THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Essays in Computational Economics

David R. Pugh

A thesis presented for the degree of
Doctor of Philosophy

School of Economics
University of Edinburgh
May 12, 2014

To my incredible wife Clarissa, sons Callan Harry and Rafsanjani, and my parents Mark and Megan. Without your continual love and encouragement, this thesis would not have been possible.

Contents

List of Figures

List of Tables

0.1 Abstract

The focus of my PhD research has been on the acquisition of computational modeling and simulation methods used in both theoretical and applied Economics.

My first chapter provides an interactive review of finite-difference methods for solving systems of ordinary differential equations (ODEs) commonly encountered in economic applications using Python. The methods surveyed in this chapter, as well as the accompanying code and IPython lab notebooks should be of interest to any researcher interested in applying finite-difference methods for solving ODEs to economic problems.

My second chapter is an empirical analysis of the evolution of the distribution of bank size in the U.S. This paper assesses the statistical support for Zipf's Law (i.e., a power law, or Pareto, distribution with a scaling exponent of $\alpha = 2$) as an appropriate model for the upper tail of the distribution of U.S. banks. Using detailed balance sheet data for all FDIC regulated banks for the years 1992 through 2011, I find significant departures from Zipf's Law for most measures of bank size inmost years. Although Zipf's Law can be statistically rejected, a power law distribution with α of roughly 1.9 statistically outperforms other plausible heavy-tailed alternative distributions.

In my final chapter, which is based on joint work with Dr. David Comerford, I apply computational methods to model the relationship between per capita income and city size. A well-known result from the urban economics literature is that a monopolistically competitive market structure combined with internal increasing returns to scale can be used to generate log-linear relations between income and population. I extend this theoretical framework to allow for a variable elasticity of substitution between factors of production in a manner similar to Zhelobodko et al. (2012). Using data on Metropolitan Statistical Areas (MSAs) in the U.S. I find evidence that supports what Zhelobodko et al. (2012) refer to as "increasing relative love for variety (RLV)." Increasing RLV generates pro-competitive effects as market size increases which means that IRS, whilst important for small to medium sized cities, are exhausted as cities become large. This has important policy implications as it suggests that focusing intervention on creating scale for small populations is potentially much more valuable than further investments to increase market size in the largest population centers.

0.2 Declaration of Own Work

I declare that this thesis was written and composed by myself and is the result of my own work unless clearly stated and references. This thesis has not been submitted for any other degree or professional qualifications. Chapter 3 of this thesis is based on joint work with Dr. David Comerford. In addition to my individual contributions to our ongoing joint work, this chapter represents a unique and significant contribution to our research agenda.

0.3 Acknowledgements

I am heavily indebted to my supervisors Profs. Andy Snell and John Hardman Moore for their support and guidance. Both of my supervisors have taken a keen interest in my work and professional development and I have learned a great deal from them. I would also like to acknowledge the support I have received from Dr. Simon Clark, Dr. Ric Holt, Prof. Jonathan Thomas, Marie Craft, and the rest of the Scottish Graduate Programme in Economics (SGPE) staff all of whom have been instrumental in helping me to set up and develop my course on *Numerical Methods for Economists*. I wish to thank Prof. Philipp Kircher, and once again, Prof. John Hardman Moore for writing references that were instrumental in helping me to secure my first academic appointment. Finally, I would like to thank Profs. Vincent Danos and Elham Kashefi for their continued friendship, support, and encouragement.

I would also like to thank my fellow PhD students for making my time here enjoyable and for providing stimulating discussion and advice. I would particularly like to thank Drs. Sean Brocklebank, David Comerford, James Best and Keshav Dogra for providing helpful comments and feedback along the way. I am very grateful to Profs. Ed Hopkins, Michèle Belot, Tim Worrall, and once again Prof. Andy Snell for creating a positive and supporting environment in which my fellow PhD students and I could work. I also wish to thank Lesley Mayné for shepherding me through the submission process.

Chapter 1

Finite-difference methods for solving ODEs: A Python-based survey

This chapter reviews finite-difference methods for solving systems of ordinary differential equations (ODEs) commonly encountered in economic applications using Python. The methods surveyed in this chapter, as well as the accompanying code and IPython lab notebooks, should be of interest to any researcher interested in applying finite-difference methods for solving ODEs to economic problems.

1.1 Introduction

This chapter explores the use of the Python programming language for solving types of ordinary differential equations (ODEs) commonly encountered in economics using finite-difference methods. The major contribution of this chapter is pedagogical. While the economics graduate curriculum often includes in-depth training in theoretical branches of mathematics such as measure theory and real analysis, very few economics department offer graduate courses in applied mathematics or numerical methods. This general lack of awareness of numerical methods has significantly limited the types of questions that economic analysis can be used to address. This chapter seeks to partially fill this significant gap in the training of economics graduate students. The methods surveyed in this chapter, as well as the accompanying Python code and IPython notebooks which implement them should be of use to any economist interested in applying finite-difference methods for

solving ODEs to economic problems.[1]

I have chosen to focus on numerical methods for solving ODEs because differential equations have been used to model a wide variety of economic phenomena. For example, the neoclassical optimal growth model of Ramsey (1928), Cass (1965), and Koopmans (1965), the Solow (1956) model, as well as the Pissarides (1985) and Mortensen and Pissarides (1994) search models of unemployment are all examples of ODEs used to model economic dynamics. Equilibria of signaling models, such as Spence (1974) and its descendants, are often characterized by ODEs. Recent models of assortative matching, such as Eeckhout and Kircher (2010), Eeckhout and Kircher (2012), and references therein, typically reduce to systems of ODEs. As a final example, as demonstrated by Hubbard and Paarsch (2009), Hubbard et al. (2011), and Hubbard et al. (2012), and references therein, equilibrium bid functions in many auction models can often be characterized as solutions to systems of ODEs.[2]

As the main focus of this chapter is pedagogical, instead of focusing on cutting edge applications of finite-difference methods in economics, I have decided to focus on three models that are taught to first-year students in most any economics graduate program: the optimal growth model of Ramsey (1928), Cass (1965), and Koopmans (1965), the Solow (1956) model, and the Spence (1974) model of "signaling." In their most general formulations, none of these models can be solved analytically and thus researchers are forced to either make use of numerical methods to approximate the full non-linear solutions, or to resort to potentially inaccurate and misleading linear approximations of their non-linear solutions. Each of the models does, however, have an analytic solution for some specific set of parameter restrictions. I will make heavy use of these analytic results to compare the accuracy and computational efficiency of the numerical methods.[3]

My results suggest a number of "best practices" that all economic researchers should adhere to when solving ODEs.

1. Classic finite difference methods with fixed step-size such as variants of Euler's method (discussed in sections 1.3.1-1.3.3) or the family of explicit Runge-Kutta methods (discussed in section 1.3.5) should be avoided. While such methods are easy to code, they tend to be computationally inefficient and will be orders of magnitude less accurate than more modern methods which implement adaptive step-size control (discussed in sections 1.3.4-1.3.4 and 1.3.5)

[1] The Python code and IPython notebooks used to obtain the results reported in this chapter are available online via the author's GitHub repository.

[2] For a smorgasbord of additional examples of economic models using ODEs see Brock and Malliaris (1989) and Zhang (2005).

[3] Full derivations of these analytic solutions are provided in the appendices.

2. Of the high-quality ODE solvers currently available via the `scipy.optimize` module, the embedded Runge-Kutta methods due to Dormand and Prince (1980), `dopri5` and `dop853` (discussed in section 1.3.5) are generally the most accurate "out of the box" (i.e., without changing any of the default tolerances).

3. When solving BVPs using finite-difference methods (discussed in section 1.5) it is important to remember that the approximation error for multi-layered algorithms is determined by the interaction between the approximation errors of the individual layers. For shooting methods (discussed in sections 1.6.1-1.6.2) it may be necessary to set a relatively loose error tolerance in the outer layer in order for the algorithm to terminate. Using an ODE solver with adaptive step-size control in the inner layer will slow down the rate at which error accumulates in the inner layer of the algorithm, which in turn, will allow the researcher to set a tighter error tolerance in the outer layer.

4. Where applicable, reverse shooting (discussed in section 1.6.2) is preferred over forward shooting for solving BVPs. Reverse shooting is more computationally efficient, more numerically stable, and significantly more accurate than forward shooting.

5. When comparing run times times across the various methods, it is the relative (and not absolute) speed which matters. While absolute speed of any particular method will vary across computers, the relative speed of various methods should be fairly stable.

The remainder of this paper proceeds as follows. Section 1.2 provides some background information on the Python programming language and formally defines many of the key mathematical concepts used throughout the paper. Section 1.3 surveys finite-difference methods for solving IVPs. Section 1.4 shows how to solve some classic IVPs from economics using Python. Sections 1.5 and 1.6 do the same for BVPs. Section 1.7 concludes.

1.2 Basic definitions

In this section I briefly summarize some of the key reasons for my use of the Python programming language, before formally defining the key mathematical concepts and notation used throughout the paper.

1.2.1 Python

Python is a modern, object-oriented programming language widely used in academia and private industry, whose clean, yet expressive syntax, makes it an easy programming language to learn while still remaining powerful enough for serious scientific computing. Python's syntax was designed from the start with the human reader in mind and generates code that is easy to understand and debug which shortens development time relative to low-level, compiled languages such as Fortran and C++. Among the high-level, general purpose languages, Python has the largest number of MATLAB®-style library modules (both installed in the standard library and through additional downloads) which meaning that one can quickly construct sophisticated programs.[4]

Python is completely free and platform independent, making it a very attractive option as a teaching platform relative to other high-level scripting languages, particularly MATLAB®. Python is also open-source, making it equally attractive as a research tool for scientists interested in generating computational results that are more easily reproducible.[5] Finally, Python comes with a powerful interactive interpreter that allows real-time code development and live experimentation. The functionality of the basic Python interpreter can be greatly increased by using the Interactive Python (or IPython) interpreter. Working via the Python or IPython interpreter eliminates the time-consuming (and productivity-destroying) compilation step required when working with low-level languages at the expense of slower execution speed.[6]

While the Python programming language has found widespread use in private industry and many fields within academia,[7] the capabilities of Python as a research tool remain relatively unknown within the economics research community.[8]

[4]The IPython, Pandas, NumPy, SciPy, SymPy, Matplotlib and Mayavi libraries form the core of the Python scientific computing stack.

[5]The Python Software Foundation License (PSFL) is a BSD-style license that allows a developer to sell, use, or distribute his Python-based application in anyway he sees fit. In addition, the source code for the entire Python scientific computing stack is available on GitHub making it possible to directly examine the code for any specific algorithm in order to better understand exactly how a result has been obtained.

[6]In many cases, it may be possible to achieve the best of both worlds using "mixed language" programming. Python can be easily extended by wrapping compiled code written in FORTRAN, C/C++ using libraries such as f2Py, Cython, or swig. See Peterson (2009), Behnel et al. (2011) and references therein for more details.

[7]A non-exhaustive list of organizations currently using Python for scientific research and teaching: MIT's legendary *Introduction to Computer Science and Programming*, CS 6.00, is taught using Python; Python is the in-house programming language at Google; NASA, CERN, Los Alamos National Labs (LANL), Lawrence Livermore National Labs (LLNL), and Industrial Light and Magic (ILM) all rely heavily on Python.

[8]Notable exceptions are Stachurski (2009) and Sargent and Stachurski (2013).

1.2.2 Ordinary differential equation (ODE)

An ODE is in equation of the form

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}) \tag{1.1}$$

where $\mathbf{f} : [t_0, \infty) \times \mathbb{R}^n \to \mathbb{R}^n$. In the case where $n = 1$, then equation 1.1 reduces to a single ODE; when $n > 1$, equation 1.1 defines a system of ODEs. ODEs are one of the most basic examples of functional equations: the solution to equation 1.1 is a function $\mathbf{y}(t) : D \subset \mathbb{R} \to \mathbb{R}^n$.

1.2.3 Initial value problems (IVPs)

An initial value problem (IVP) has the form

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \ t \geq t_0, \ \mathbf{y}(t_0) = \mathbf{y}_0 \tag{1.2}$$

where $\mathbf{f} : [t_0, \infty) \times \mathbb{R}^n \to \mathbb{R}^n$ and the initial condition $\mathbf{y}_0 \in \mathbb{R}^n$ is a given vector.[9] The unknown in this problem is the function $\mathbf{y}(t) : [t_0, T] \subset \mathbb{R} \to \mathbb{R}^n$ that satisfies the initial condition $\mathbf{y}(t_0) = \mathbf{y}_0$. So long as the function $\mathbf{f}$ is reasonably well-behaved, the function $\mathbf{y}(t)$ exists and is unique.[10]

1.2.4 Boundary value problems (BVPs)

First order differential equations in one variable constitute IVPs by default: with only a single equation, I can fit the function $y(t)$ at only one $t \in [t_0, T]$. However, with $n > 1$, the auxiliary conditions can fit the various components of the function $\mathbf{y}(t)$ at different values of t. The key difference between an initial value problem and a boundary value problem is that with initial value problems the side conditions pin down the solution, $y(t)$, at a single point, whereas with boundary value problems pin down $y(t)$ at several points.

A two-point boundary value problem (2PBVP) imposes n conditions on the function $\mathbf{y}(t)$

[9] Alternatively, I could also specify an initial value problem by imposing a terminal condition of the form $\mathbf{y}(T) = \mathbf{y}_T$. The key point is that the solution $\mathbf{y}(t)$ is pinned down at one $t \in [t_0, T]$.

[10] Brock and Malliaris (1989) provide an existence proof for a solution to 1.1 under very general conditions. Uniqueness of the solution requires that the function $\mathbf{f}$ satisfy a Lipschitz condition of the form

$$||\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})|| \leq \lambda ||\mathbf{x} - \mathbf{y}|| \text{ for all } \mathbf{x}, \ \mathbf{y} \in \mathbb{R}^n, \ t \geq t_0.$$

Very readable proofs of existence and uniqueness when the function $\mathbf{f}$ satisfies this Lipschitz condition can be found in Iserles (2009).

of the form

$$g_i(\mathbf{y}(t_0)) = 0, i = 1, \ldots, n' \tag{1.3}$$
$$g_i(\mathbf{y}(T)) = 0, i = n' + 1, \ldots, n \tag{1.4}$$

where $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$. More generally, a multi-point BVP imposes

$$g_i(\mathbf{y}(t_i)) = 0 \tag{1.5}$$

for a set of points $t_i, t_0 \leq t_i \leq T, 1 \leq i \leq n$ where $T = \infty$ denotes some condition on the $\lim_{t\to\infty} \mathbf{y}(t)$.

With the auxiliary, or boundary, conditions defined as above the general formulation of a BVP is

$$\begin{aligned} \mathbf{y}' &= \mathbf{f}(t, \mathbf{y}),\ t \geq t_0, \\ g_i(\mathbf{y}(t_i)) &= 0,\ 1 \leq i \leq n \end{aligned} \tag{1.6}$$

for a set of points $t_i, t_0 \leq t_i \leq \infty$. The solution to this boundary value problem is the function $\mathbf{y}(t) : [t_0, T] \subset \mathbb{R} \to \mathbb{R}^n$ that satisfies the boundary conditions. Unlike IVPs, with BVPs neither existence of a solution nor its uniqueness is guaranteed.

1.3 Finite-difference methods for IVPs

This section provides a non-technical survey of finite-difference methods for approximating solutions to IVPs of the form

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}),\ t \geq t_0,\ \mathbf{y}(t_0) = \mathbf{y}_0 \tag{1.7}$$

where $\mathbf{f} : [t_0, \infty) \times \mathbb{R}^n \to \mathbb{R}^n$ is some function, and $\mathbf{y}_0 \in \mathbb{R}^n$ is an initial condition. Finite-difference methods are perhaps the most commonly used class of numerical methods for approximating solutions to ODEs. The basic idea behind all finite-difference methods is to construct a difference equation

$$\mathbf{y}(t_i)' = \mathbf{f}(t_i, \mathbf{y}(t_i)) \tag{1.8}$$

which is "similar" to the differential equation at some grid of values $t_0 < \cdots < t_N$. Finite-difference methods then "solve" the original differential equation by finding for each

$n = 0, \ldots, N$ a value $\mathbf{y}_n$ that approximates the value of the solution $\mathbf{y}(t_n)$.[11] The literature on finite-difference methods for solving ODEs is vast and there are many excellent reference texts. Those interested in a more in-depth treatment of these topics, including formal proofs of convergence, order, and stability of the methods discussed in this section, should consult Hairer et al. (1993), Butcher (2008), and Iserles (2009). Chapter 10 of Judd (1998) covers a subset of this material with a specific focus on economic applications.

1.3.1 The forward Euler method

The simplest scheme for approximating the solution to 1.7 is called the explicit, or forward, Euler method. The basic idea behind the forward Euler method is to estimate the solution $\mathbf{y}(t)$ by making the approximation

$$\mathbf{f}(t, \mathbf{y}(t)) \approx \mathbf{f}(t_0, \mathbf{y}(t_0)) \tag{1.9}$$

for $t \in [t_0, t_0 + h]$ and some sufficiently small $h > 0$. Integrating 1.7 and applying the Euler approximation yields the following.

$$\mathbf{y}(t) = \mathbf{y}(t_0) + \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{y}(\tau))d\tau \approx \mathbf{y}_0 + (t - t_0)\mathbf{f}(t_0, \mathbf{y}_0) \tag{1.10}$$

Applying equation 1.10 to the initial condition yields the Euler estimate for $\mathbf{y}_1$

$$\mathbf{y}_1 = \mathbf{y}_0 + (t_1 - t_0)\mathbf{f}(t_0, \mathbf{y}_0). \tag{1.11}$$

Repeated application of this approximation scheme yields a general, recursive formulation for the forward Euler method

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(t_n, \mathbf{y}_n), \ n = 0, 1, \ldots, N \tag{1.12}$$

where $h = t_{n+1} - t_n > 0$ denotes the size of the step. The forward Euler method is an order one method: as the step-size, h, shrinks to zero, the approximation error using the forward Euler method decays as $\mathcal{O}(h)$.[12] A graphical illustration of the forward Euler method is

[11] Note that finite-difference methods only approximate the solution $\mathbf{y}$ at the N grid points. In order to approximate $\mathbf{y}$ between grid points I must resort to some form of interpolation. See chapter 6 of Judd (1998) for an introduction to interpolation methods commonly used in economics research.

[12] Order of convergence is related to the speed of decay of the approximation error. Let $\mathbf{y}_{n,h}$ be the numerical approximation of $\mathbf{y}(t_n)$ when the size of the step is h. The approximation error of an order p method satisfies

$$\lim_{h \to 0^+} \max_{n=0,1,\ldots,\lfloor \frac{t^*}{h} \rfloor} \frac{||\mathbf{y}_{n,h} - \mathbf{y}(t_n)||}{h^p} < \infty. \tag{1.13}$$

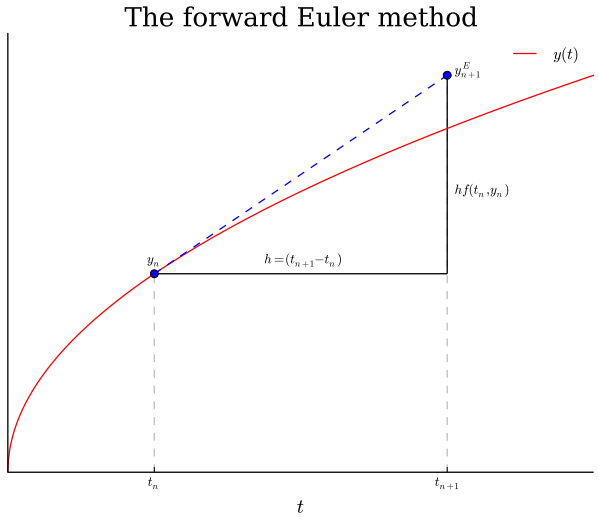

Figure 1.1: The forward Euler method assumes that the slope of the solution, $\mathbf{y}(t)$, over the interval $[t_n, t_{n+1}]$ is constant and equal to its value at the point $(t_n, \mathbf{y}_n)$.

provided in figure 1.1.

1.3.2 The backward Euler method

The forward Euler method approximates the slope of the solution $\mathbf{y}(t)$ over the interval $[t_n, t_{n+1}]$ using the slope of the solution at the point $(t_n, \mathbf{y}_n)$. An obvious alternative to this procedure would be to approximate the slope of $\mathbf{y}(t)$ over the interval using the slope of the solution at $(t_{n+1}, \mathbf{y}_{n+1})$. This is the basic idea behind the backward, or implicit, Euler method.

Starting from some initial condition $\mathbf{y}_0$, the backward Euler method estimates $\mathbf{y}(t_1)$ as

$$\mathbf{y}_1 = \mathbf{y}_0 + h\mathbf{f}(t_1, \mathbf{y}_1). \tag{1.14}$$

Equation 1.14 is a non-linear equation in the unknown $\mathbf{y}_1$ where $\mathbf{y}_1$ is defined only implicitly in terms of t_0, t_1, and $\mathbf{y}_0$ (all of which are taken as given).[13] Repeated application of this

[13] Numerical routines for solving non-linear equations typically require the user to initially guess the value of the solution $\mathbf{y}_1$. A good strategy for generating an guess in this context is to use a simple explicit method, such as the forward Euler method, to generate a"predicted" value of $\mathbf{y}_1$. The implicit method can then be thought of as "correcting" the prediction of the explicit method. This common implementation strategy is called a "predictor-corrector" approach in the numerical analysis literature.

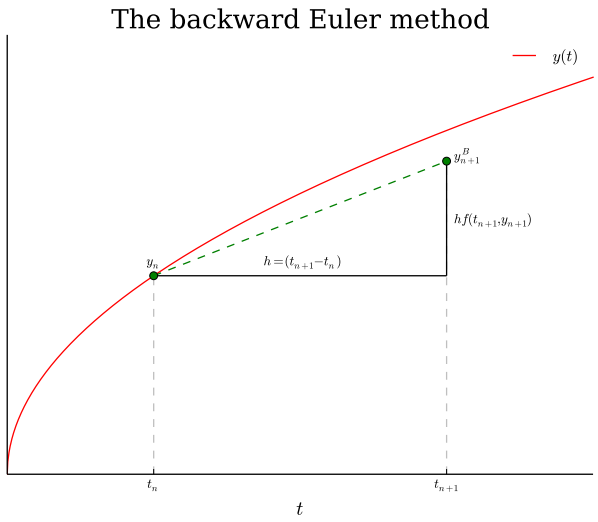

Figure 1.2: The backward Euler method assumes that the slope of the solution, $\mathbf{y}(t)$, over the interval $[t_n, t_{n+1}]$ is constant and equal to its value at the point $(t_{n+1}, \mathbf{y}_{n+1})$.

idea yields a recursive formulation of the backward Euler method

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(t_n, \mathbf{y}_n) \tag{1.15}$$

where $h = t_{n+1} - t_n > 0$ denotes the size of the step. Like the forward Euler method, the approximation error of the backward Euler method decays as $\mathcal{O}(h)$. A graphical illustration of the backward Euler method can be found in figure 1.2.

1.3.3 The trapezoidal rule

The forward (backward) Euler method assumes that the slope of the solution in the interval $[t_n, t_{n+1}]$ is constant and equal to its value at the left (right) end point. One might guess that a better way to approximate the derivative of the solution by a constant over an interval would be to take a simple average of the slope of the solution at both end points. This simple intuition is formalized by the following approximation scheme

$$\begin{aligned}\mathbf{y}(t) =& \mathbf{y}(t_n) + \int_{t_n}^{t} \mathbf{f}(\tau, \mathbf{y}(\tau))d\tau \\ \approx& \mathbf{y}(t_n) + \frac{1}{2}(t - t_n)\mathbf{f}(t_0, \mathbf{y}(t_n)) + \frac{1}{2}(t - t_n)\mathbf{f}(t_{n+1}, \mathbf{y}(t_{n+1}))\end{aligned} \tag{1.16}$$

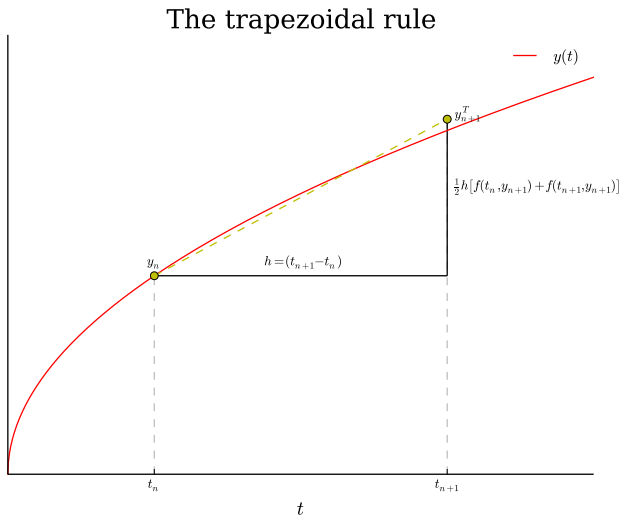

Figure 1.3: The trapezoidal rule is an implicit, order two method, that exhibits quadratic convergence to the solution $\mathbf{y}(t)$ as the step-size, h, shrinks to zero.

which leads to the trapezoidal rule. A graphical illustration of the trapezoidal rule is provided in figure 1.3.[14]

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{1}{2}h\mathbf{f}(t_n, y_n) + \frac{1}{2}h\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) \tag{1.17}$$

The trapezoidal method is quadratically convergent implying that the approximation error decays globally as $\mathcal{O}(h^2)$. Put more simply, when using the trapezoidal rule, reducing the step-size by a factor of 10 will reduce the approximation error by a factor of 10^2.[15] The higher order of convergence means that the trapezoidal rule can often make due with a larger step-size than the Euler method while still maintaining the same level of approximation error. Like the backward Euler method, the trapezoidal method is an implicit method and each step requires solving a non-linear equation in the unknown

[14] Both Euler method and the trapezoidal rule are special cases of a more general approximation scheme which approximates the derivative of the solution in the interval $[t_n, t_{n+1}]$ by taking a linear combination of the slope of the solution at both end points.

$$y_{n+1} = y_n + h\Big[\theta f(t_n, y_n) + (1-\theta)f(t_{n+1}, y_{n+1})\Big],\ n = 0, 1, \ldots$$

To recover the forward Euler method simply set $\theta = 1$ and to recover the trapezoidal rule set $\theta = \frac{1}{2}$. Setting $\theta = 0$ recovers the backward Euler method.

[15] Contrast this with both the forward and backward Euler methods whose linear rates of convergence imply that reducing the step-size by a factor of 10 will only reduce the approximation error by a factor of 10.

$\mathbf{y}_{n+1}$ as well as extra evaluation of the function $\mathbf{f}$. In practice, however, the more rapid convergence of the trapezoidal method and ability to use a larger step-size more than compensate for this additional computational burden.

1.3.4 Linear multi-step methods

All of the methods discussed above are called single-step methods as they use information about a single point and its derivative to approximate the next point of the solution. In general, multi-step methods attempt to gain efficiency by incorporating information used to compute previous points of the solution trajectory and leveraging it to approximate the current point of the solution. Linear multi-step methods use a linear combination of $\mathbf{y}_n$ and $\mathbf{f}(t_n, \mathbf{y}_n)$ from the previous s steps to approximate the next step in the solution, $\mathbf{y}_{n+1}$.

$$\begin{aligned} a_s\mathbf{y}_{n+s} + a_{s-1}\mathbf{y}_{n+s-1} + \cdots + a_0\mathbf{y}_n = \\ h\Big[b_s\mathbf{f}(t_{n+s}, \mathbf{y}_{n+s}) + b_{s-1}\mathbf{f}(t_{n+s-1}, \mathbf{y}_{n+s-1}) + \cdots + b_0\mathbf{f}(t_n, \mathbf{y}_n)\Big] \end{aligned} \tag{1.18}$$

There are three main classes of linear multi-step methods: the Adams-Bashforth (AB) methods, which are explicit, and the Adams-Moulton (AM) and backwards differentiation formulae (BDF) methods, both of which are implicit. These classes are differentiated from one another by the restrictions each imposes on the coefficients $a_0, \ldots, a_s$ and $b_0, \ldots, b_s$ and by how the values of the remaining, unrestricted, coefficients are chosen.

Adams-Bashforth (AB) methods set $a_s = 1, a_{s-1} = -1$ and impose $a_{s-2}, \ldots, a_0 = 0$ and $b_s = 0$. Applying these restrictions yields a reduced-form version of 1.18

$$\mathbf{y}_{n+s} = \mathbf{y}_{n+s-1} + h\Big[b_{s-1}\mathbf{f}(t_{n+s-1}, \mathbf{y}_{n+s-1}) + \cdots + b_0\mathbf{f}(t_n, \mathbf{y}_n)\Big] \tag{1.19}$$

where $h > 0$ is the step-size. The Adams-Moulton (AM) methods impose that same restrictions on the coefficients $a_0, \ldots, a_s$ but, being implicit methods, allow $b_s \neq 0$.

$$\mathbf{y}_{n+s} = \mathbf{y}_{n+s-1} + h\Big[b_s\mathbf{f}(t_{n+s}, \mathbf{y}_{n+s}) + b_{s-1}\mathbf{f}(t_{n+s-1}, \mathbf{y}_{n+s-1}) + \cdots + b_0\mathbf{f}(t_n, \mathbf{y}_n)\Big] \tag{1.20}$$

To derive both the coefficients $b_0, \ldots, b_{s-1}$ for the s-step Adams-Bashforth method and the coefficients $b_0, \ldots, b_s$ for the s-step Adams-Moulton method, start by integrating 1.7

over the interval $[t_{n+s-1}, t_{n+s}]$.

$$\mathbf{y}(t_{n+s}) = \mathbf{y}(t_{n+s-1}) + \int_{t_{n+s-1}}^{t_{n+s}} \mathbf{f}(\tau, \mathbf{y}(\tau))d\tau \tag{1.21}$$

The idea behind both the AB and AM methods is to use past values of the solution to construct an interpolating polynomial, $p(t)$, that approximates $\mathbf{f}$ over the interval of integration, $[t_{n+s-1}, t_{n+s}]$. The s-step AB method, which restricts $b_s = 0$, uses a degree $s-1$ polynomial, while the s-step AM method uses a degree s polynomial.[16]

The interpolating polynomial, $p(t)$, of degree $s-1$ used in the s-step AB method matches the function $\mathbf{f}(t_m, y_m)$ for $m = n, n+1, \ldots, n+s-1$ and is defined as

$$p(t) = \sum_{m=0}^{s} p_m(t)\mathbf{f}(t_{n+m}, \mathbf{y}_{n+m}) \tag{1.22}$$

where the functions

$$p_m(t) = \prod_{l=0, l\neq m}^{s-1} \frac{t - t_{n+l}}{t_{n+m} - t_{n+l}} = \frac{(-1)^{s-m-1}}{m!(s-m-1)!} \prod_{l=0, l\neq m}^{s-1} \left(\frac{t - t_n}{h} - l\right) \tag{1.23}$$

for all $m = 0, 1, \ldots, s-1$ are the Lagrange interpolating polynomials. To obtain the coefficients $b_0, \ldots, b_{s-1}$ substitute the interpolating polynomial, $p(t)$ for $\mathbf{f}$ in equation 1.21 and note that since p is a polynomial, the integral on the right-hand side of this expression can be solved exactly. After integrating along an interval of length h, and replacing $\mathbf{y}(t_{n+s})$ and $\mathbf{y}(t_{n+s-1})$ with their numerical counterparts $\mathbf{y}_{n+s}$ and $\mathbf{y}_{n+s-1}$, one obtains

$$\mathbf{y}_{n+s} = \mathbf{y}_{n+s-1} + h \sum_{m=0}^{s-1} b_m \mathbf{f}(t_{n+m}, \mathbf{y}_{n+m}) \tag{1.24}$$

where the coefficients $b_0, b_1, \ldots, b_{s-1}$ are defined as

$$b_m = \frac{1}{h} \int_{t_{n+s-1}}^{t_{n+s}} p_m(\tau)d\tau = \frac{1}{h} \int_0^h p_m(t_{n+s-1} + \tau)d\tau, \;\; m = 0, 1, \ldots, s-1 \tag{1.25}$$

Equations 1.24 and 1.25 comprise the s-step AB method which has order s implying that the approximation error decays globally as $\mathcal{O}(h^s)$. Note that the Adams-Bashforth method with $s = 1$ is equivalent to the forward Euler method.

Similar arguments can be used to derive expressions for the coefficients $b_0, \ldots, b_s$ used in

[16] The extra restriction imposed by the explicit Adams-Bashforth method reduces the available degrees of freedom by one and results in an interpolating polynomial of only degree $s-1$.

the s-step AM method:

$$\mathbf{y}_{n+s} = \mathbf{y}_{n+s-1} + h \sum_{m=0}^{s} b_m \mathbf{f}(t_{n+m}, \mathbf{y}_{n+m}) \tag{1.26}$$

$$b_m = \frac{1}{h} \int_{t_{n+s-1}}^{t_{n+s}} p_m(\tau) d\tau = \frac{1}{h} \int_0^h p_m(t_{n+s-1} + \tau) d\tau, \ m = 0, 1, \ldots, s \tag{1.27}$$

where the functions

$$p_m(t) = \prod_{l=0, l \neq m}^{s} \frac{t - t_{n+l}}{t_{n+m} - t_{n+l}} = \frac{(-1)^{s-m}}{m!(s-m)!} \prod_{l=0, l \neq m}^{s} \left(\frac{t - t_n}{h} - l \right) \tag{1.28}$$

are the Lagrange interpolating polynomials. The s-step AM method has order $s + 1$ implying that the approximation error decays globally as $\mathcal{O}(h^{s+1})$. Note that the Adams-Moulton methods nest the backward Euler method and the trapezoidal rule as special cases. To recover the backward Euler method set $s = 1$; to recover the trapezoidal rule set $s = 2$.

Backwards differentiation formula (BDF) methods take a completely different approach. BDF methods normalize $a_s = 1$ and impose $b_{s-1} = b_{s-2} = \cdots = b_0 = 0$ yielding the following approximation formula.

$$\mathbf{y}_{n+s} + a_{s-1} \mathbf{y}_{n+s-1} + \cdots + a_0 \mathbf{y}_n = h b_s \mathbf{f}(t_{n+s}, \mathbf{y}_{n+s}) \tag{1.29}$$

When choosing the values of the remaining, unrestricted, coefficients $a_0, \ldots, a_{s-1}$ the guiding principal is to choose values that maximize the order at which the approximation error decays subject to the constraint that the resulting BDF method remains convergent.[17] These considerations yields a simple formula for the coefficient b_s.

$$b_s = \beta = \left(\sum_{m=1}^{s} \frac{1}{m} \right)^{-1} \tag{1.30}$$

Given b_s, the coefficients $a_0, \ldots, a_{s-1}$ are then set equal to the coefficients of the following polynomial.

$$\rho(w) = \beta \sum_{m=1}^{s} \frac{1}{m} w^{s-m} (w - 1)^m \tag{1.31}$$

BDF methods have order s but are convergent only for $1 \leq s \leq 6$. In practice, however, only s-step BDF methods with $1 \leq s \leq 5$ are commonly used. I have already discussed

[17] For more details on the subtle trade-offs between order and convergence for multi-step methods in general and BDF methods in particular, see Iserles (2009).

the simplest example of a BDF. The single step BDF, which has order $s = 1$, is just the backward Euler method. The formulas for the remaining BDF methods can be found in Hairer et al. (1993).

Adaptive step-size control

The linear multi-step methods that I have discussed so far in this section have all used a fixed step-size, h. Most high quality linear multi-step integrators, however, exert significant adaptive control over their own progress by frequently adjusting the step-size based on estimates of the local truncation error.[18]

The objective of adaptive step-size control is to minimize the computational burden of approximating the solution by adjusting the step-size, h, in order to achieve some predetermined level of accuracy. The basic principle behind all adaptive step-size schemes is best demonstrated via a hillwalking analogy. An experienced hillwalker takes relatively few large strides when moving over smooth, uninteresting countryside, and a large number of short, carefully-placed steps when moving over treacherous mountain terrain.

The same underlying logic extends to integrating ODEs. When the solution trajectory is relatively flat, a good integrator understands that it can take larger steps without sacrificing too much accuracy and thus increases the step-size accordingly. However, when the solution trajectory is rapidly varying, a good integrator knows that it may need to adjust the step-size down in order to maintain its predetermined level of accuracy.

The literature on adaptive step-size control is vast with optimal approaches tending to be integrator specific. Those interested in a more detailed discussion of these, and other, methods for achieving adaptive step-size control using linear multi-step integrators should consult, Hairer et al. (1993), Butcher (2008), or Press et al. (2009). The remainder of this section illustrates a general approach to adaptive step-size control, called the Milne device, variants of which are commonly found in off the shelf linear multi-step integrators.

The idea behind the Milne device is to use two order p linear multi-step methods, one explicit and one implicit, to estimate the local truncation error of the implicit method. Consider the following example. Suppose that I wish to estimate the local truncation error of an implicit second-order Adams-Moulton method (i.e., the trapezoidal rule)

$$\mathbf{y}_{n+1}^{TR} = \mathbf{y}_n + \frac{1}{2}h_n\mathbf{f}(t_n, y_n) + \frac{1}{2}h_n\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})$$

[18] Local truncation error is the error incurred when integrating the system from t_n to t_{n+1} under the standard assumption that the value of $\mathbf{y}_n$ is exact.

using an explicit 2nd-order Adams-Bashforth method

$$\mathbf{y}_{n+1}^{AB} = \mathbf{y}_n + \frac{3}{2}h_n\mathbf{f}(t_n, y_n) - \frac{1}{2}h_n\mathbf{f}(t_{n-1}, \mathbf{y}_{n-1}).$$

Deriving the local truncation error for these methods using a Taylor expansion yields

$$\mathbf{y}(t_{n+1}) - \mathbf{y}_{n+1}^{TR} = -\frac{1}{12}h_n^3\mathbf{y}'''(t_n) + \mathcal{O}(h_n^4)$$
$$\mathbf{y}(t_{n+1}) - \mathbf{y}_{n+1}^{AB} = \frac{5}{12}h_n^3\mathbf{y}'''(t_n) + \mathcal{O}(h_n^4)$$

where $c_{TR} = -\frac{1}{12}$ and $c_{AB} = \frac{5}{12}$ are the error constants for the trapezoidal rule and the second-order AB method, respectively. Subtracting the bottom expression from the top yields an estimate of $h_n\mathbf{y}'''(t_n)$

$$h_n\mathbf{y}'''(t_n) \approx -2(\mathbf{y}_{n+1}^{AB} - \mathbf{y}_{n+1}^{TR})$$

which can be substituted back into the trapezoidal rule formula obtain an estimate of its local truncation error.

$$\mathbf{e}_{n+1} = \mathbf{y}(t_{n+1}) - \mathbf{y}_{n+1}^{TR} \approx \frac{1}{6}(\mathbf{y}_{n+1}^{AB} - \mathbf{y}_{n+1}^{TR})$$

Now that I have an estimate of the local truncation error for the trapezoidal rule, to complete my adaptive step-size scheme I need a rule for updating the step-size. Perhaps the most widely used approach for updating the step-size, called error control per unit step, computes h_{n+1} using

$$h_{n+1} = \beta h_n \left[\frac{\delta}{\left|\left|\frac{\mathbf{e}_{n+1}}{h_n}\right|\right|_\infty}\right]^{\frac{1}{2}}$$

where $\delta > 0$ is some predefined error tolerance and $0 < \beta < 1$ is a tuning parameter.

The above example is generalized for arbitrary linear multi-step methods in Iserles (2009). The general formula for estimating local truncation error is

$$\mathbf{y}(t_{n+s}) - \mathbf{y}_{n+s} \approx \frac{c}{c - \tilde{c}}(\mathbf{y}_{n+s} - \tilde{\mathbf{y}}_{n+s}),\ s = 0, 1, \ldots \tag{1.32}$$

where c and $\tilde{c}$ are error constants associated with the order p linear multi-step methods used to obtain $\mathbf{y}$ and $\tilde{\mathbf{y}}$. After computing the estimated local truncation error, the step-size

h_n is updated according to

$$h_{n+1} = \beta h_n \left[\frac{\delta}{\left|\left| \frac{\mathbf{e}_{n+1}}{h_n} \right|\right|_\infty} \right]^{\frac{1}{p}}. \tag{1.33}$$

1.3.5 Runge-Kutta (RK) methods

Like linear multi-step methods, RK methods can be thought of as a generalization of the basic Euler methods. However, unlike linear multi-step methods, which use current and previous values of the solution to approximate the value of the solution at the next step, RK methods approximate the value of the solution at the next step by combining multiple evaluations of the derivative at various points within the current step.

Explicit RK methods

To construct an explicit RK method, start by applying the fundamental theorem of calculus to integrate the solution from t_n to $t_{n+1} = t_n + h$.

$$\begin{aligned} \mathbf{y}(t_{n+1}) =& \mathbf{y}(t_n) + \int_{t_n}^{t_{n+1}} \mathbf{f}(\tau, \mathbf{y}(\tau)) d\tau \\ =& \mathbf{y}(t_n) + \int_0^1 \mathbf{f}(t_n, \mathbf{y}(t_n + h\tau)) d\tau \end{aligned}$$

The next step is to approximate the integral on the right-hand side by the following weighted average.[19]

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h \sum_{j=1}^{\nu} b_j \mathbf{f}(t_n + c_j h, \mathbf{y}(t_n + c_j h)), \ n = 0, 1, \ldots \tag{1.34}$$

The final step in constructing an explicit RK rule is to approximate each $\mathbf{y}(t_n + c_j h)$ for $j = 1, \ldots, \nu$. Let ξ_j denote the numerical estimate of $\mathbf{y}(t_n + c_j h)$ for $j = 1, \ldots, \nu$. Explicit RK methods approximate each of the ξ_j by updating $\mathbf{y}_n$ with a linear combination of $\mathbf{f}(t_n, \xi_1), \mathbf{f}(t_n + hc_2, \xi_2), \ldots, \mathbf{f}(t_n + c_{j-1}h, \xi_{j-1})$.[20] Specifically, define the RK stages ξ_j for

[19] In the literature on numerical integration the above weighted average is called a quadrature rule. The collection of points $c_1, \ldots, c_\nu \in [t_n, t_{n+1}]$ are called quadrature nodes; and the values $b_1, \ldots, b_\nu$ are called quadrature weights.

[20] By convention $c_1 = 0$ which implies that $\xi_1 = \mathbf{y}_n$.

$j = 1, \ldots, \nu$ as follows.

$$\begin{aligned}
\xi_1 =& \mathbf{y}_n \\
\xi_2 =& \mathbf{y}_n + h a_{2,1} \mathbf{f}(t_n, \xi_1) \\
\xi_3 =& \mathbf{y}_n + h a_{3,1} \mathbf{f}(t_n, \xi_1) + h a_{3,2} \mathbf{f}(t_n + c_2 h, \xi_2) \\
& \vdots \\
\xi_\nu =& \mathbf{y}_n + h \sum_{i=1}^{\nu-1} a_{\nu,i} \mathbf{f}(t_n + c_i h, \xi_i)
\end{aligned}$$

The lower triangular matrix $A = (a_{j,i})_{j,i=1,2,\ldots,\nu}$, where missing elements are defined to be zero, is called the RK matrix, while

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_\nu \end{pmatrix}, \text{ and } \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_\nu \end{pmatrix}$$

are called the RK weights and RK nodes, respectively. Combining the ν stages with the weights, $\mathbf{b}$, and nodes, $\mathbf{c}$, leads to the following general formula for an explicit RK methods.

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h \sum_{j=1}^{\nu} b_j \mathbf{f}(t_n + c_j h, \xi_j) \tag{1.35}$$

RK methods are commonly represented by a partitioned tableau with the following form.

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \end{array} \tag{1.36}$$

There are an almost endless number of explicit RK methods. See Gear (1971), Shampine (1986), and particularly Butcher (2008) for a comprehensive listing of formulas for different explicit RK methods. The tableaus for several of the more commonly encountered methods: a two-stage, second-order RK method called the mid-point rule, and the classic third and

fourth-order RK methods are provided below.

Second-order RK:

0		
$\frac{1}{2}$	$\frac{1}{2}$	
	0	1

Third-order RK:

0			
$\frac{1}{2}$	$\frac{1}{2}$		
1	-1	2	
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Fourth-order RK:

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

Variants of the fourth-order RK method have been used as default solvers in many software packages for decades. The sustained popularity of the fourth-order RK method can be attributed to the fact that it is both highly accurate and cheap to compute.[21]

Implicit RK methods

Implicit Runge-Kutta (IRK) methods allow the RK stages $\xi_1, \ldots, \xi_\nu$ to depend on each other in a more general manner than allowed for in equation 1.35. In general, IRK schemes are formulated as follows.

$$\begin{aligned} \xi_j =& \mathbf{y}_n + h \sum_{i=1}^{\nu} a_{j,i} \mathbf{f}(t_n + c_i h, \xi_i), \ j = 1, \ldots, \nu \\ \mathbf{y}_{n+1} =& \mathbf{y}_n + h \sum_{j=1}^{\nu} b_j \mathbf{f}(t_n + c_j h, \xi_j) \end{aligned} \tag{1.37}$$

[21] The approximation error of fourth-order Runge-Kutta decays globally as $\mathcal{O}(h^4)$: reducing the step-size by a factor of 10 reduces the global approximation error of by a factor of 10^4. Compare this result to the basic forward Euler method whose approximation error would only fall by a factor of 10 for a similar reduction in step-size. This substantial improvement in accuracy comes at the trivial computational cost of only four extra evaluations of the function $\mathbf{f}$ per step.

Note that the matrix $A = (a_{j,i})_{j,i=1,2,\ldots,\nu}$, is now an arbitrary matrix whereas in equation 1.35 it was strictly lower triangular.[22]

IRK methods have superior stability properties but require substantially more computational effort than explicit RK methods. For a general IRK matrix A the implicit RK formula defined by equation 1.37 defines a system of νn, generally non-linear, coupled algebraic equations.[23] As was the case with explicit RK methods , IRK methods are typically expressed in tableau form.

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{11} & \ldots & a_{1s} \\ c_2 & a_{21} & a_{21} & \ldots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \ldots & a_{2s} \\ \hline & b_1 & b_2 & \ldots & b_s \\ & \tilde{b}_1 & \tilde{b}_2 & \ldots & \tilde{b}_s \end{array}$$

An important subset of IRK methods are called collocation methods. Suppose that the solution has already been integrated up to the point $(t_n, \mathbf{y}_n)$ and that I now want to advance the integration to $(t_{n+1}, \mathbf{y}_{n+1})$, where $t_{n+1} = t_n + h$. Implementation of a collocation method begins with the specification of ν distinct collocation nodes $c_1, c_2, \ldots, c_\nu \in [0, 1]$. The next step is to find a degree ν polynomial that obeys the given initial conditions and satisfies the ODE exactly at the ν distinct collocation nodes.

$$\mathbf{u}(t_n) = \mathbf{y}_n, \tag{1.39}$$

$$\mathbf{u}'(t_n + c_j h) = \mathbf{f}(t_n + c_j h, \mathbf{u}(t_n + c_j h)), \ j = 1, 2, \ldots, \nu \tag{1.40}$$

A collocation method finds such a polynomial $\mathbf{u}$ and sets

$$\mathbf{y}_{n+1} = \mathbf{u}(t_{n+1}) \tag{1.41}$$

[22] By convention the rows of matrix A satisfy

$$c_j = \sum_{i=1}^{\nu} a_{ji}, \ j = 1, \ldots, \nu \tag{1.38}$$

which is necessary for the resulting method to be of non-trivial order.

[23] As a result of the significant computational cost of evaluating the RK stages of fully implicit RK schemes, several authors have put forward various "semi-implicit" RK methods. Popular methods in the class are the diagonally implicit Runge-Kutta (DIRK) and the singly-diagonally implicit Runge-Kutta (SDIRK) methods. For a detailed discussion of these methods and other semi-implicit Runge-Kutta methods, as well as practical issues associated with implementing implicit Runge-Kutta methods see sections 34-36 of Butcher (2008)

While in principle one could choose any set of collocation nodes $c_1, c_2, \ldots, c_\nu$, a common approach is to set $c_1, c_2, \ldots, c_\nu$ equal to the roots of a suitably transformed ν-degree Legendre polynomial resulting in a ν-stage IRK method of order 2ν.[24] The following are the RK tableaux for $\nu = 1, 2, 3$ and orders $p = 2, 4, 6$, respectively.

$$\nu = 1, p = 2, \quad \begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array};$$

$$\nu = 2, p = 4, \quad \begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array};$$

$$\nu = 3, p = 6, \quad \begin{array}{c|ccc} \frac{1}{2} - \frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\ \frac{1}{2} + \frac{\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}.$$

The computation of the νn non-linear algebraic systems generated by fully implicit Runge-Kutta methods can be expensive. Usually, however, the additional computational burden is more then compensated by the increased order of the resulting method. Although there are no general rules for when the law of diminishing returns forces the use of a lower order method, Iserles (2009) suggests that the three stage, order six Gauss-Legendre method is likely the largest order method that is consistent with reasonable implementation costs.

Adaptive step-size control

As was the case with linear multi-step methods, a high quality Runge-Kutta integrator should implement some form of adaptive step-size control. Recall that an explicit order p Runge-Kutta method consists of a matrix, $A = (a_{ji})_{j,i=1,\ldots,\nu}$, weights, $\mathbf{b}$, and nodes, $\mathbf{c}$ that

[24] The classic ν-degree Legendre polynomial, P_ν, is orthogonal with the weighting function $\omega(t) = 1$ on the domain $-1 < t < 1$. In order to insure that the nodes $c_1, c_2, \ldots, c_\nu \in [0, 1]$, I need to apply a linear transformation to the polynomial P_ν to create a new polynomial $\tilde{P}_\nu$ that is orthogonal with the weight function $\omega(t)$ on the domain $0 < t < 1$.

$$\tilde{P}_\nu(t) = P_\nu(2t - 1) = \frac{(\nu)!^2}{(2\nu)!} \sum_{k=0}^{\nu} (-1)^{\nu-k} \binom{\nu}{k} \binom{\nu + k}{k} t^k \tag{1.42}$$

are combined using

$$\mathbf{y}_{n+1} = \mathbf{y} + h\sum_{i=1}^{\nu} b_i \xi_i \tag{1.43}$$

where

$$\begin{aligned}
\xi_1 =& \mathbf{y}_n \\
\xi_2 =& \mathbf{y}_n + h a_{2,1}\mathbf{f}(t_n + c_1 h, \xi_1) \\
\xi_3 =& \mathbf{y}_n + h a_{3,1}\mathbf{f}(t_n + c_1 h, \xi_1) + h a_{3,2}\mathbf{f}(t_n + c_2 h, \xi_2) \\
&\vdots \\
\xi_\nu =& \mathbf{y}_n + h\sum_{i=1}^{\nu-1} a_{\nu,i}\mathbf{f}(t_n + c_i h, \xi_i).
\end{aligned}$$

An embedded Runge-Kutta (ERK) method adds an additional vector of weights $\tilde{\mathbf{b}}$ that can be used to construct an explicit order $p-1$ method

$$\tilde{\mathbf{y}}_{n+1} = \tilde{\mathbf{y}} + h\sum_{i=1}^{\nu} \tilde{b}_i \xi_i \tag{1.44}$$

where $\xi_i,\ i = 1, \ldots, \nu$ are the same stages used to construct the order p method. The difference between the two estimates of $\mathbf{y}(t_{n+1})$,

$$\mathbf{e}_{n+1} = \mathbf{y}_{n+1} - \tilde{\mathbf{y}}_{n+1} = h\sum_{i=1}^{\nu} (b_i - \tilde{b}_i)\xi_i, \tag{1.45}$$

is an estimate of the local truncation error for the order p method which can be used in conjunction with equation 1.32 to adjust the step-size.

ERK methods are generally expressed using an extended form of the Butcher tableau.[25]

$$\begin{array}{c|cccc}
c_1 & a_{11} & a_{11} & \ldots & a_{1s} \\
c_2 & a_{21} & a_{21} & \ldots & a_{2s} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
c_s & a_{s1} & a_{s2} & \ldots & a_{2s} \\
\hline
 & b_1 & b_2 & \ldots & b_s \\
 & \tilde{b}_1 & \tilde{b}_2 & \ldots & \tilde{b}_s
\end{array}$$

[25] Although the first to propose combining two RK methods of different orders into a single tableau was Merson (1957), embedded Runge-Kutta methods were widely popularized by Fehlberg (1968).

The most basic ERK method combines Heun's second-order explicit RK method with the forward Euler method.

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \\ & 1 & 0 \end{array}$$

In the above tableau, the first row of $\mathbf{b}$ coefficients gives the second-order accurate solution, and the second row has the first-order solution.

The classic example of a third-order ERK method is the Bogacki and Shampine (1989) method.

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{3}{4} & 0 & \frac{3}{4} & & \\ 1 & \frac{2}{9} & \frac{1}{3} & \frac{4}{9} & \\ \hline & \frac{2}{9} & \frac{1}{3} & \frac{4}{9} & 0 \\ & \frac{7}{24} & \frac{1}{4} & \frac{1}{3} & \frac{1}{8} \end{array}$$

In the above tableau, the first row of $\mathbf{b}$ coefficients gives the third-order accurate solution, and the second row has the second-order solution.

The most popular ERK methods are the Fehlberg (1968) method, Cash and Karp (1990) method, and the Dormand and Prince (1980) method.[26] All three of these methods are

[26]The Dormand and Prince (1980) method is the default ODE integrator in many commercial software packages including MATLAB©.

fourth-order accurate.

Fehlberg 4(5) =

0						
$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$			
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$		
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	
	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$
	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0

Cash-Karp 4(5) =

0						
$\frac{1}{5}$	$\frac{1}{5}$					
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$				
$\frac{3}{5}$	$\frac{3}{10}$	$-\frac{9}{10}$	$\frac{6}{5}$			
1	$-\frac{11}{54}$	$\frac{5}{2}$	$-\frac{70}{27}$	$-\frac{35}{27}$		
$\frac{7}{8}$	$\frac{1631}{55296}$	$\frac{175}{512}$	$\frac{575}{13824}$	$\frac{44275}{110592}$	$\frac{253}{4096}$	
	$\frac{37}{378}$	0	$\frac{250}{621}$	$\frac{125}{594}$	0	$\frac{512}{1771}$
	$\frac{2825}{27648}$	0	$\frac{18575}{48384}$	$\frac{13525}{55296}$	$\frac{277}{14336}$	$\frac{1}{4}$

Dormand-Prince 4(5) =

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$

In the above tableaus, the first row of $\mathbf{b}$ coefficients gives the fifth-order accurate solution,

and the second row has the fourth-order solution.

1.4 Solving IVPs using Python

1.4.1 Solow model

A classic example of an initial value problem in economics is the Solow (1956) model of economic growth. The ket equation of the Solow (1956) model is the equation of motion for capital per effective worker:

$$\dot{k} = sf(k(t)) - (n + g + \delta)k(t),\ k(0) = k_0 \tag{1.46}$$

where k is capital per effective worker, $f(k(t))$ is output per effective worker and is assumed to be concave, twice differentiable and satisfy the Inada conditions. The parameters $s,\ n,\ g,\ \delta$ are the savings rate, population growth rate, technology growth rate, and depreciation rate of physical capital, respectively.[27]

The Solow (1956) model with Cobb-Douglas production consists of the following autonomous, first-order non-linear ODE

$$\dot{k} = sk(t)^{\alpha} - (n + g + \delta)k(t),\ k(0) = k_0 \tag{1.47}$$

This special case of the model happens to have an analytic solution.[28]

$$k(t) = \left[\left(\frac{s}{n + g + \delta}\right)\left(1 - e^{-(n+g+\delta)(1-\alpha)t}\right) + k_0^{1-\alpha}e^{-(n+g+\delta)(1-\alpha)t}\right]^{\frac{1}{1-\alpha}} \tag{1.48}$$

The existence of a closed-form solution for this special case allows me to directly compare the true solution with various numerical approximations.

Consider the simple forward Euler method. Figure 1.4 plots the approximation errors for the forward Euler method for various fixed step-sizes ranging from $h = 1.0$ to $h = 0.001$. In agreement with theory, my results indicate that the global approximation error of the forward Euler method falls roughly linearly with h. Note that for all h the approximation errors declines monotonically over the interval of interest. This undesirable behavior is a product of the fixed step-size and the Solow (1956) model's monotonic convergence towards

[27] The parameters of the model are calibrated to the UK using a variation of the growth accounting procedure of data of Hall and Jones (1999) and data from Feenstra et al. (2013).

[28] A complete derivation of the differential equation describing the evolution of capital per effective worker for the Solow (1956) model as well as its solution for the case of Cobb-Douglas production are included in appendix 1.A.

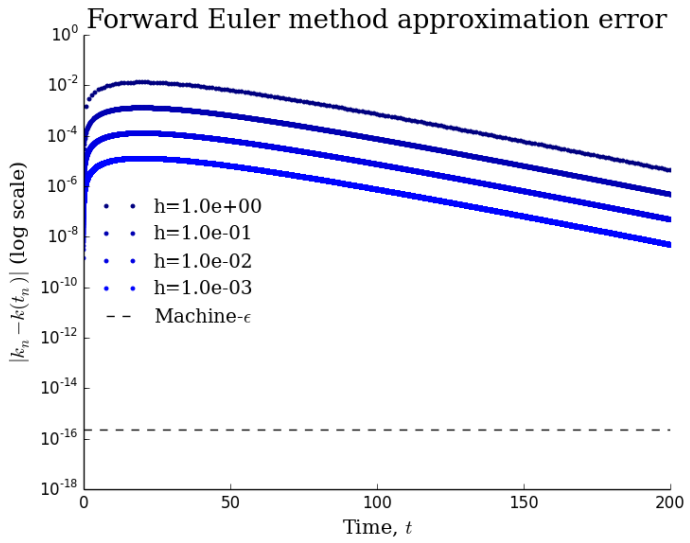

Figure 1.4: Solow (1956) approximation errors using `forward_euler`.

Table 1.1: Comparison of the first-order Euler methods with the second-order trapezoidal rule for solving the Solow (1956) model.

	`forward_euler`		`backward_euler`		`trapezoidal_rule`	
h	L^∞ error	Time (sec.)	L^∞ error	Time (sec.)	L^∞ error	Time (sec.)
1.0(0)	1.8(-2)	1.0(-2)	1.7(-2)	6.9(-2)	7.7(-5)	8.9(-2)
1.0(-1)	1.8(-3)	1.0(-1)	1.8(-3)	5.9(-1)	7.7(-7)	7.5(-1)
1.0(-2)	1.8(-4)	1.1(0)	1.8(-4)	5.9(0)	7.7(-9)	7.7(0)
1.0(-3)	1.8(-5)	5.6(1)	1.8(-5)	1.0(2)	7.7(-11)	1.2(2)

its fixed-point attractor k^*. Numerical approximation errors computed using the L^∞ norm as well as the run-times required to integrate equation 1.47 forward from an initial condition of $k_0 = \frac{1}{2}k^*$ where

$$k^* = \left(\frac{s}{n+g+\delta}\right)^{\frac{1}{1-\alpha}} \tag{1.49}$$

over the interval $0 \leq t \leq 200$ using the `forward_euler`, `backward_euler`, as well as the `trapezoidal_rule` are given in table 1.1 [29] When comparing run times times across the various methods, it is the relative (and not absolute) speed which matters. While absolute speed of any particular method will vary across computers, the relative speed of various methods should be fairly stable.

Next I approximate $k(t)$ using two high-quality, linear multi-step integrators both of which implement adaptive step-size control. Figure 1.5 plots approximation errors for `lsoda`,

[29] Estimated run times for each of the methods were computed using the IPython `%timeit` magic command. The value reported is the fastest of three successive integrations.

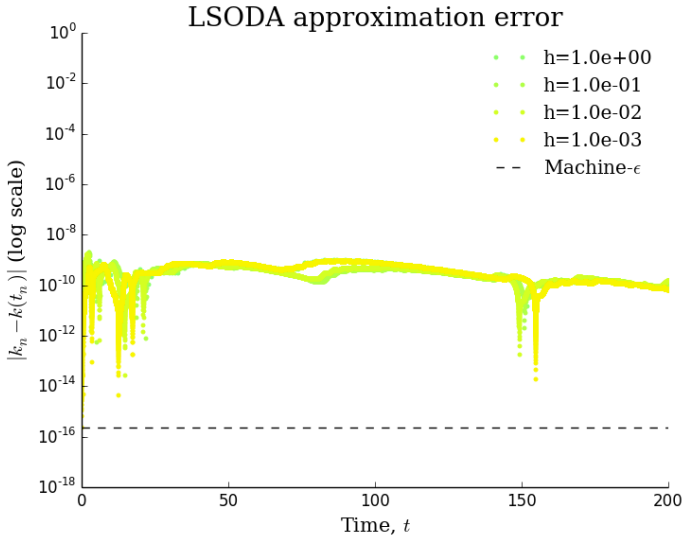

Figure 1.5: In addition to adaptive step size control, the `lsoda` integrator also adaptively switches between linear multi-step methods of different orders. The sporadic large drops in the level of the approximation error are caused by `lsoda` switching between integrators of different orders.

which is part of the Livermore Solvers for Ordinary Differential Equations (LSODE) package of ODE integrators developed by Hindmarsh and Radhakrishnan (1993). Figure 1.6 plots approximation errors for `vode`, a variable coefficient ODE solver developed by Brown et al. (1989). Both of these integrators are available via the `scipy.integrate.ode` module and their control parameters have been tuned in order to insure that the local truncation error stays in the neighborhood of $1(-9)$. L^∞ errors and run-times for `lsoda`, and two different implementations of the `vode` solver are given in table 1.2. These results clearly demonstrate the importance of using high-quality integrators with adaptive step size control. Both the `lsoda` and `vode` integrators achieve a significantly higher level of accuracy compared with the simple Euler methods without sacrificing computational efficiency.

Table 1.2: Comparison of various linear multi-step methods with adaptive step-size control for solving the Solow (1956) model.

	`lsoda`		`vode` (Adams-Moulton)		`vode` (BDF)	
h	L^∞ error	Time (sec.)	L^2 error	Time (sec.)	L^2 error	Time (sec.)
1.0	1.4(-8)	1.3(-2)	2.9(-9)	1.3(-2)	9.5(-8)	1.2(-2)
1.0(-1)	3.4(-9)	1.0(-1)	2.9(-9)	1.1(-1)	9.6(-8)	1.2(-1)
1.0(-2)	2.4(-9)	1.1(0)	6.7(-9)	1.2(0)	5.8(-8)	1.2(0)
1.0(-3)	1.8(-9)	5.5(1)	6.6(-9)	5.6(1)	7.1(-8)	5.5(1)

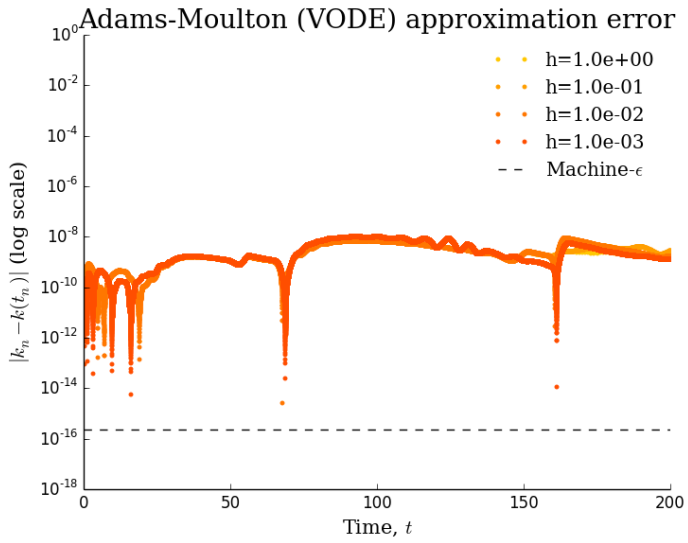

Figure 1.6: Like `lsoda`, the `vode` integrator also adaptively switches between linear multi-step methods of different orders. The sporadic large drops in the level of the approximation error when using the `vode` integrator are caused by such switches.

1.4.2 The Spence (1974) model

The education signaling model of Spence (1974) is another classic example of an IVP problem in economics. The Spence (1974) model assumes that worker differ in their ability, and that a worker's individual ability $n \in [n_L, n_H]$, is private information thus can not be observed by any other economic agent. Workers acquire y years of education at a total cost of $C(y, n)$ where $C_y > 0 > C_n$. After acquiring some level of education, workers are offered a job. A worker of ability n and y years of education produces $S(y, n)$ output where $S_n > 0$ and $S_y > 0$. The critical assumption of the Spence (1974) model is that a prospective employer can not directly observe a worker's ability, but instead must infer it by observing a worker's level of education. Spence showed that the equilibrium function $n(y)$ satisfies the following first-order non-linear differential equation.

$$n'(y) = \frac{C_y(y, n(y)) - S_y(y, n(y))}{S_n(y, n(y))} \tag{1.50}$$

Given n_L and assuming that the lowest ability workers obtain the socially optimal level of education, then the y_L that solves

$$S_y(y_L, n_L) = C_y(y_L, n_L) \tag{1.51}$$

pins down the initial condition, $n(y_L) = n_L$.

In the special case of the above model analyzed in Spence (1974) $C(y,n) = \frac{y}{n}$ and $S(y,n) = ny^{\alpha}$ with $0 < \alpha < 1$. Under these restrictions, the above differential equation reduces to a non-autonomous, first-order, non-linear ODE

$$n'(y) = \frac{n(y)^{-1} - \alpha n(y) y^{\alpha-1}}{y^{\alpha}}, \; n(y_L) = n_L \tag{1.52}$$

where $y_L = (\alpha n^2)^{\frac{1}{1-\alpha}}$ is the socially optimal level of education for the lowest ability workers. This special case also happens to have a closed-form solution for $n(y)$.

$$n(y) = y^{-\alpha} \left(\frac{2(y^{1+\alpha} + \left[\left(\frac{1+\alpha}{2}\right) \left(\frac{n_L}{y_L^{-\alpha}}\right)^2 - y_L^{1+\alpha} \right]}{1+\alpha} \right)^{\frac{1}{2}} \tag{1.53}$$

The existence of a closed-form solution for this special case allows us to directly compare the true solution with various numerical approximations.

Consider the trapezoidal rule. Given that the trapezoidal rule is an order two method, reducing the step-size, h by a factor of ten should reduce the approximation error by a factor of 10^2. In practice, I find that the approximation error of the trapezoidal rule does not decay quite as fast. Figure 1.7 plots the absolute errors for various h. L^{∞} errors and run-times for the `forward_euler`, `backward_euler`, and the `trapezoidal_rule` are given in table 1.3 for a model with $\alpha = 0.25$ and $n_L = 0.1$ (which implies $y_L = 0.00034$).

Table 1.3: Comparison of the first-order Euler methods with the second-order trapezoidal rule for solving the Spence (1974) model.

	`forward_euler`		`backward_euler`		`trapezoidal_rule`	
h	L^{∞} error	Time (sec.)	L^{∞} error	Time (sec.)	L^{∞} error	Time (sec.)
1.0(-1)	1.2(0)	1.0(-2)	1.2(-1)	5.2(-2)	2.1(-1)	7.0(-2)
1.0(-2)	1.3(-1)	1.0(-1)	2.6(-2)	5.0(-1)	5.9(-2)	6.5(-1)
1.0(-3)	1.9(-2)	1.1(0)	2.6(-3)	4.8(0)	6.6(-3)	6.2(0)
1.0(-4)	9.3(-4)	5.3(1)	6.9(-4)	9.0(1)	9.4(-5)	1.0(2)

Figures 1.8 and 1.9 plot the approximation errors for two widely used embedded Runge-Kutta methods developed by Dormand and Prince (1980). L^{∞} errors and average run-times for a fourth-order explicit RK method, `rk4`, as well as the two embedded RK methods, `dopri5`, and `dop853` are given in table 1.4. The performance gains from using embedded Runge-Kutta methods with adaptive step size control are impression. Although embedded Runge-Kutta methods are about as fast as the classic Euler methods, they are roughly 10

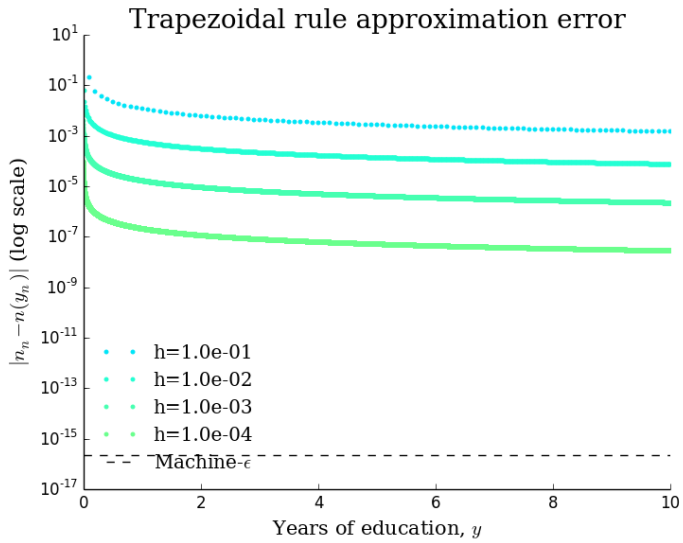

Figure 1.7: The performance of the trapezoidal rule is disappointing: the approximation error decays much more slowly than theory would predict.

orders or magnitude more accurate.

Table 1.4: Comparison of the fourth-order Runge-Kutta method with embedded Runge-Kutta methods with adaptive step-size control for solving the Spence (1974) model.

	`rk4`		`dopri5`		`dop853`	
h	L^∞ error	Time (sec.)	L^∞ error	Time (sec.)	L^∞ error	Time (sec.)
1(-1)	3.1(-2)	7.3(-2)	8.3(-11)	3.2(-2)	9.0(-13)	4.8(-2)
1(-2)	2.3(-2)	7.2(-1)	6.0(-11)	2.8(-1)	6.9(-12)	4.5(-1)
1(-3)	7.4(-5)	7.4(0)	3.4(-10)	2.9(0)	8.7(-13)	4.6(0)
1(-4)	1.0(-7)	1.2(2)	2.4(-11)	7.1(1)	1.9(-12)	8.7(1)

1.5 Finite-difference methods for BVPs

Initial value problems are typically straightforward to solve because each point of the solution depends on only local conditions which allows for the use of local approximation methods discussed in section 1.3. In contrast, boundary value problems impose restrictions on the solution at multiple points and thus the solution at each point no longer depends only on local conditions. The need to resort to the more sophisticated global approximation schemes makes solving BVPs inherently more challenging than solving IVPs. In this section I discuss "shooting" methods for solving two-point boundary value problems (2PBVPs)

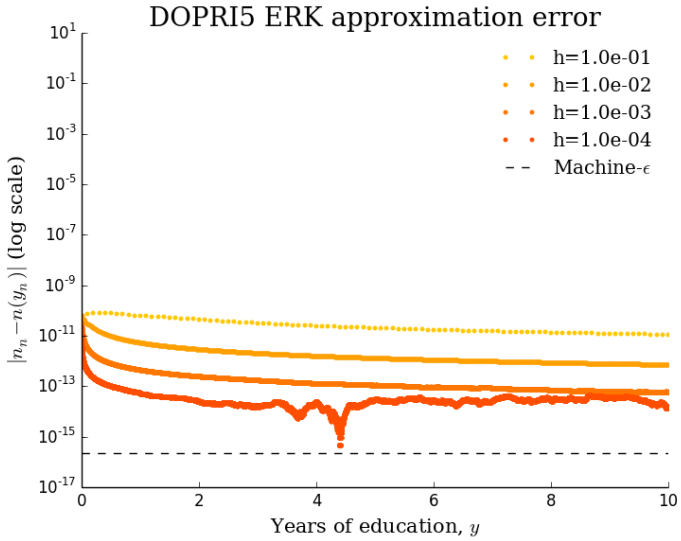

Figure 1.8: Approximation errors for various step-sizes when solving the Spence (1974) model using a fourth-order embedded Runge-Kutta method due to Dormand and Prince (1980).

commonly encountered in the economics literature.[30]

Consider the following two-point boundary value problem.

$$\begin{aligned} \mathbf{y}' &= \mathbf{f}(t, \mathbf{y}), \ t \geq t_0 \\ g_i(y_i(t_0)) &= 0, i = 1, \ldots, n' \\ g_i(y_i(T)) &= 0, i = n' + 1, \ldots, n \end{aligned} \tag{1.54}$$

where $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$. Note that the auxiliary conditions of the boundary value problem defined in 1.54 provide initial conditions for n' components of the solution and terminal conditions for the remaining $n - n'$ components.

The idea behind shooting methods is to turn the boundary value problem into an initial value problem by guessing appropriate initial conditions for the remaining $n - n'$ components of the solution and then, using an appropriate initial value method, integrating the system forward in order to see what this guess implies about the value of the $\mathbf{y}(T)$ for the $n - n'$ components of the solution. If the conjectured $n - n'$ initial conditions lead to a value of $\mathbf{y}(T)$ that is sufficiently close to satisfying the given terminal conditions for the $n - n'$ components of the solution then procedure terminates. Otherwise a new guess for appropriate initial conditions is generated and the above process repeats until a value for

[30] An excellent discussion of simple shooting methods can be found in Chapter 10 of Judd (1998).

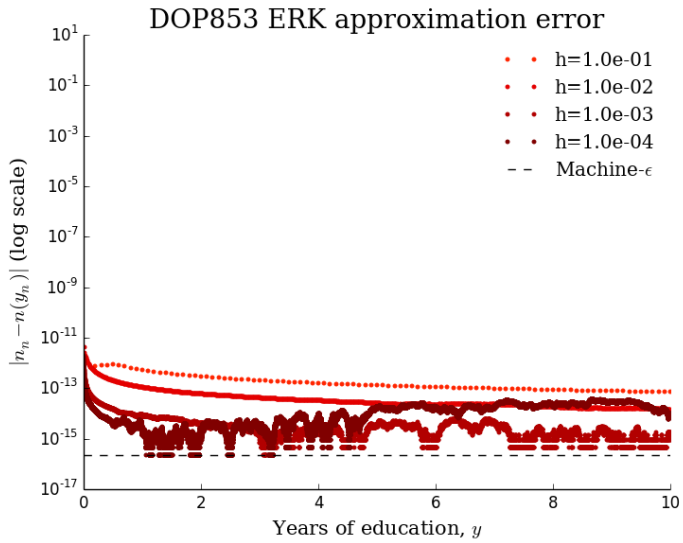

Figure 1.9: Approximation errors for various step-sizes when solving the Spence (1974) model using an eighth-order embedded Runge-Kutta method due to Dormand and Prince (1980).

$\mathbf{y}(t_0)$ is found that is roughly consistent with the given terminal conditions.

This heuristic description suggests a shooting algorithm is comprised of two basic pieces. The first piece is a method for solving the initial value problem

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \ t \geq t_0, \ \mathbf{y}(t_0) = \mathbf{y}_0^j \tag{1.55}$$

in order to compute the implied value for $\mathbf{y}_T$ for some conjectured initial condition $\mathbf{y}(t_0) = \mathbf{y}_0^j$. In order to make explicit the dependence of $\mathbf{y}_T$ on the initial guess $\mathbf{y}_0^j$, let $\mathbf{y}_T(\mathbf{y}_0^j)$ denote the value of the solution at time T given a guess of $\mathbf{y}_0^j$ as the initial condition. The second piece is a method for finding a value $\mathbf{y}_0$ that is roughly consistent with the $n' - n'$ terminal conditions $g_i(y_i(T)) = 0, i = n'+1, \ldots, n$. Specifically, this requires finding values $y_{i,0}$ such that

$$g_i(y_{i,T}(y_{i,0})) = 0, i = n' + 1, \ldots, n.$$

Since this is a system of generally non-linear equation in the $n' - n$ unknowns $y_{i,0}$, the second piece of any shooting algorithm for solving a two-point BVP is a method for solving systems of non-linear equations. These ideas are summarized in algorithm 1

The generic shooting algorithm described above is an example of a two-layer algorithm. The inner layer, defined on line 4, uses an appropriate method approximating the solution of an IVP given an initial conditions $\mathbf{y}_0^j$. The accuracy of this inner layer will depend

Algorithm 1: Simple shooting algorithm

1 Objective: Solve the two-point BVP defined by 1.54.
2 Guess some $\mathbf{y}_0^j$ and specify a stopping tolerance, $tol > 0$.
3 **while** *True* **:**
4 Solve the IVP defined by 1.55 for $\mathbf{y}_T(\mathbf{y}_0^j)$.
5 **if** $||g_i(y_{i,T}(y_{i,0}^j)) - y_{i,T}|| < tol,\ i = n' + 1, \ldots, n$ **:**
6 break
7 **else:**
8 Choose $\mathbf{y}_0^{j+1}$ based on $\mathbf{y}_0^j,\ \mathbf{y}_0^{j-1}$, etc.

on the numerical method used as well as the choice of step-size, h. The outer layer, represented by the while loop, updates $\mathbf{y}_0^j$ in order to solve the system of non-linear equation $g_i(y_{i,T}(y_{i,0})) = 0, i = n' + 1, \ldots, n$. Any method for solving non-linear equations can be used in line 7 to choose the next iterate $\mathbf{y}_0^{j+1}$ based on previous values $\mathbf{y}_0^j$ and/or derivatives of $\mathbf{y}_T(\mathbf{y}_0)$.[31]

It is important to remember that, in general, the approximation error for multi-layered algorithms is determined by the interaction between the approximation errors of the individual layers. For shooting methods, in particular, depending on the speed with which error accumulates in the inner layer it may be necessary to set a relatively loose error tolerance in the outer layer in order for the algorithm to terminate.

1.6 Solving BVPs using Python

1.6.1 The optimal growth model

In this section I compare the relative speed and accuracy of three different shooting methods for solving a version of the optimal growth model of Ramsey (1928), Cass (1965), and Koopmans (1965) with constant relative risk aversion (CRRA) preferences and a Cobb-

[31] The literature on numerical methods for solving non-linear equations is vast. Chapter 5 of Judd (1998) discusses some of the classic techniques.

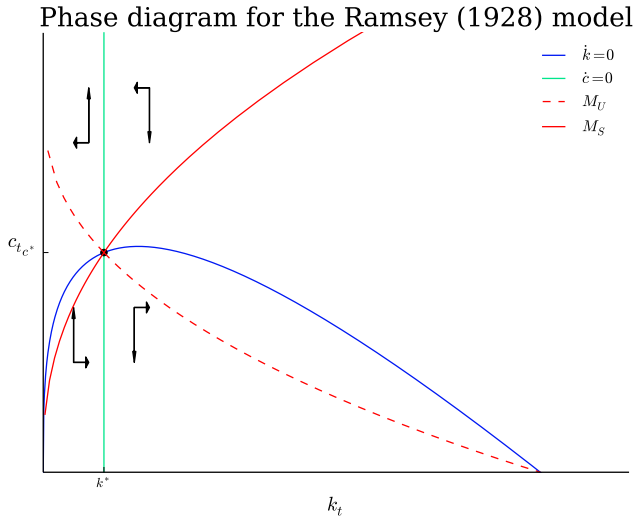

Figure 1.10: Since the optimal growth model is saddle point stable, "solving" the model is really just computing the stable manifold (i.e., saddle path).

Douglas production technology.[32]

$$\begin{aligned}\dot{k} =& k(t)^{\alpha} - (n+g+\delta)k(t) - c(t)\\ \frac{\dot{c}}{c(t)} =& \frac{\alpha k(t)^{\alpha-1} - \delta - \rho - \theta g}{\theta}\\ k(0) =& k_0,\ 0 < \lim_{t\to\infty} |k(t)| < \infty\end{aligned} \tag{1.56}$$

Although there is no general, analytic solution for the optimal growth model with CRRA utility and Cobb-Douglas production, it is possible to obtain an analytic solution under the assumption that the discount rate, ρ satisfies

$$\rho = \alpha\theta(n+g+\delta) - (\delta+\theta g) > 0. \tag{1.57}$$

This assumption implies that the model exhibits a constant gross savings rate which, in

[32] See appendix 1.B for a complete description of the model including a derivation of this system of equations and boundary conditions.

turn, leads to the following analytic solution for $k(t)$, $c(t)$.[33]

$$k(t) = \left[k_0^{1-\alpha} e^{-\lambda t} + \left(\frac{1}{\theta(n+g+\delta)}\right)\left(1 - e^{-\lambda t}\right)\right]^{\frac{1}{1-\alpha}} \tag{1.58}$$

$$c(t) = \left(\frac{\theta - 1}{\theta}\right) k(t)^{\alpha} \tag{1.59}$$

The existence of a closed-form solution for this, admittedly rather specific, parameter restriction allows me to directly compare the shooting approximations with the exact solution. In what follows I assume the following parameter values: $\alpha = 0.33$, $\delta = 0.04$, $\theta = 3.0$, $g = 0.01$, $n = 0.025$ which together imply that $\rho = 0.00425$.

Forward shooting

Figure 1.10 is a phase diagram for this system of differential equations. Because the model is saddle point stable there exist two invariant manifolds: a stable manifold, M_S and an unstable manifold, M_U. These manifolds are called invariant because any path that begins on $M_S(M_U)$ will remain on $M_S(M_U)$. M_S is called the stable manifold because any path that begins on the stable manifold will eventually converge to the steady state; M_U is called the unstable manifold because any path the begins on M_U will diverge away from the steady state. In order to solve the optimal growth model I need to compute its stable manifold, M_S.

The forward shooting method for finding M_S begins by guessing a feasible value for c_0 and then using some IVP scheme to generate the implied solution trajectories for the variables $k(t)$ and $c(t)$.[34] If the initial choice of c_0 is too small, then the solution trajectory eventually crosses the $\dot{c} = 0$ locus and $c(t)$ begins to fall. Similarly, if the choice of c_0 is too large, then our path eventually crosses the $\dot{k} = 0$ curve at which point $k(t)$ will start to fall. These observations motivate the forward shooting scheme described in algorithm 2 for an initial condition $k(0) = k_0 < k^*$.[35]

Although algorithm 2 uses bisection search to update the guesses for the initial condition, c_0 in order to solve the non-linear equation $c^* = c_T(c_0)$, this is not strictly necessary. Any non-linear equation solver could, in principle, be used.[36] Figure 1.11 gives a sense of

[33] See appendix 1.B for a complete derivation of this solution.

[34] The choice of consumption per effective worker must be non-negative and be less than the sum total of output and un-depreciated capital per effective worker. This implies the feasible range for the choice of c_0 is the interval $[0, k_0^{\alpha} - (n+g+\delta)k_0]$.

[35] The algorithm for solving the case where $k(0) = k_0 > k^*$ is almost identical.

[36] Bisection search is not the most efficient technique for solving non-linear equations. However, so long as I correctly bracket the true initial condition in line 1 of algorithm 2, bisection search is guaranteed to converge to a solution. For a more detailed discussion of the bisection search method, as well as several

Algorithm 2: Forward shooting the optimal growth model

Bracket the true $c(0)$ by setting $c_L = 0$ and $c_H = (1-\delta)k_0 + k_0^{\alpha}$.
Guess that $c_0 = \frac{1}{2}(c_H + c_L)$ and specify a stopping tolerance, $tol > 0$.
while *True* **:**
    Solve the model as an IVP with $k(0) = k_0$ and $c(0) = c_0$
    if $\dot{c} < 0$ **:**
        if $|c(T) - c^*| < tol$ **:**
            break
        else:
            $c_L = c_0$
            $c_0 = \frac{1}{2}(c_H + c_L)$
    elif $\dot{k} < 0$ **:**
        if $|c(T) - c^*| < tol$ **:**
            break
        else:
            $c_H = c_0$
            $c_0 = \frac{1}{2}(c_H + c_L)$
    else:
        continue

how the update process using bisection search converges to an initial condition, c_0^*, that approximates the stable manifold, M_S.

Approximation errors and estimated run-times for computing M_S for $k_0 = \frac{1}{2}k^*$ using algorithm 2 with the `dopri5` integrator are given in table 1.5 for various step-sizes and convergence tolerances. When approximating the value of the policy function between grid points, I use third-order B-spline interpolation. Note that, for a given convergence tolerance, decreasing the step size actually increases the approximation error! When using an integrator that implements some form of adaptive step size control, such as `lsoda`, `vode`, or `dopri5`, the value of *tol* is likely the major determinant of the overall approximation error for the forward shooting algorithm. In a sense, decreasing h only results in better approximation of a trajectory that differs from M_S by *tol*, as opposed to a better approximation of M_S itself. Forward shooting is also numerically unstable: for sufficiently tight convergence tolerance, the algorithm fails to converge when $h = 1(-2)$ and $h(1-3)$. The results reported in table 1.5 suggest that a good strategy for accurately approximating the M_S using forward shooting is to choose a tight convergence tolerance and a relatively large step size.

other techniques for solving non-linear equations widely used in the economics literature, see Chapter 5 of Judd (1998).

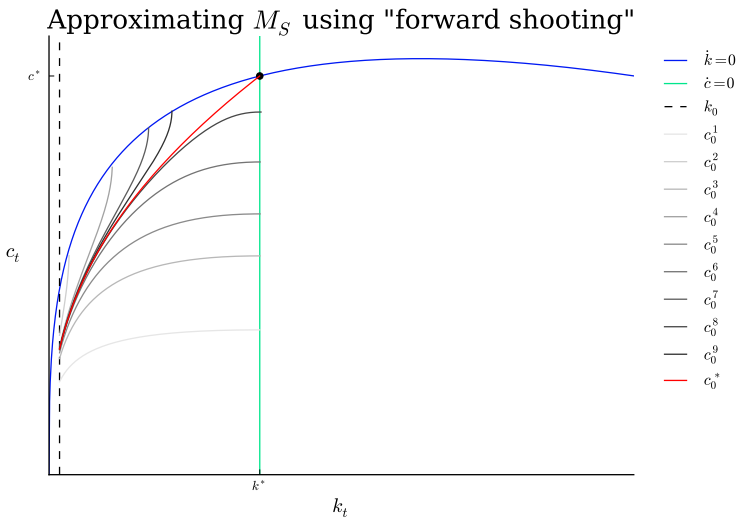

Figure 1.11: Forward shooting with bisection search to find an initial condition c_0^* consistent with the stable manifold, M_S.

Table 1.5: Forward shooting approximation errors and run times.

	h = 1(0)		h = 1(-1)		h = 1(-2)	
tol	L^2 error	Time (sec.)	L^2 error	Time (sec.)	L^2 error	Time (sec.)
1(-1)	1.22(0)	4.3(-2)	3.26(0)	2.1(-1)	1.0(1)	2.1(0)
1(-2)	4.4(-1)	1.2(-1)	1.3(0)	7.8(-1)	4.2(0)	7.7(0)
1(-3)	6.8(-1)	1.3(-1)	2.0(0)	8.0(-1)	6.4(0)	8.2(0)
1(-4)	6.8(-1)	1.3(-1)	1.3(0)	2.6(0)	4.2(0)	2.7(1)
1(-5)	8.5(-3)	2.4(1)	-	-	-	-
1(-6)	1.3(-4)	1.1(2)	-	-	-	-
1(-7)	1.6(-4)	1.1(2)	-	-	-	-
1(-8)	6.0(-6)	3.0(2)	-	-	-	-
1(-9)	5.3(-7)	4.0(2)	-	-	-	-

1.6.2 Reverse shooting

The phase diagram in figure 1.10 suggests that using forward shooting to compute the stable manifold, M_S, of the model might be difficult for at least some parameter values and initial conditions $k(0)$. Any initial deviation from M_S, however small, is magnified over time resulting in a path that increasingly departs from the exact solution. However, suppose that instead of computing the stable manifold, I wished instead to compute the unstable manifold, M_U. As figure 1.10 suggests deviations from M_U, however large, become smaller over time. In fact, all I would need to do in order to compute a path that lies on the unstable manifold is to choose a point near the steady state and then integrate the system. The basic idea behind reverse shooting is to transform the system of equations

in such a way so that the stable manifold of the original system becomes the unstable manifold of the transformed system and then solve for the unstable manifold by shooting "backwards" from the steady state.

Since time plays no direct role in the model, the household's choice of consumption per effective worker at time t depends only on the value of capital per effective worker at time t. I can express this by defining a policy function, $c(k)$, such that $c(k) = c(k(t)) = c(t)$. Furthermore, 1.56 implies that the consumption policy function satisfies the following differential equation.

$$c'(k) = \frac{\dot{c}}{\dot{k}} = \left(\frac{c(k)}{\theta}\right)\left(\frac{\alpha k^{\alpha-1} - \delta - \rho - \theta g}{k^{\alpha} - (n+g+\delta)k - c(k)}\right) \tag{1.60}$$

Since optimality requires the economy to converge to its steady state, the solution $c(k)$ must satisfy the boundary condition $c(k^*) = c^*$. The reverse shooting approach solves for the lower portion of the consumption policy function by choosing some initial step-size $\epsilon > 0$, setting

$$\begin{aligned} k_0 =& k^* - \epsilon \\ c_0 =& c(k^* - \epsilon) \approx c^* - \epsilon c'(k^*) \end{aligned}$$

and then integrating equation 1.60 backward using some IVP scheme. To compute the upper portion of the policy function using reverse shooting, simply integrate equation 1.60 forward from the initial condition

$$\begin{aligned} k_0 =& k^* + \epsilon \\ c_0 =& c(k^* + \epsilon) \approx c^* + \epsilon c'(k^*). \end{aligned}$$

Table 1.6 displays the L^2 approximation errors for various choices of the initial step-size, ϵ, and the regular step-size, h, when using the `dopri5` integrator from Dormand and Prince (1980) to solve the IVPs on the interval $[\frac{1}{2}k^*, 2k^*]$. To approximate the value of the policy function between grid points I use third-order B-spline interpolation. Reverse shooting is both faster, more accurate, and more numerically stable than forward shooting. Also, unlike forward shooting, reverse shooting has the desirable property that the approximation error is declining with the step size h.

Table 1.6: Reverse shooting approximation errors and run-times.

	h = 1(-1)		h = 1(-2)		h = 1(-3)	
eps	L^2 error	Time (sec.)	L^2 error	Time (sec.)	L^2 error	Time (sec.)
1(-1)	5.8(-5)	3.9(-2)	7.9(-5)	3.6(-1)	8.3(-5)	3.6(0)
1(-2)	8.3(-8)	3.8(-2)	1.8(-7)	3.6(-1)	2.4(-7)	3.7(0)
1(-3)	7.0(-9)	3.9(-2)	2.9(-10)	3.6(-1)	2.5(-10)	3.9(0)
1(-4)	7.1(-9)	3.9(-2)	2.0(-10)	3.6(-1)	2.6(-13)	3.8(0)
1(-5)	7.1(-9)	4.3(-2)	2.5(-10)	3.7(-1)	8.8(-14)	4.0(0)
1(-6)	7.1(-9)	5.3(-2)	2.5(-10)	3.9(-1)	8.1(-14)	4.2(0)

1.7 Conclusions

My results suggest a number of "best practices" that all economic researchers should adhere to when solving ODEs. Classic finite difference methods with fixed step-size such as variants of Euler's method or the family of explicit Runge-Kutta methods should be avoided. While such methods are easy to code, they tend to be computationally inefficient and will be orders of magnitude less accurate than more modern methods which implement adaptive step-size control. There are several high-quality ODE solvers currently available via the `scipy.optimize` module: the embedded Runge-Kutta methods due to Dormand and Prince (1980), `dopri5` and `dop853` and the linear multi-step integrators `lsoda`, which is part of the Livermore Solvers for Ordinary Differential Equations (LSODE) package of ODE integrators developed by Hindmarsh and Radhakrishnan (1993), and `vode`, a variable coefficient ODE solver developed by Brown et al. (1989). Of these methods `dopri5` and `dop853` are generally the most accurate "out of the box" (i.e., without changing any of the default tolerances).

When solving BVPs using finite-difference methods it is important to remember that the approximation error for multi-layered algorithms is determined by the interaction between the approximation errors of the individual layers. For shooting methods it may be necessary to set a relatively loose error tolerance in the outer layer in order for the algorithm to terminate. Using an ODE solver with adaptive step-size control in the inner layer will slow down the rate at which error accumulates in the inner layer of the algorithm, which in turn, will allow the researcher to set a tighter error tolerance in the outer layer. Where applicable, reverse shooting is preferred over forward shooting for solving BVPs. Reverse shooting is more computationally efficient, more numerically stable, and significantly more accurate than forward shooting.

This chapter explored the use of the Python programming language for solving types of ordinary differential equations (ODEs) commonly encountered in economics using finite-difference methods. The methods surveyed in this chapter, as well as the accompanying Python code and IPython notebooks which implement them should be of use to any economist interested in applying finite-difference methods for solving ODEs to economic problems.

Appendix

1.A The Solow growth model

The Solow (1956) model focuses on four variables: output, Y, capital, K, labor L, and technology or "effectiveness of labor", A. Capital, labor, and technology are combined to produce output according to some production function F.

$$Y(t) = F(K(t), A(t)L(t)) \tag{1.61}$$

F is assumed to exhibit constant returns to scale which allows us to work with the intensive form of the production function

$$y(t) = \frac{Y(t)}{A(t)L(t)} = F\left(\frac{K(t)}{A(t)L(t)}, 1\right) = f(t, k(t)) \tag{1.62}$$

The initial values of capital, K_0, labor, L_0, and technology, A_0, are taken as given and assumed to be strictly positive. Labor and technology are assumed to grow at constant rates:

$$\dot{L}(t) = nL(t), \tag{1.63}$$
$$\dot{A}(t) = gA(t), \tag{1.64}$$

where n and g are exogenous parameters representing the growth rates of the labor force and technology, respectively.

Output is divided between consumption and investment. The fraction of output invested at each point in time, s, is exogenous and constant. One unit of output devoted to investment yields one unit of new capital. Existing capital is assumed to depreciate at a constant rate δ. Putting all of this together, the aggregate capital stock evolves according to

$$\dot{K}(t) = sY(t) - \delta K(t). \tag{1.65}$$

No restrictions are placed on the parameters n, g, and δ other than the assumption that their sum is positive.

The key equation of the Solow (1956) model is the equation of motion for capital per effective worker, $k(t) = \frac{K(t)}{A(t)L(t)}$. Application of the chain rule to $k(t)$ yields

$$\begin{aligned}\dot{k}(t) =& \frac{\dot{K}A(t)L(t) - K(t)\left[A(t)\dot{L}(t) + \dot{A}(t)L(t)\right]}{[A(t)L(t)]^2} \\ =& \frac{\dot{K}}{A(t)L(t)} - \frac{K(t)}{A(t)L(t)}\left(\frac{\dot{L}(t)}{L(t)} + \frac{\dot{A}(t)}{A(t)}\right).\end{aligned} \tag{1.66}$$

From here we need only substitute the equation of motion for capital and the expressions for the exogenous growth rates of labor and technology to obtain a first-order non-linear differential equation describing the evolution of capital per effective worker, $k(t)$.

$$\begin{aligned}\dot{k}(t) =& \frac{sY(t) - \delta K(t)}{A(t)L(t)} - \frac{K(t)}{A(t)L(t)}\left(\frac{\dot{L}(t)}{L(t)} + \frac{\dot{A}(t)}{A(t)}\right) \\ =& s\frac{Y(t)}{A(t)L(t)} - (n + g + \delta)\frac{K(t)}{A(t)L(t)} \\ =& sf(k(t)) - (n + g + \delta)k(t)\end{aligned} \tag{1.67}$$

I assume throughout this chapter that the production function F is a constant returns to scale Cobb-Douglas technology:

$$F(K(t), A(t)L(t)) = K(t)^{\alpha}[A(t)L(t)]^{1-\alpha} \tag{1.68}$$

where α is capital's share of income/output. Under this assumption, the intensive form of the production function is simply $f(k(t)) = k(t)^{\alpha}$, and equation 1.67 is reduces to

$$\dot{k} = sk(t)^{\alpha} - (n + g + \delta)k(t) \tag{1.69}$$

for some given initial condition $k(0) = k_0$.

1.A.1 Analytic solution

The Solow (1956) with Cobb-Douglas production is known to have a general analytic solution. The solution method presented here follows Chiang and Wainwright (2005) and is intentionally pedestrian. The basic idea is to use a clever change of variables to transform equation 1.69 into a linear, first-order differential equation which can be solved using

standard methods. Start by defining a new variable, $z(t)$, as follows.[37]

$$z(t) = \frac{k(t)}{y(t)} = k(t)^{1-\alpha} \tag{1.70}$$

Next, differentiate equation 1.70 with respect to t to obtain the following relationship between $\dot{z}$ and $\dot{k}$

$$\dot{z} = (1-\alpha)k(t)^{-\alpha}\dot{k} \implies \dot{k} = \dot{z}(1-\alpha)^{-1}k(t)^{\alpha} \tag{1.71}$$

which can be used to substitute for $\dot{k}$ in equation 1.69 in order to yield the following linear, first-order differential equation

$$\dot{z} + (n+g+\delta)(1-\alpha)z(t) = s(1-\alpha) \tag{1.72}$$

with $z(0) = k_0^{1-\alpha}$.

The solution to equation 1.72, which is a non-homogenous, first-order linear differential equation with constant coefficient and constant term, will consist of the sum of two terms called the complementary function, z_c and the particular integral, z_p, both of which have significant economic interpretation.

Mathematically, the complementary function, z_c, is simply the general solution of the following reduced form, homogenous version of equation 1.72.

$$\dot{z} + (n+g+\delta)(1-\alpha)z(t) = 0 \tag{1.73}$$

Standard techniques for solving homogenous, first-order linear differential equations demonstrate that the general solution of equation 1.73 must be of the form

$$z_c = Ce^{-(n+g+\delta)(1-\alpha)t} \tag{1.74}$$

where C is some, as yet unknown, constant.

The particular integral, z_p, is any particular solution of 1.72. Suppose that $z(t)$ is some constant function. In this case $\dot{z} = 0$ and equation 1.72 becomes

$$z_p = \frac{s}{n+g+\delta} \tag{1.75}$$

which is a valid solution so long as $n + g + \delta \neq 0$.

The sum of the complementary function and the particular integral constitutes the general

[37] This clever change of variables was originally published in Sato (1963).

solution to equation 1.72.

$$z(t) = z_c + z_p = Ce^{-(n+g+\delta)(1-\alpha)t} + \left(\frac{s}{n+g+\delta}\right) \tag{1.76}$$

Using the initial condition, $z(0) = k_0^{1-\alpha}$, to solve for the constant C yields

$$C = k_0^{1-\alpha} - \left(\frac{s}{n+g+\delta}\right) \tag{1.77}$$

which can be combined with equation 1.76 to give the closed for solution for the capital-output ratio, $z(t)$.

$$z(t) = \left(\frac{s}{n+g+\delta}\right)\left(1 - e^{-(n+g+\delta)(1-\alpha)t}\right) + k_0^{1-\alpha}e^{-(n+g+\delta)(1-\alpha)t} \tag{1.78}$$

At this point it is worth digressing slightly to discuss the economic interpretation of the complementary function and the particular integral. The particular integral, z_p, is the inter-temporal equilibrium value for the capital-output ratio, $z(t)$, whilst the complementary function, z_c, represents deviations from this long-run equilibrium. Dynamic stability of $z(t)$ requires that deviations from equilibrium described by z_c die out as $t \to \infty$. In order for $\lim_{t\to\infty} z_c = 0$, I require that $(n+g+\delta)(1-\alpha) > 0$.

Finally, from equation 1.78 it is straightforward to obtain a closed form solution for the time path of $k(t)$ by substituting $z(t) = k(t)^{1-\alpha}$ and then solving for $k(t)$.

$$k(t) = \left[\left(\frac{s}{n+g+\delta}\right)\left(1 - e^{-(n+g+\delta)(1-\alpha)t}\right) + k_0^{1-\alpha}e^{-(n+g+\delta)(1-\alpha)t}\right]^{\frac{1}{1-\alpha}} \tag{1.79}$$

1.B The optimal growth model

The classic optimal growth model due to Ramsey (1928), Cass (1965), and Koopmans (1965) model.

1.B.1 Assumptions

Firms

As in the Solow (1956) model, there are a large number of identical firms each having access to the same constant returns to scale (CRTS) production technology, F, which combines capital, $K(t)$, labor, $L(t)$, and technology, $A(t)$, to produce output, $Y(t)$. Firms hire workers and rent capital in competitive factor markets, and sell their output in a competitive output market. Firms take the path of technology $A(t)$ as given; as in the Solow (1956) model $A(t)$ grows exogenously at rate g. Firms are owned by households and are assumed to maximize profits.

Households

There are a large number of identical, infinitely-lived households. The size of each household grows at rate n. Each member of the household is assumed to supply one unit of labor at every point in time (i.e., labor supply is inelastic). In addition, each household rents whatever capital it owns to firms. Each household has initial capital holdings of $\frac{K(0)}{H}$, where $K(0)$ is the initial amount of total capital in the economy and H is the number of households. $K(0)$ is assumed to be strictly positive. At each point in time a household divides its income between consumption and saving in order to maximize its lifetime utility, U.

$$U = \int_0^\infty e^{-\rho t} u(C(t)) \frac{L(t)}{H} dt \tag{1.80}$$

$C(t)$ is the consumption of each member of the household at time t. The function u is the instantaneous utility function giving each household member's utility at time t. $L(t)$ is the total number of workers in the economy at time t; therefore $\frac{L(t)}{H}$ is the total number of workers per household. Thus

$$u(C(t)) \frac{L(t)}{H} \tag{1.81}$$

represents a household's total instantaneous utility at time t. Finally, $\rho > 0$, is the discount rate. The greater is ρ the less a household values future consumption relative to current

consumption. The instantaneous utility function, u, is assumed to take the constant relative risk aversion (CRRA) form.

$$u(C(t)) = \frac{C(t)^{1-\theta}}{1-\theta},\ \theta > 0 \tag{1.82}$$

Since there is no uncertainty in the model, a household's attitudes toward risk are not directly relevant. However, CRRA utility implies that a household's elasticity of substitution of consumption between different points in time is $\frac{1}{\theta}$. Thus θ also determines a household's willingness to shift consumption between different points in time. For θ close to zero instantaneous utility is almost linear and a household would be willing to shift large amounts of consumption across time to take advantage of difference between its discount rate and the prevailing interest rate. For large values of θ, a household requires large differences between the interest rate and the discount rate in order to shift even small amounts of consumption across time.

1.B.2 Behavior of households and firms

Firms

The behavior of firms is relatively simple. At each point in time each firm employs stocks of capital and labor, pays them their respective marginal products, and sell the resulting output. The joint assumption of competitive markets and constant returns implies that all firms earn zero profits.

Firms solve the following optimization problem at each point in time.

$$\max_{K(t),L(t)} \Pi(t) = Y(t) - (r(t)+\delta)K(t) - W(t)L(t) \tag{1.83}$$

subject to the constraint imposed by the production function.

$$\begin{aligned} Y(t) =& F(K(t), A(t)L(t)) \\ =& A(t)L(t)F\left(\frac{K(t)}{A(t)L(t)}, 1\right) \\ =& A(t)L(t)f\left(\frac{K(t)}{A(t)L(t)}\right) \end{aligned} \tag{1.84}$$

The first-order conditions for the problem can be used to derive expression for both the

real return to capital and the real wage.

$$\begin{aligned}
\frac{\partial \Pi(t)}{K(t)} =& f'\left(\frac{K(t)}{A(t)L(t)}\right) - (r(t)+\delta) = 0 \implies \\
r(t) =& f'\left(\frac{K(t)}{A(t)L(t)}\right) - \delta \\
=& f'(k(t)) - \delta \qquad (1.85) \\
\frac{\partial \Pi(t)}{L(t)} =& A(t)\left[f\left(\frac{K(t)}{A(t)L(t)}\right) - \left(\frac{K(t)}{A(t)L(t)}\right) f'\left(\frac{K(t)}{A(t)L(t)}\right)\right] - W(t) = 0 \implies \\
W(t) =& A(t)\left[f\left(\frac{K(t)}{A(t)L(t)}\right) - \left(\frac{K(t)}{A(t)L(t)}\right) f'\left(\frac{K(t)}{A(t)L(t)}\right)\right] \\
w(t) =& \frac{W(t)}{A(t)} = f(k(t)) - k(t)f'(k(t)) \qquad (1.86)
\end{aligned}$$

Households' budget constraint

Each household takes the time-paths of r and W as given when solving its maximization problem. At each point in time a household's consumption and investment must not exceed its total income. Formally, at each point in time the household faces the following flow budget constraint.

$$\frac{C(t)L(t)}{H} + \frac{\dot{K}(t)}{H} \leq \frac{W(t)L(t)}{H} + \frac{r(t)K(t)}{H} \qquad (1.87)$$

Because the marginal utility from consumption is always positive, the constraint will bind with equality for all t.

Household's maximization problem

The representative household wants to maximize its lifetime utility subject to its budget constraint. As in the Solow (1956) model, it is easier to work with variables normalized by the quantity of effective labor. To do this we need to express both the objective function and the budget constraint in terms of consumption and labor income per unit of effective labor.

Start with the objective function. Define $c(t)$ to be consumption per unit of effective labor. Thus consumption per worker, $C(t)$, equals $c(t)A(t)$. The household's instantaneous utility

function becomes

$$\begin{aligned}\frac{C(t)^{1-\theta}}{1-\theta} =& \frac{[c(t)A(t)]^{1-\theta}}{1-\theta} \\ =& \frac{\left[c(t)A(0)e^{gt}\right]^{1-\theta}}{1-\theta} \\ =& A(0)^{1-\theta}e^{(1-\theta)gt}\frac{c(t)^{1-\theta}}{1-\theta}\end{aligned} \tag{1.88}$$

Substituting 1.88 and the fact that $L(t) = L(0)e^{nt}$ into the household's objective function yields

$$\begin{aligned}U =& \int_0^\infty e^{-\rho t}\frac{C(t)^{1-\theta}}{1-\theta}\frac{L(t)}{H}dt \\ =& \int_0^\infty e^{-\rho t}\left[A(0)^{1-\theta}e^{(1-\theta)gt}\frac{c(t)^{1-\theta}}{1-\theta}\right]\frac{L(0)e^{nt}}{H}dt \\ =& A(0)^{1-\theta}\frac{L(0)}{H}\int_0^\infty e^{-(\rho-n-(1-\theta)g)t}\frac{c(t)^{1-\theta}}{1-\theta}dt \\ =& B\int_0^\infty e^{-\beta t}\frac{c(t)^{1-\theta}}{1-\theta}dt\end{aligned} \tag{1.89}$$

where $B \equiv A(0)^{1-\theta}L(0)/H$ and $\beta \equiv \rho - n - (1-\theta)g$. Note that $\beta > 0$ is required in order for the household's lifetime utility to converge.

Now consider the budget constraint, 1.87. Start by dividing both sides of the budget constraint by $A(t)L(t)$.

$$\begin{aligned}\frac{C(t)}{A(t)} + \frac{\dot{K}(t)}{A(t)L(t)} =& \frac{W(t)}{A(t)} + r(t)\frac{K(t)}{A(t)L(t)} \\ c(t) + \frac{\dot{K}(t)}{A(t)L(t)} =& w(t) + r(t)k(t)\end{aligned} \tag{1.90}$$

To derive and expression for the second term in equation 1.90, note that

$$\begin{aligned}\dot{K}(t) =& \frac{\partial(k(t)A(t)L(t))}{\partial t} = \dot{k}(t)A(t)L(t) + k(t)[\dot{A}(t)L(t) + A(t)\dot{L}(t)] \\ \frac{\dot{K}(t)}{A(t)L(t)} =& \dot{k} + \left[\frac{\dot{A}}{A(t)} + \frac{\dot{L}}{L(t)}\right]k(t) \\ \frac{\dot{K}(t)}{A(t)L(t)} =& \dot{k} + (n+g)k(t).\end{aligned} \tag{1.91}$$

Substituting this result into equation 1.90 yields the flow budget constraint for the repre-

sentative household in efficiency units.

$$\dot{k} = w(t) + [r(t) - (n+g)]k(t) - c(t) \tag{1.92}$$

In addition to the flow budget constraint, we need to impose the following transversality condition that rules out exploding time paths of consumption per effective worker.

$$\lim_{t\to\infty} |c(t)| < \infty \tag{1.93}$$

Household behavior

The household's objective is to choose the path of $c(t)$ in order to

$$\max_{c(t)} B \int_0^\infty e^{-\beta t} \frac{c(t)^{1-\theta}}{1-\theta} dt \tag{1.94}$$

subject to

$$\dot{k} = w(t) + [r(t) - (n+g)]k(t) - c(t), \; k(0) = k_0 \tag{1.95}$$

$$\lim_{t\to\infty} |c(t)| < \infty \tag{1.96}$$

where $B \equiv A(0)^{1-\theta} L(0)/H$ and $\beta \equiv \rho - n - (1-\theta)g$.

The solution to the household's optimization problem can be easily characterized using the theory of optimal control. The main tool for solving problems of optimal control is known as Pontryagin's maximum principle. The maximum principle states that the first-order, necessary condition for optimality requires the representative household to choose a feasible value of the control so as to maximize the Hamiltonian, H, at each point in time.[38]

The Hamiltonian for the representative household's problem is

$$H(t,k,c,\lambda) = Be^{-\beta t}\left(\frac{c(t)^{1-\theta}}{1-\theta}\right) + \lambda(t)\left[(r(t) - (n+g))k(t) + w(t) - c(t)\right] \tag{1.97}$$

where the variables k, c, and λ are respectively referred to as the state, control, and costate variables. So long as H is concave in the control, c, the maximum of H corresponds to an

[38] The classic reference for the theory of optimal control is Pontryagin (1959). Chiang and Wainwright (2005) and the mathematical appendix of Barro and Sala-i Martin (2003) provide nice, easy introductions to the core material. Kamien and Schwartz (2012) covers everything about optimal control theory that even the most mathematically inclined economist might ever want to know.

interior solution and therefore we can find the optimal choice of c as follows.

$$\frac{\partial H}{\partial c} = Be^{-\beta t}c(t)^{-\theta} - \lambda(t) = 0 \implies \lambda(t) = Be^{-\beta t}c(t)^{-\theta} \tag{1.98}$$

Since, along with the control, c, H depends on both k and λ, the maximum principle also stipulates how k and λ should evolve over time. These equations are referred to as the state equation and the costate equation, respectively.

$$\dot{k} = \frac{\partial H}{\partial \lambda} = (r(t) - (n+g))k(t) + w(t) - c(t) \tag{1.99}$$

$$\dot{\lambda} = -\frac{\partial H(t)}{\partial k} = -\lambda(t)(r(t) - (n+g)) \tag{1.100}$$

These two equations of motion, together with equation 1.98, define the optimal time-paths for c, k, and λ.

To derive the consumption Euler equation, differentiate equation 1.98 with respect to time

$$\begin{aligned}\dot{\lambda} =& B\left[e^{-\beta t}\left(-\theta c(t)^{-\theta-1}\dot{c}\right) - [\rho - n - g(1-\theta)]e^{-\beta t}c(t)^{-\theta}\right] \\ =& -Be^{-\beta t}c(t)^{-\theta}\left[\theta\left(\frac{\dot{c}}{c(t)}\right) + [\rho - n - g(1-\theta)]\right]\end{aligned} \tag{1.101}$$

and then equate the resulting expression for $\dot{\lambda}$ with the costate equation.

$$-Be^{-\beta t}c(t)^{-\theta}\left[\theta\left(\frac{\dot{c}(t)}{c(t)}\right) + [\rho - n - g(1-\theta)]\right] = -\lambda(t)(r(t) - (n+g))$$

Finally, after substituting for $\lambda(t)$ using equation 1.98 (and a bit of algebra) we arrive at the consumption Euler equation.

$$\begin{aligned}-Be^{-\beta t}c(t)^{-\theta}\left[\theta\left(\frac{\dot{c}}{c(t)}\right) + [\rho - n - g(1-\theta)]\right] =& -Be^{-\beta t}c(t)^{-\theta}(r(t) - (n+g)) \\ \theta\left(\frac{\dot{c}(t)}{c(t)}\right) + [\rho - n - g(1-\theta)] =& (r(t) - (n+g)) \\ \frac{\dot{c}(t)}{c(t)} =& \frac{r(t) - \rho - \theta g}{\theta}\end{aligned} \tag{1.102}$$

Equilibrium

The decentralized equilibrium of the optimal growth model is completely characterized by the two first-order conditions from the representative firm's profit maximization problem,

and the consumption Euler equation and budget constraint of the representative household.

These four equations can be combined to yield a two-dimensional system of non-linear differential equations

$$\dot{k}(t) = f(k(t)) - c(t) - (n + g + \delta)k(t) \tag{1.103}$$

$$\frac{\dot{c}(t)}{c(t)} = \frac{f'(k(t)) - \delta - \rho - \theta g}{\theta} \tag{1.104}$$

with the following boundary conditions.

$$k(0) = k_0 \tag{1.105}$$

$$\lim_{t \to \infty} |c(t)| < \infty \tag{1.106}$$

1.B.3 Analytic solution

Although the Ramsey (1928), Cass (1965), Koopmans (1965) model with Cobb-Douglas production does not have analytic solution for generic parameter values, Smith (2006) derives an analytic solution for the case where the inverse of the inter-temporal elasticity of substitution, θ, equals capital's share of income, α. Unfortunately, this parameter restriction leads to a linear saddle path which is of little use when the objective is to compare approximation errors for across numerical methods. Instead, in order to obtain a closed-form solution of the model that is sufficiently non-linear, I focus on the a special case of the model with a constant gross savings rate.

Start by defining the capital-output ratio, $z(t)$, and the consumption-output ratio, $\chi(t)$, as follows.

$$z(t) = \frac{k(t)}{y(t)} = k(t)^{1-\alpha} \tag{1.107}$$

$$\chi(t) = \frac{c(t)}{y(t)} = \frac{c(t)}{k(t)^{\alpha}} \tag{1.108}$$

Differentiating equation 1.107 with respect to t yields a relation between $\dot{z}$ and $\dot{k}$

$$\dot{z} = (1 - \alpha)k(t)^{-\alpha}\dot{k} \tag{1.109}$$

which can be used to transform equation 1.103 into a linear differential equation.

$$\dot{z} + (1 - \alpha)(n + g + \delta)z(t) + (1 - \alpha)\chi(t) = (1 - \alpha) \tag{1.110}$$

Taking logarithms and then differentiating equation 1.108 with respect to t yields a relation between the growth rate of $\chi(t)$ and those of $c(t)$ and $k(t)$

$$\frac{\dot{\chi(t)}}{\chi(t)} = \frac{\dot{c(t)}}{c(t)} - \alpha\frac{\dot{k(t)}}{k(t)} \tag{1.111}$$

which can be used to transform equation 1.104 into

$$\frac{\dot{\chi(t)}}{\chi(t)} = \left[\frac{\alpha}{\theta} - \alpha(1-\chi(t))\right] k(t)^{\alpha-1} + \alpha(n+g+\delta) - \left(\frac{\delta+\rho+\theta g}{\theta}\right). \tag{1.112}$$

Now conjecture that the consumption-output ratio, $\chi(t)$, is constant along the saddle-path. This conjecture implies the following.

$$\chi^* = 1 - \frac{1}{\theta} = \frac{\theta-1}{\theta},\ \theta > 1 \tag{1.113}$$

$$\rho = \alpha\theta(n+g+\delta) - (\delta+\theta g) > 0 \tag{1.114}$$

This conjecture makes equation 1.110 into a linear differential equation

$$\dot{z} + \lambda z(t) = (1-\alpha)(1-\chi^*) \tag{1.115}$$

where

$$\lambda = (1-\alpha)(n+g+\delta).$$

Given the initial condition $z(0) = z_0$, the above equation can be solved using standard methods to obtain the following solution for $z(t)$.[39]

$$z(t) = z_0 e^{-\lambda t} + \left(\frac{1}{\theta(n+g+\delta)}\right)\left(1 - e^{-\lambda t}\right) \tag{1.116}$$

Finally, transforming back to $k(t)$ using $z(t) = k(t)^{1-\alpha}$ and $z_0 = k_0^{1-\alpha}$ yields the analytic solution for the time path of capital per effective worker.

$$k(t) = \left[k_0^{1-\alpha} e^{-\lambda t} + \left(\frac{1}{\theta(n+g+\delta)}\right)\left(1 - e^{-\lambda t}\right)\right]^{\frac{1}{1-\alpha}} \tag{1.117}$$

With the solution for $k(t)$ in hand, the solution for $c(t)$ is implied by the definition of the consumption-output ratio, $\chi(t)$, which I have conjectured to be constant along the saddle path.

$$c(t) = \left(\frac{\theta-1}{\theta}\right) k(t)^{\alpha} \tag{1.118}$$

[39] See Chiang and Wainwright (2005) for an introductory discussion of such methods.

One can confirm that the original conjecture of $\chi(t) = \chi^*$ for all t is correct by checking that equations 1.117 and 1.118 jointly satisfy equations 1.103 and 1.104.

Chapter 2

Characterizing the size distribution of U.S. banks

Using detailed balance sheet data for all FDIC regulated banks for the years 1992 through 2011 this paper assesses the statistical support for Zipf's Law (i.e., a the power law distribution with a scaling exponent of $\alpha = 2$) as an appropriate model for the upper tail of the size distribution of U.S. banks. Although I find statistically significant departures from Zipf's Law for most measures of bank size in most years, a power law distribution with $\hat{\alpha} \approx 1.9$ out performs other plausible heavy-tailed alternative distributions.

2.1 Introduction

The past 20 years have seen significant agglomeration within the U.S. banking sector. In 1992 there were 13,973 banks regulated by the Federal Deposit Insurance Corporation (FDIC), the largest of which, Citibank, controlled roughly 3.5% of all U.S. banking assets. By the end of 2011, the number of banks under FDIC regulation had fallen by almost 50% to 7,366, and the largest remaining bank, JP Morgan-Chase, controlled approximately 13% of all U.S. banking assets. Figure 2.1 shows the fraction of total U.S. banking assets, loans, liabilities, deposits, equity, and employees controlled by the single largest bank from 1992 to 2011.[1] The extent of agglomeration within the banking sector appears even more dramatic when one examines the market shares held by the ten largest U.S. banks. Figure 2.2 plots these market shares for the same measures of bank size used in figure 2.1. The

[1]Monetary figures are first deflated and then re-scaled by dividing through by banking sector totals relative to 2011. See section 2.2 for the details.

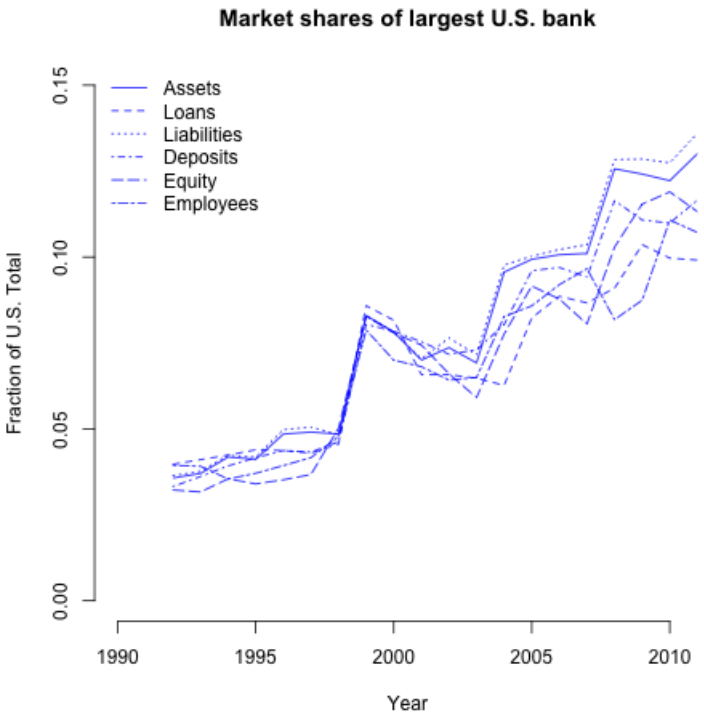

Figure 2.1: The market shares of the largest U.S. bank have increased markedly over the past 20 years.

market share of the 10 largest banks has more than doubled over the last 20 years from less than 20% to roughly 50%.[2]

The objective of this paper is to assess whether the levels of concentration observed in the data are consistent with Zipf's Law (i.e., a the power law distribution with a scaling exponent of $\alpha = 2$) using statistical techniques advocated by Clauset et al. (2009).[3] Although the banking literature has long been interested in the size distribution of banks,[4]

[2]The "jumps" in the share of assets and liabilities controlled by the largest U.S. bank all occurred because of mergers between large banks. In 1999, Nations Bank, the largest U.S. bank at the time, merged with Bank of America. JPMorgan-Chase, itself the largest bank at end of 2003, purchased the sixth largest bank in the U.S., Bank One, in 2004. Finally, JPMorgan-Chase acquired both Bear Stearns and Washington Mutual in 2008 in the aftermath of the global financial crisis. These "jumps" in market shares also hint at another important stylized fact: there seems to be no preferred scale for bank mergers. Small banks merge with other small banks; small banks merge with larger banks; and large banks also frequently merge with other extremely large banks.

[3]Given some measure of bank size Zipf's Law states that the size of a bank should be inversely proportional to its rank. Put another way, under Zipf's law, the largest bank should be approximately twice as large as the second largest bank, three times as large as the third largest bank, etc. The original exposition of Zipf's law can be found in Zipf (1949). More modern treatment of Zipf's Law can be found in Gabaix (1999) and Gabaix (2008).

[4]Recent studies include Berger et al. (1995), Ennis (2001), Goddard et al. (2004), Jones and Critchfield (2005), and Janicki and Prescott (2006), Benito (2008).

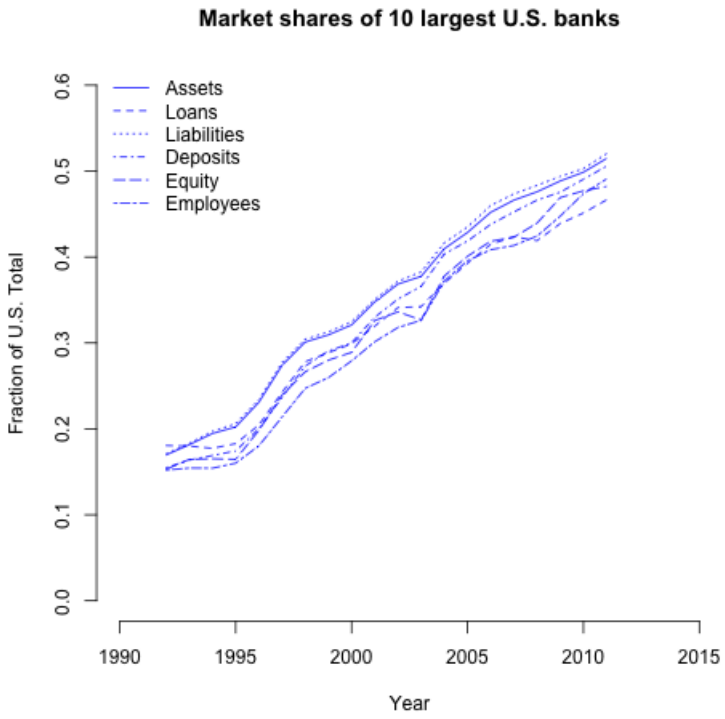

Figure 2.2: Irrespective of the measure of bank size, the 10 largest U.S. banks currently control roughly 50% of the market.

this literature has not generally focused on fitting specific statistical models to data on bank size.[5] I have chosen to focus on characterizing the size distribution of U.S. banks because the distribution contains important information on the pace and nature of concentration within the U.S. banking sector. Concentration, particularly in excess of what can be supported by economic fundamentals would call into question the competitiveness of the banking sector.

Using detailed balance sheet data for all FDIC regulated banks for the years 1992 through 2011, I find significant departures from Zipf's Law for most measures of bank size in most years. Although Zipf's Law can be statistically rejected, a power law distribution with $\hat{\alpha} \approx 1.9$ outperforms other plausible heavy-tailed alternative distributions. Power law distributions with scaling exponents $\alpha < 2$ display some startling mathematical properties that have important economic implications.[6] For example, such distributions have no well-defined expected values. The failure of the expected value to be well-defined implies that the fraction of U.S. banking sector totals in the top *anything* of the size distribution (even

[5] Janicki and Prescott (2006) is an important exception.

[6] The mathematical properties of power laws are discussed in detail in appendix 2.A.

the top 1% of banks) should tend to unity in the limit of an infinite number of banks. In practice, a scaling exponent of $\alpha < 2$ implies that one should expect effectively all of the assets, loans, liabilities, deposits, and equity in the U.S. banking sector to be controlled by a small number of banks in the extreme tail of the size distribution.

I view characterizing the bank size distribution as an important first-step toward developing an empirically relevant theory of the banking. A natural point of departure for such a theory would be Gabaix (2011). Gabaix (2011) posits that, so long as the upper tail of the firm size distribution is sufficiently heavy, much of the variation in aggregate macroeconomic time series data over the business cycle can be explained by idiosyncratic shocks to individual firms. The "granularity hypothesis" of Gabaix (2011) constrasts sharply with existing research on business cycles which has focused almost exlcusively on the role played by aggregate shocks. An empirical characterization of the upper tail of the bank size distribution is a necessary precusor to any future attempt to develop a "granularity hypothesis" for banking.

The remainder of this paper proceeds as follows. In the following section I discuss the data used in the analysis. Section 2.3 details the statistical methodology that I use for fitting the power law model to the data. Section 2.4 presents the empirical results and section 3.5 provides some concludes and discusses some avenues for future research.

2.2 Data

All bank data used in this study are taken from the Statistics on Depository Institutions (SDI) database maintained by the U.S. Federal Deposit Insurance Corporation (FDIC). The SDI data set contains aggregate demographic and financial information about the U.S. banking sector, as well detailed data on individual bank (or bank holding company) balance sheets, income statements, performance ratios, etc., dating back to 1992. In analyzing the evolution of the size distribution of U.S. banks I look at six separate measures of bank size: total assets, total loans, total liabilities, total deposits, total equity, and number of employees.[7] In order to make sure my results are comparable across years, I deflate and

[7] Definitions of variables are as follows:

- Total assets (*asset*): The sum of all assets owned by the institution including cash, loans, securities, bank premises and other assets. This total does not include off-balance-sheet accounts.
- Total loans (*lnlsnet*): Total loans and lease financing receivables minus unearned income and loan loss allowances.
- Total liabilities (*liab*): Deposits and other borrowings, subordinated notes and debentures, limited-life preferred stock and related surplus, trading account liabilities and mortgage indebtedness.
- Total deposits (*dep*): The sum of all deposits including demand deposits, money market deposits, other savings deposits, time deposits and deposits in foreign offices.

then re-scale each measure of bank size by dividing by banking sector totals relative to 2011.[8]

Perhaps the most salient feature of the FDIC data is the enormous heterogeneity in the size of U.S. banking institutions irrespective of how size is measured. Figure 2.3 plots density estimates of the size distribution of U.S. banks where size is measured in terms of assets, deposits, liabilities, loans, and equity. The data range over six orders of magnitude (depending a bit on the measure of size). One can see that the distributions of bank assets, deposits, and liabilities lie almost on top of one another. The distributions of bank size measured in terms of equity or loans are similar in form to the distributions using the other measures, but are both shifted to the left (equity more so than loans). This rough ordering of the distributions is consistent for each year in the sample.[9]

Five of the six size measures included in this study are based on the various components of banks' balance sheets. The final size measure, number of employees, is a measure of size often considered in the literature on the size distribution of firms. Figure 2.4 shows the evolution of the bank size distribution from 1992 to 2011 under this alternative measure.

To give a sense of how the extreme upper tail of the size distribution of banks has evolved over the last 20 years it is useful to examine the survival functions for the various measures of bank size. Figure 2.5 plots estimates of the survival function of the size distributions from 1992 to 2011 where size is again measured by normalized assets. Overlaying the survival functions for each year clearly documents a thickening of the extreme upper tail consistent with the agglomeration hinted at in figure 2.2 above. There also appears to be a kink (i.e., change in slope) of the survival functions at about $1 billion in total assets. Interestingly, the FDIC defines "community banks" to be exactly those banks with less than $1 billion in total assets and manages a number of government programs whose purpose is to encourage lending by these banks.

- Total equity (*toteq*): Total equity capital on a consolidated basis.
- Number of employees (*numemp*): The number of full-time employees on the payroll of the bank and its subsidiaries at the end of the quarter.

More details can be found in the SDI online documentation.

[8] Specifically, let $S_{i,t}^{raw}$ denote the raw size of bank i in year t based on one of the six size measures detailed above. The normalized size of bank i relative to the base year, $t = 2011$, is defined as follows:

$$S_{i,t}^{norm} = \left(\frac{S_{i,t}^{raw}}{\sum_j S_{j,t}^{raw}} \right) \sum_j S_{i,2011}^{raw} \tag{2.1}$$

where $\sum_j S_{j,t}^{raw}$ is the banking sector total of some size measure in year t (i.e., total banking sector assets in year t), and $\sum_j S_{j,2011}^{raw}$ is the banking sector total of the same size measure in the year 2011.

[9] Density estimates of the size distribution for a given measure exclude banks with non-positive or unreported values.

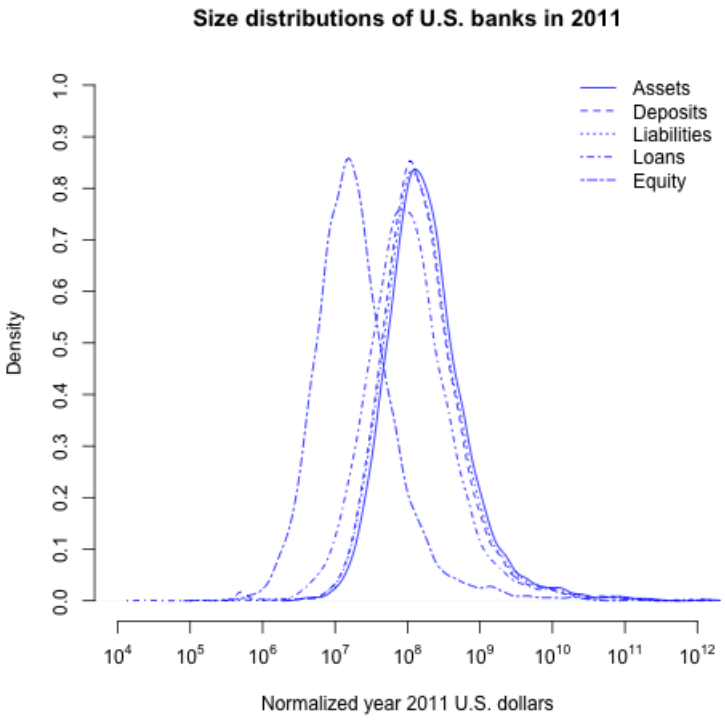

Figure 2.3: There is enormous heterogeneity in the distribution of bank size in the U.S. irrespective of the measure of size used. In 2011 the largest bank, in terms of assets, was almost six orders of magnitude larger than the smallest bank!

2.3 Methodology

Power law distributions have been used to fit data in a wide variety of scientific fields from astrophysics to linguistics,[10] and have a long intellectual tradition in the social sciences as a model for heavy-tailed phenomena (particularly in economics where it is better known as the Pareto distribution).[11]

Given that this paper deals with empirical measures of bank size that are (at least approximately) continuous, I restrict attention to the continuous version of the power law distribution. The density function, $p(x)$, for the power law distribution can be written as

[10] Newman (2005) surveys various applications of the power law model including word frequency, citations of scientific papers, web hits, copies of books sold, telephone calls, earthquakes, moon craters, solar flares, intensity of wars, individual wealth, frequency of family names, and the population of cities.

[11] Early reference are Pareto (1896), Champernowne (1953), Simon (1955), Mandelbrot (1963) and Steindl (1965). Gabaix (2008) is an excellent, recent review of applications of power laws in economics.

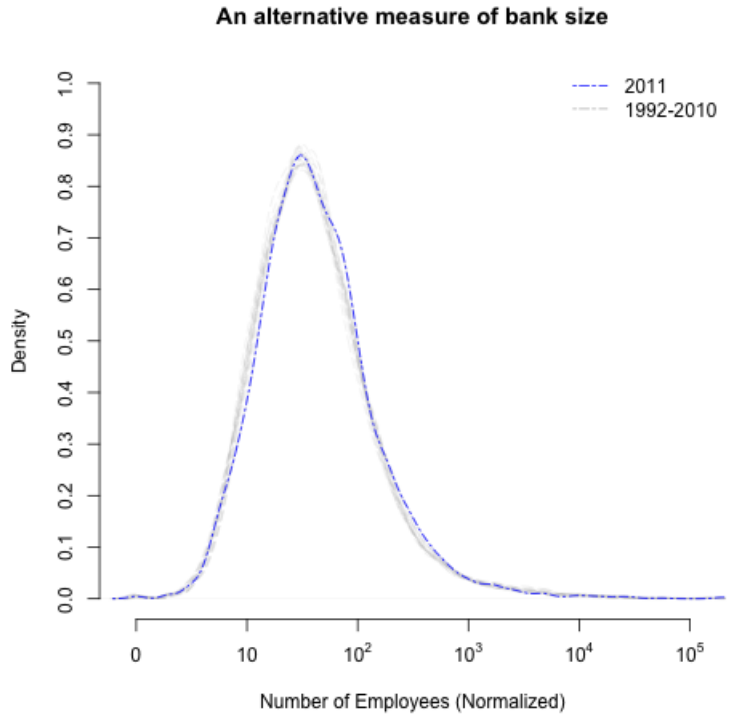

Figure 2.4: There is even substantial heterogeneity in the distribution of bank size in the U.S. when size is measured by number of employees.

follows.[12]

$$p(x) = \left(\frac{\alpha - 1}{x_{min}}\right)\left(\frac{x_{min}}{x}\right)^{\alpha} \tag{2.2}$$

note that normalization requires $\alpha > 1$. The cumulative distribution function, $P(x)$, can be derived from integrating the density function derived above:

$$P(X) = Pr(X \leq x) = 1 - \left(\frac{x_{min}}{x}\right)^{-(\alpha-1)} \tag{2.3}$$

One of the most useful properties of the power law distribution is that the survival function (sometimes also referred to as the upper cumulative distribution function) also follows a power law:

$$Pr(X > x) = 1 - P(x) = \left(\frac{x_{min}}{x}\right)^{-(\alpha-1)} \tag{2.4}$$

Note that the power law scaling exponent of the survival function is $\alpha - 1$, which is one less than the scaling exponent of the power law density function.

[12]Complete derivations of the density function, cumulative distribution function, survival function, as well as many other interesting mathematical results about power laws can be found in appendix 2.A.

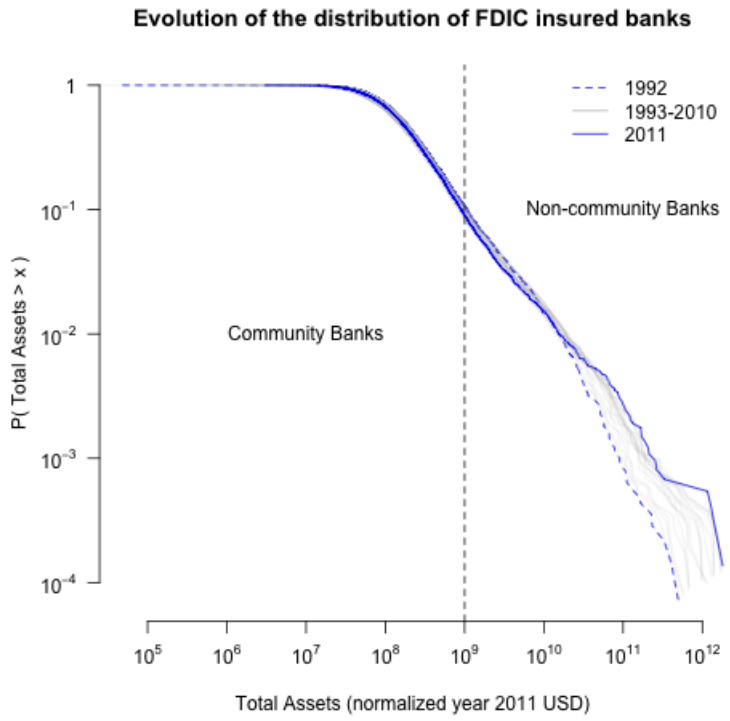

Figure 2.5: The upper tail of the bank size distribution has become "heavier" over the last 20 years. Plots of the survival functions for the other measures of bank size show a similar pattern.

In the remainder of this section I discuss, in detail, the procedure for fitting power law models to data advocated by Clauset et al. (2009). The procedure has three stages: fitting the power law distribution to the data, validating the model fit using goodness-of-fit tests, and finally testing the fitted model against plausible alternatives.

2.3.1 Parameter Estimation

Previous empirical research typically uses either OLS or the Hill (1975) procedure to estimate the scaling exponent, α, of the power law model. In this section I briefly summarize and critique these approaches before detailing the maximum likelihood procedure advocated by Clauset et al. (2009).

OLS Methods

Perhaps the most obvious method for estimating the power law scaling exponent, α, would be to take logarithms of equation 2.2 in order to obtain:

$$\ln\, p(x) = [\ln\, (\alpha - 1) + (\alpha - 1)\ln\, x_{min}] - \alpha \ln\, x + \epsilon \tag{2.5}$$

where ϵ is a mean zero disturbance that is uncorrelated with $\ln\, x$. If the quantity of interest, x, is power law distributed above some threshold x_{min} with scaling exponent α, then the above equation implies that its density function will be log-linear with a slope equal $-\alpha$ above the threshold x_{min}. A procedure to estimate the scaling exponent would be to obtain an estimate of the true density $p(x)$ using a simple histogram, plot the histogram on double logarithmic scales, estimate the lower bound x_{min} visually, and then estimate α by applying OLS to the observations above x_{min}.

A second method for estimating α, one encountered frequently in economics literature on power laws, takes logarithms of equation 2.4 to obtain:

$$\ln\, P(x) = [(\alpha - 1)\ln\, x_{min}] - (\alpha - 1)\ln\, x + \epsilon \tag{2.6}$$

where ϵ is a mean zero disturbance that is uncorrelated with $\ln\, x$. Again, if the quantity of interest, x, is power law distributed above some threshold x_{min} with scaling exponent α, then the above equation implies that the complementary cumulative distribution function, $P(x)$, will be log-linear with a slope equal $-\zeta = -(\alpha - 1)$ above the threshold x_{min}. A procedure to estimate the scaling exponent would be to obtain an estimate of $P(x)$ by constructing a simple rank ordering of the data, plot $P(x)$ on double logarithmic scales, estimate the lower bound x_{min} visually, and then estimate α by applying OLS to the observations above x_{min} to obtain an estimate of ζ from which the estimate of α can be backed out. In practice, researchers using this method often report their estimates for ζ as the power law scaling exponent rather than backing out the estimate of α.

A final method for fitting power law models to data using OLS, which is closely related to the approach used to estimate equation 2.6, is called a "log-rank, log-size" regression. First, order the data by size, S, letting $S_{(1)} \geq S_{(2)} \geq \cdots \geq S_n = x_{min}$ denote the observation of rank n. As with the other OLS methods, the choice of x_{min} is fairly arbitrary and based on either a visual assessment of goodness-of-fit of a linear model for the observations above x_{min} or by simply restricting the analysis to the upper 5% of observed values. One can then estimate the power law scaling exponent ζ by regressing log-rank i on log size using the follow specification

$$\ln\, (i - s) = \text{constant} - \zeta \ln\, S_i + \epsilon \tag{2.7}$$

where ϵ is a mean zero disturbance that is uncorrelated with ln S, and the parameter s is a shift correction.[13]

Criticisms of OLS-based methods

There are several issues with the above OLS based procedures (in addition to the arbitrary nature of the choice for x_{min}). Clauset et al. (2009) demonstrate via simulation that estimates of the scaling exponent ζ (or α) obtained using OLS as described in the above procedures are significantly biased even in large (i.e., $N \approx 10e4$) samples. Estimating equation 2.5 using OLS is particularly problematic as the magnitude of the bias is quite sensitive to the choice of binning scheme used in constructing the histogram (i.e., estimate of the density function $p(x)$). For OLS fits of the power law model using either equation 2.6 or 2.7 the classic OLS standard errors for the scaling exponent are no longer valid as adjacent values of both $P(x)$ or log-rank are highly correlated by construction and this introduces significant correlations into the disurbance term. Finally, there is no guarantee that the OLS estimate of the scaling exponent will, when combined with the researcher's choice of x_{min}, result in a valid probability distribution.

In addition to difficulties in estimating the scaling exponent, researchers using any of the above OLS procedures will not easily be able to assess the goodness-of-fit of the power law model. Typically one assesses goodness-of-fit for a linear model estimated using OLS by examining the R^2. However, the R^2 associated with OLS estimation of any of 2.5, 2.6, or 2.7 can be a misleading measure of goodness-of-fit for the power law model. Informally, the problem is that there are many non power law distributions that will appear to be "roughly linear" when plotted on doubly logarithmic scales, and in these instances OLS estimation of will likely yield a very large R^2. In simulation experiments, large measures of R^2 can be obtained even when the true underlying distribution is not well-approximated by a power law. As such a high value for R^2 provides little indication of how well the power law model actually fits the data.[14]

[13] Typically, the shift correction is simply $s = 0$, however Gabaix and Ibragimov (2011) notes that OLS estimates of ζ can be heavily biased in small samples and argues that $s = \frac{1}{2}$ is optimal to reduce the bias. Furthermore, in this case a correction of

$$(\hat{\alpha} - 1)\left(\frac{n}{2}\right)^{-\frac{1}{2}} \tag{2.8}$$

to the classic OLS standard error is required in order to account for the auto-correlation in the data introduced by the ranking procedure.

[14] Gabaix (2008) provides an alternative test for the goodness-of-fit for the power law model within the OLS regression framework. The approach is to estimate a modified form of equation 2.7 that includes a quadratic term designed to capture deviations from a pure power law.

Method of Maximum Likelihood

Clauset et al. (2009), suggest estimating the scaling exponent α of the power law model using a maximum likelihood procedure which is identical to the Hill (1975) estimator popular in the economics and finance literature, particularly the subset of this literature that uses techniques from a branch of statistical theory known as extreme value theory.[15]

Given some data set containing n observations $x_i \geq x_{min}$ and a particular value of α the likelihood that the data were drawn from a power law model is proportional to

$$p(x|\alpha) = \prod_{i=1}^{n} \left(\frac{\alpha - 1}{x_{min}} \right) \left(\frac{x_{min}}{x} \right)^{\alpha}. \tag{2.9}$$

Using the method of maximum likelihood the data are most likely to have been generated by a power law model with a scaling parameter α that maximizes this function. Taking logarithms yields

$$\begin{aligned} \mathcal{L} = \ln p(x|\alpha) =& \sum_{i=1}^{n} \left[\ln(\alpha - 1) - \ln x_{min} + \alpha \ln \left(\frac{x_{min}}{x_i} \right) \right] \\ =& n \ln(\alpha - 1) - n \ln x_{min} - \alpha \sum_{i=1}^{n} \ln \left(\frac{x_i}{x_{min}} \right). \end{aligned} \tag{2.10}$$

Taking the derivative of the log-likelihood function with respect to α and setting the result equal to zero yields the maximum likelihood estimator for the power law scaling exponent.

For a given value of x_{min}, the maximum likelihood estimator for the scaling exponent is

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \left(\frac{x_i}{x_{min}} \right) \right]^{-1}. \tag{2.11}$$

Equation 2.11, is equivalent to the Hill (1975) estimator, and has been shown to be asymptotically normal Hall (1982) and consistent Mason (1982). The standard error of $\hat{\alpha}$ is

$$\sigma = \frac{\hat{\alpha} - 1}{\sqrt{n}} + \mathcal{O}\left(n^{-1}\right) \tag{2.12}$$

Note, however, that equation 2.11 holds only for a given value of x_{min}.[16] Thus in order to

[15] For more on applications of extreme value theory to economics and finance see Embrechts et al. (1997), Kotz and Nadarajah (2000), Beirlant (2004), or Resnick (2007).

[16] It is not possible to estimate the threshold parameter x_{min} using maximum likelihood as maximizing the likelihood function would require setting $x_{min} = \infty$.

estimate α using the method of maximum likelihood, the researcher must take a stand on the "optimal" choice of the scaling threshold, x_{min}. In this respect the maximum likelihood estimation procedure is no different than the OLS methods detailed above. However, instead of employing the *ad hoc* selection procedures typically used in the literature, Clauset et al. (2009) suggest estimating x_{min} from the data by choosing a value $\hat{x}_{min}$ that minimizes the "distance" between the empirical distribution of the observed data and the best-fit power law model above $\hat{x}_{min}$.[17]

Roughly speaking, the Clauset et al. (2009) procedure for choosing x_{min} works as follows. First a set of candidate threshold parameters, Θ, is chosen from the observed data. Then for each $x^c_{min} \in \Theta$ one obtains an estimate of the scaling exponent $\hat{\alpha}^c$ using maximum likelihood and then calculates the Kolmogorov-Smirnov (KS) distance between the empirical distribution of the data above x^c_{min} and the theoretical distribution of a power law with parameters $\hat{\alpha}^c$ and x^c_{min}. The optimal choice for the threshold parameter, x_{min}, is the $x^*_{min} \in \Theta$ which minimizes the KS distance between the observed data above x^*_{min} and the theoretical power law distribution with scaling exponent, $\hat{\alpha}^*$ (i.e., the maximum likelihood estimate of α obtained by applying equation 2.11 with $x_{min} = x^*_{min}$), and threshold parameter, x^*_{min}.

In order to get estimates of parameter uncertainty that accurately take into account the flexibility introduced by the joint estimation of α and x_{min}, standard errors and confidence intervals for the parameter estimates are estimated using a basic non-parametric bootstrap procedure detailed in Davison (1997).

2.3.2 Assessing Goodness-of-fit

By definition the maximum likelihood estimation procedure described above will find the "best-fitting" power law model for the upper tail of the size distribution of U.S. banks. However, the estimation procedure itself says nothing about how good an approximation the power law model actually is to the tail of that distribution.[18] I would like to ask whether, given data on some measure of bank size, the power law model is a plausible model for the upper tail of the size distribution.

[17] See Beirlant (2004) for a detailed discussion of the costs and benefits of the various strategies *ad hoc* methods for choosing x_{min}. The specific procedure for joint selecting x_{min} used in Clauset et al. (2009) was first suggested and implemented in Clauset et al. (2007).

[18] As pointed out by Clauset et al. (2009), in general, authors rarely assess the goodness-of-fit for the power law model. While Janicki and Prescott (2006) avoid committing the sin of using an OLS-based procedure for estimating the scaling exponent, they do fail to assess (or at least report) any test of the plausibility of the power law distribution as a model for their data.

I test the plausibility of the power law model using the following simulation procedure.[19] First, I fit the power law model to data using the above maximum likelihood procedure to find the optimal parameter estimates $\hat{\alpha}$, and $\hat{x}_{min}$. I then extract the KS test statistic for the optimal fit, which I will use as my "observed" test statistic. I then generate a large number, say $B = 2500$, synthetic data sets that mimic the empirical data below $\hat{x}_{min}$, but follow a true power law with scaling parameter $\hat{\alpha}$ in the tail.[20] For each of the $i \in 1, \ldots, B$ synthetic data sets, I fit the power law model to i-th synthetic data set and find the optimal parameters using the above maximum likelihood procedure and calculate the KS test statistic for this optimal fit. As the p-value for my goodness-of-fit test, I take the fraction of the B KS statistics larger than the "observed" KS statistic. A large p-value (i.e., > 0.10) indicates power law model is plausible; small p-values (i.e., p-values ≤ 0.05) indicates power law can be rejected as plausible given the data. To provide some context, a p-value of 0.05 indicates that there is roughly a 1 in 20 chance that I would observe data on bank sizes for a given year that agree as poorly (as measured by the KS test statistic) with the power law model as the data that I actually observe.

2.3.3 Testing Alternative Hypotheses

Once I have determined the plausibility of the power law model using goodness-of-fit testing, the next step in the methodology is to rigorously test the power law model against several alternative hypotheses using Vuong (1989) likelihood ratio testing procedures as advocated in Clauset et al. (2009). The Vuong (1989) likelihood procedure comes in two flavors depending on whether or not the distribution chosen as the null hypothesis is "nested" within the class of distributions chosen as the alternative hypothesis. I consider three common alternative distributions for heavy-tailed data: the log-normal, stretched-exponential, and power law with an exponential cut-off; and one thin-tailed alternative: the exponential distribution. Of the alternatives considered, all but the power law with an exponential cut-off make use of the "non-nested" flavor of the Vuong (1989) test. The density functions for the power law and each of the considered alternatives are listed in table 2.1.

In comparing the power law null to the non-nested alternatives, I implement the Vuong

[19] For a more complete description of the goodness-of-fit testing procedure see Clauset et al. (2009).

[20] Specifically, suppose that I have n observations of bank size in a given year, and based on the maximum likelihood fit of the power law model I have found that $n_{tail} < n$ of the data points lie in the positive tail of the size distribution. I then generate a synthetic data set of length n by selecting with probability $\frac{n_{tail}}{n}$ a random draw from a true power law distribution with parameters $\hat{\alpha}$ and $\hat{x}_{min}$, otherwise with probability $1 - \frac{n_{tail}}{n}$ I select uniformly at random an observed value for bank size strictly less than $\hat{x}_{min}$. Repeating the selection procedure n times generates a synthetic data set that mimics the observed data below the estimated threshold $\hat{x}_{min}$, but that follows a true power law with scaling exponent $\hat{\alpha}$ above $\hat{x}_{min}$.

	$p(x) = Cf(x)$	
Distribution	C	$f(x)$
Power law	$\frac{\alpha-1}{x_{min}} x_{min}^{\alpha}$	$x^{-\alpha}$
Truncated power law	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha,\lambda x_{min})}$	$x^{-\alpha}e^{-\lambda x}$
Exponential	$e^{-\lambda x}$	$\gamma e^{\lambda x_{min}}$
Stretched exponential	$x^{\beta-1}e^{-\lambda x^{\beta}}$	$\beta\lambda e^{\lambda x_{min}^{\beta}}$
Log-normal	$\sqrt{\frac{2}{\pi\sigma^2}}\left[\text{erfc}\left(\frac{\ln x_{min}-\mu}{\sqrt{2}\sigma}\right)\right]^{-1}$	$\frac{1}{x}e^{-\left(\frac{\ln x-\mu}{2\sigma}\right)^2}$

Table 2.1: Mathematical definitions of the power law distributions and the various alternatives considered in this paper. For each of the distributions, I give the basic functional form, $f(x)$, along with the normalization constant, C.

(1989) likelihood ratio test in two steps. First I consider the two-sided null hypothesis that each of the power law and the alternative are equally far from the true distribution against a general alternative. If I reject this two-sided null hypothesis, then I conclude that one of the power law or the alternative is preferred (given the data). Following a rejection of the two-sided null, I then move to test a one-sided null hypothesis of a power law against the alternative distribution directly. A rejection of the two-sided null hypothesis, followed by a failure to reject the one-sided null hypothesis of a power law leads me to conclude that the data are sufficient to distinguish between the power law and the considered alternative, and the the power law model if preferred. If, however, I fail to reject the two-sided null hypothesis, then the test is indeterminate: there is simple not enough data to distinguish between the power law and the alternative. A major advantage of using the Vuong (1989) likelihood ratio approach over the classic Wilks (1938) likelihood ratio test is that the former will alert the researcher to the possibility that the data are not sufficient to favor either the power law or the alternative model.

2.4 Results

2.4.1 Fitting a log-normal distribution

Figure 2.6 shows kernel density estimates of the size distribution of banks, where size is measured by normalized assets, for each year from 1992-2011. The distributions look remarkably similar and are, at a glance, consistent with a log-normal distribution. Upon closer inspection, however, the distributions appear to have a heavy upper tail and a slight

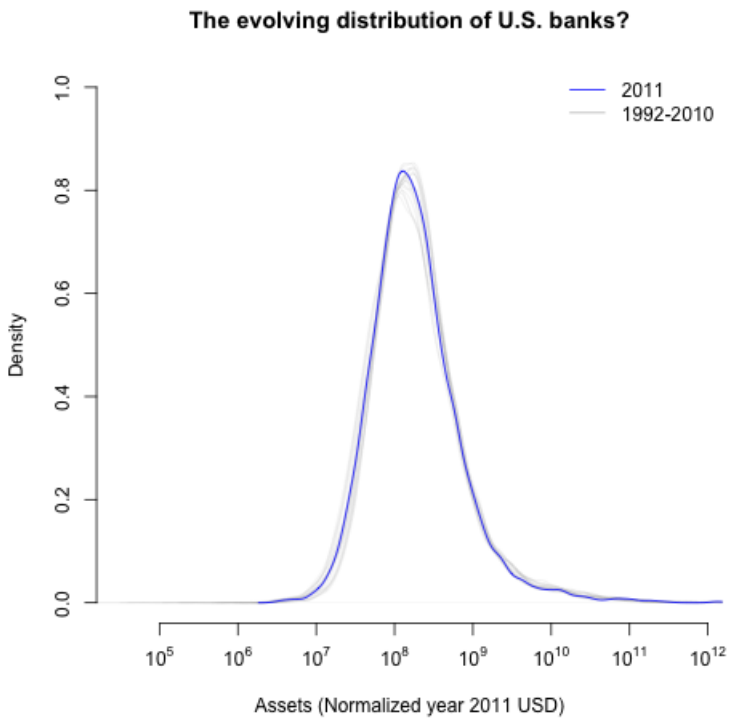

Figure 2.6: Kernel density estimates of the size distributions appear remarkably constant across time. However, these plots obscure the agglomeration occurring in the upper tail of the distributions.

right-skew. Given the logarithmic scale on the horizontal axis, any significant right-skew would be inconsistent with a log-normal distribution: the log-normal distribution should have thin, exponentially decaying tails when plotted on a logarithmic scaled horizontal axis.

Indeed, the log-normal distribution does a particularly poor job of fitting the right tail of the size distribution. The log-normal distribution systematically underestimates the probability of observing "large" banks. This fact is brutally demonstrated by figure 2.7 which plots the upper cumulative distribution function of the size distribution of U.S. banks in 2011 and then overlays the upper cumulative distribution functions for 2500 synthetic log-normal data sets.[21] Figure 2.7 clearly indicates that the log-normal distribution is a poor fit: it over-predicts the number of small to medium size banks and substantially

[21] Specifically, I generate 2500 different sets of parameter estimates (i.e., log-μ and log-σ) for the best-fitting log-normal distribution using a parametric bootstrap procedure, and then create each of the synthetic log-normal data sets using a unique set of bootstrap parameter estimates. The goal is to generate a "log-normal cloud" that gives a visual indication of the variability of the log-normal distribution in the upper tail while taking into account the statistical uncertainty in the estimated parameters of the distribution itself.

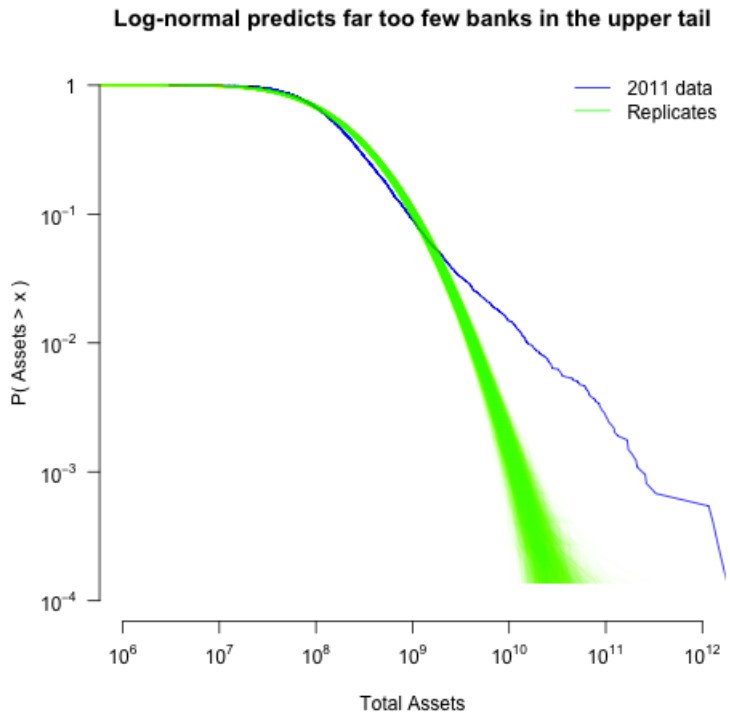

Figure 2.7: The log-normal distribution is unable to reproduce the extremely fat upper tail, which is a key feature of the data irrespective of the measure of bank size.

under-predicts the number of very large banks.

2.4.2 Fitting the power law distribution

Parameter Estimates

Estimation results for the scaling exponent α and the threshold parameter x_{min} of the power law distribution are reported for each of the six measures of bank size in tables 2.2 to 2.7. The results are summarized in figure 2.8. I find statistically significant departures from Zipf's Law (i.e., a power law with $\alpha = 2$) for most measures of bank size in most years for which I have data. Of the six measures considered, the only measure of bank size broadly consistent with Zipf's Law is number of employees.

When either total assets, net loans and leases, or total liabilities are used as the measure of bank size the estimated scaling exponent of $\hat{\alpha} \approx 1.9$ is roughly constant (within 95% confidence bands) across time. When total deposits are used as the measure of bank size the estimated scaling exponent is slightly larger (though still significantly less than $\alpha = 2$),

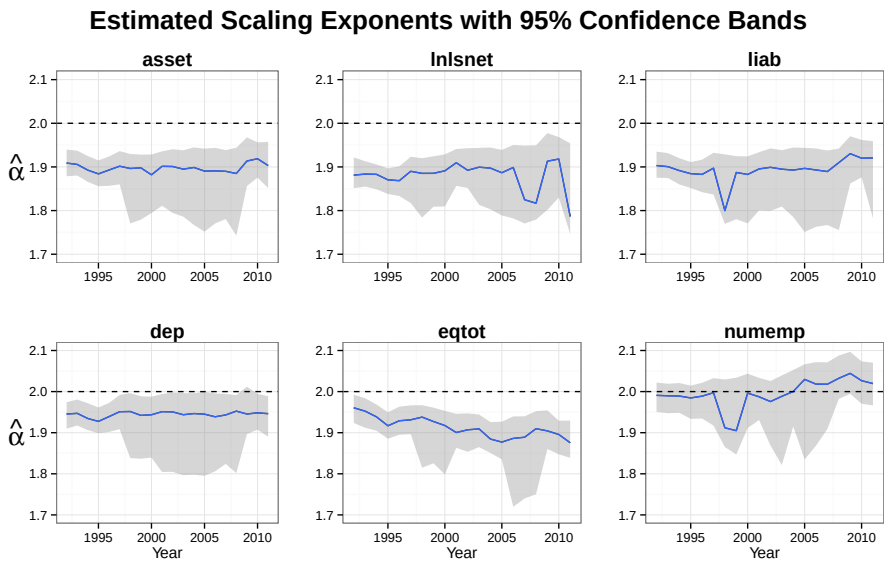

Figure 2.8: There are significant departures from Zipf's Law for most measures of bank size in most years. The tail of the bank size distribution is simply *too heavy* to be consistent with Zipf's Law.

than those obtained using either assets, loans, or liabilities.[22] The only measure of bank size for which there is a clear trend in the estimated scaling exponent is equity. The negative trend in $\hat{\alpha}$ when equity is used as the proxy for bank size indicates that equity holdings are becoming increasingly concentrated in the tail of the distribution.

These results indicate that the upper tail of the bank size distribution is generally too heavy (i.e., there are simply too many extremely large banks) to be consistent with Zipf's Law. In fact, the tail of the size distribution is so heavy (i.e., $\hat{\alpha} < 2$) that the mean of the best-fitting power law model is undefined. Put another way, conditional on being in the upper tail of the distribution, there is no such thing as an "average" sized bank!

Goodness-of-Fit Testing

In order to assess whether or not the power law model is a plausible model for the upper tail of the size distribution I implement a bootstrap version of the Kolmogorov-Smirnov (KS) goodness-of-fit test advocated in Clauset et al. (2009). Perhaps because goodness-of-fit testing has been strongly neglected in the empirical literature on power laws, Clauset et al. (2009) make a point to emphasize the importance of assessing the goodness-of-fit of the

[22] The difference between estimated scaling exponents when using deposits versus assets, loans, and liabilities as the measure of bank size is statistically significant in only half of the years in the sample.

power law model.[23] Table 2.8 reports the observed KS distance between the best fitting power law model as well as one-sided 95% confidence intervals for this statistic computed using the non-parametric bootstrap procedure discussed in section 2.3. The results of these goodness-of-fit tests are not supportive of the power law distribution as a model for the upper tail of the bank size distribution: in the majority of cases the observed KS distance between the best-fit power law model and the data is larger than any of the simulated KS distances.[24]

While the goodness-of-fit tests document statistically significant deviations from the power law distribution, the reason that the power law distribution is rejected as plausible differs depending on the year. In the 1990's the upper tail of the distribution of U.S. banks, though too heavy to be consistent with Zipf's Law, was *too thin* relative to the best-fitting power law distribution for the power law to be a plausible model for the data. However, following two decades of consolidation within the U.S. banking sector, the power law model is rejected for the completely opposite reason: in late 2000's, the extreme upper tail of the bank size distribution was *too heavy* relative to the best-fit power law model.

Testing Alternative Hypotheses

Although the power law model is statistically rejected for most measures of bank size in most years by the goodness-of-fit tests, without a structural model of the size distribution of banks it is difficult to assess whether or not the observed deviations are economically meaningful. In this section I test the relative performance of the power law distribution against several alternative distributions using the Vuong (1989) likelihood ratio test. Recall from the discussion in section 2.3 that the test comes in two flavors depending on whether or not the distribution chosen as the null hypothesis is "nested" within the class of distributions chosen as the alternative hypothesis. I consider three common alternative distributions for heavy-tailed data: the log-normal, stretched-exponential, and power-law with an exponential cut-off; and one thin-tailed alternative distribution: the exponential. Of the alternatives considered, all but the power-law with a cut-off make use of the "non-nested" flavor of the Vuong (1989) likelihood ratio test. The results are reported in tables 2.9 through 2.14 and summarized graphically by figures 2.11 and 2.12.

Both the exponential distribution and the stretched exponential distribution are inferior

[23] Most studies fail to assess (or at least report) results of any goodness-of-fit tests. Studies that use OLS to estimate the scaling exponent often report the R^2 from the regression as a measure of goodness-of-fit, however, as discussed in section 2.3 above (and in Clauset et al. (2009)) the R^2 is often a misleading indicator of goodness-of-fit for the power law model.

[24] Recall that the simulated KS distances were generated under the null hypothesis that the best-fit power law was the "correct" model for the data.

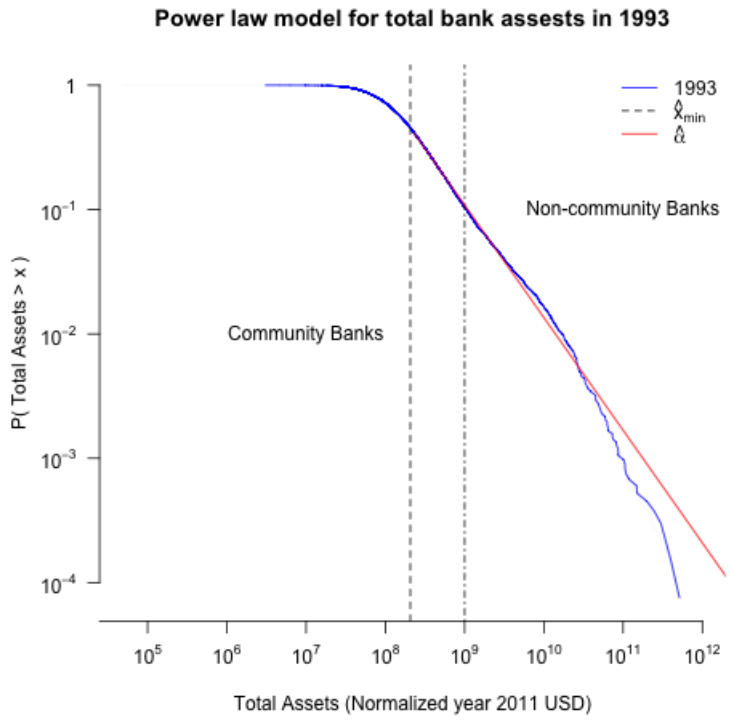

Figure 2.9: In the 1990s the best-fit power law distribution is rejected because it over predicts the number of large banks.

to the power law model. The data are sufficient to distinguish between the power law and both alternatives,[25] and, in general, the power law is statistically preferred to both the exponential and stretched exponential alternatives.[26] Given the data it is generally not possible to distinguish between the power law and the log-normal alternative.[27] Finally, while the power law distribution with an exponential cut-off is preferred over the pure power law model in the early 1990's, this distribution becomes less plausible over time. I reject the power law in favor of the truncated power law for most measures and in most years between 1992-1997. Starting in the late 1990's however, I can no longer reject the power law in favor of the truncated power law.

[25] Formally, I reject the two-sided null hypothesis of the Vuong (1989) test that the power law and either the exponential or the stretched exponential are equally far from the "true" distribution.

[26] Formally, I fail to reject the power law null hypothesis in the one-sided Vuong (1989) test.

[27] Remember that the log-normal alternative is fit to the tail and is not the log-normal distribution that we all know and love and that was definitively ruled out above.

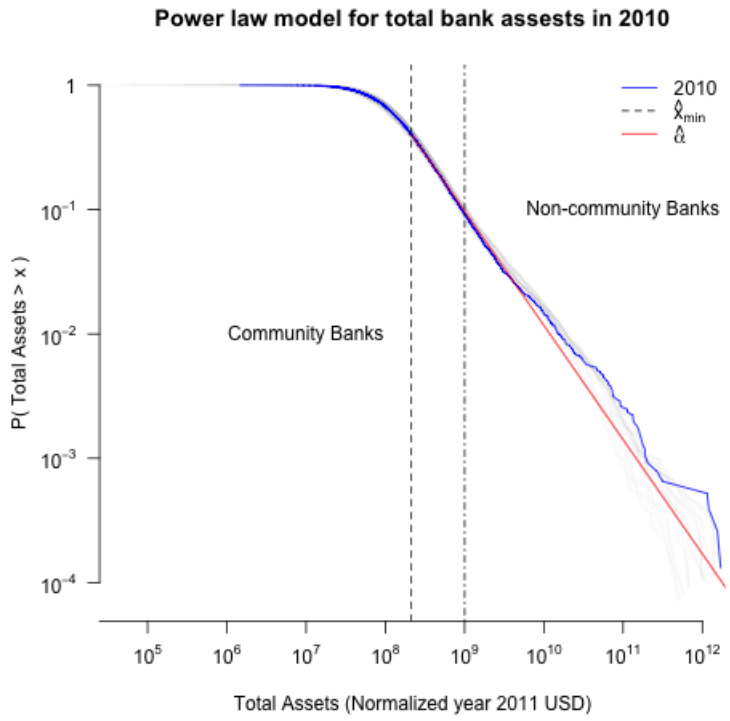

Figure 2.10: In the 2000s the best-fit power law distribution is rejected because it under predicts the number of large banks.

2.5 Conclusions

Most of the plausible benefits of banking sector concentration, such as increased profitability, economies of scale, and increased diversification accrue primarily to individual banks (and their shareholders). Many important costs of increased concentration, meanwhile, are social. These might include increased propensity to take risks due to moral hazard/adverse selection issues created by explicit deposit insurance arrangements as well as implicit bail-out arrangements for banks deemed "too big to fail," decreased competitiveness in the banking sector, etc. Characterizing the behavior of the upper tail of the size distribution of banks is important in order to quantitatively assess the potential costs and benefits from continued concentration of the U.S. banking sector.

The parameter estimates reported in tables 2.2 through 2.7 document statistically significant departures from Zipf's Law for most measures of bank size in most years. Estimated scaling exponents of $\hat{\alpha} \approx 1.9 < 2$ indicates that the upper tail of the empirical distribution of U.S. is too heavy (i.e., U.S. the banking sector is too concentrated) to be consistent with the predictions of Zipf's Law. While the goodness-of-fit tests document statistically

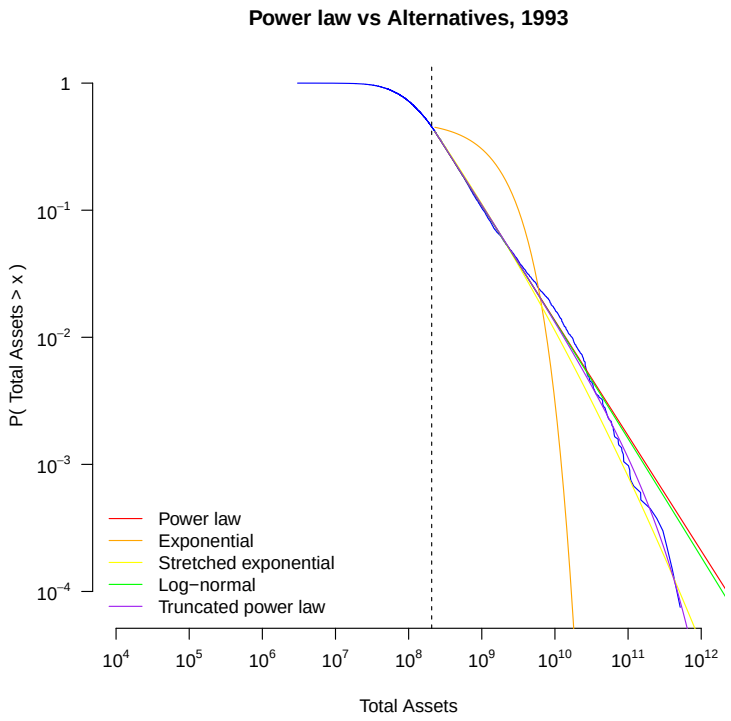

Figure 2.11: Pure power law model is rejected in favor of power law with exponential cut-off.

significant deviations from the power law model, it is not clear whether or not the observed deviations are economicially significant. Furthermore, the Vuong (1989) likelihood ratio testing results indicate that the power law model out-performs other distributions commonly used to model heavy-tailed data. I therefore conclude that a power law distribution with $\hat{\alpha} \approx 1.9$ is a reasonable null model for the upper tail of the size distribution of U.S. banks.

Power law distributions with scaling exponents $\alpha < 2$ have no well-defined expected values. The failure of the expected value to be well-defined implies that the fraction of U.S. banking sector totals in the top *anything* of the size distribution (even the top 1% of banks) should tend to unity in the limit of an infinite number of banks. In practice, a scaling exponent of $\alpha < 2$ implies that one should expect effectively all of the assets, loans, liabilities, deposits, and equity in the U.S. banking sector to be controlled by a small number of banks in the extreme tail of the size distribution.

The extreme levels of concentration within the U.S. banking sector has important implications for empirical and theoretical business cycle research. Gabaix (2011) posits that, so long as the upper tail of the firm size distribution is sufficiently heavy, much of the varia-

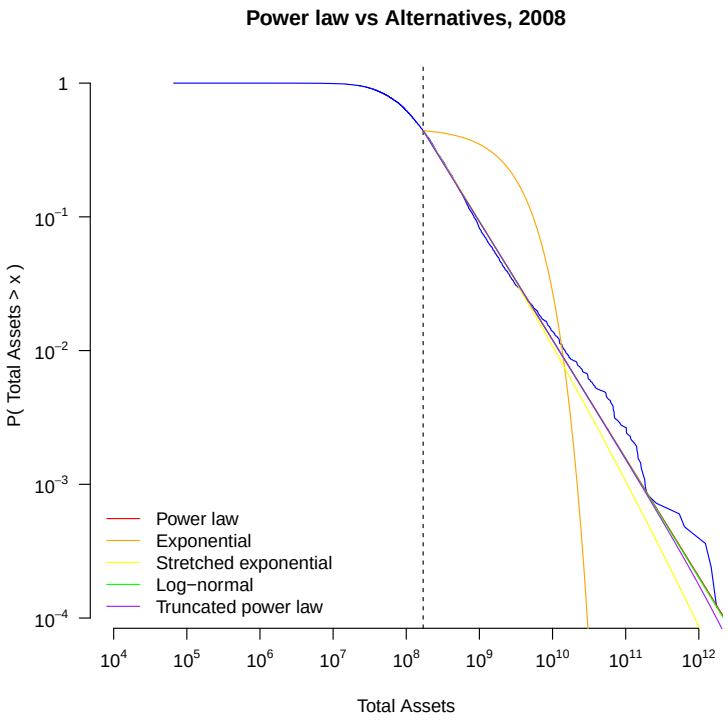

Figure 2.12: Power law is the preferred model, but LR tests unable to distinguish between plausible alternatives like log-normal.

tion in aggregate macroeconomic time series data over the business cycle can be explained by idiosyncratic shocks to individual firms. The "granularity hypothesis" of Gabaix (2011) constrasts starkly with existing research on business cycles which has focused almost exlcusively on the role played by aggregate shocks. Extending the Gabaix (2011) model to incorporate a banking sector in order to explore the theoretical and empirical support for idiosyncratic shocks to individual banks would seem to be a fruitful direction for future research.[28]

[28] Bremus et al. (2013), in a recent working paper, modify the Gabaix (2011) framework and explore idiosyncratic shocks to loan growth are a plausible source of aggregate fluctuations.

Appendix

2.A Power law primer

This appendix is a self-contained summary of basic theoretical results for the power law distribution and is intended to assist readers who may be unfamiliar with this literature to interpret the results of this paper.[29]

2.A.1 The mathematics of power laws

The density function, $p(x)$

Given that this paper deals with empirical measures of bank size that are (at least approximately) continuous, I restrict attention to the continuous version of the power law distribution described by a probability density function, $p(x)$ such that

$$p(x)\mathrm{d}x = \Pr\left(x \leq X < x + \mathrm{d}x\right) = Cx^{-\alpha}\mathrm{d}x \tag{2.13}$$

where X is the observed value (in this paper a normalized measure of bank size) and C is the normalization constant. Note that this density diverges as $x \to 0$, thus equation 2.13 cannot hold for all $x \geq 0$ and there must be some lower bound, x_{min}, for the power law distribution. To derive the normalization constant, simply integrate the density function

[29] Definitions and presentation of formal results mostly follow Newman (2005).

and set the result equal to one.

$$\begin{aligned}
1 &= \int_{x_{min}}^{\infty} p(x)\mathrm{d}x \\
1 &= \int_{x_{min}}^{\infty} Cx^{-\alpha}\mathrm{d}x \\
1 &= C\left[\frac{1}{-\alpha+1}x^{-\alpha+1}\Big|_{x_{min}}^{\infty}\right] \\
1 &= C\left[0 - \frac{1}{-\alpha+1}x_{min}^{-\alpha+1}\right] \\
C &= \left(\frac{\alpha-1}{x_{min}}\right)x_{min}^{\alpha}
\end{aligned} \tag{2.14}$$

After deriving the normalization constant, the density function, $p(x)$, can be written as

$$p(x) = \left(\frac{\alpha-1}{x_{min}}\right)\left(\frac{x_{min}}{x}\right)^{\alpha} \tag{2.15}$$

note that normalization requires $\alpha > 1$.

The distribution function, $P(x)$

The cumulative distribution function, $P(x)$, can be derived from integrating the density function derived above:

$$\begin{aligned}
P(X) = Pr(X \leq x) &= \int_{x_{min}}^{x} p(x')\mathrm{d}x' \\
&= \int_{x_{min}}^{x} \left(\frac{\alpha-1}{x_{min}}\right)\left(\frac{x_{min}}{x'}\right)^{\alpha}\mathrm{d}x' \\
&= \left(\frac{\alpha-1}{x_{min}}\right)x_{min}^{\alpha}\int_{x_{min}}^{x}\left(\frac{1}{x'}\right)^{\alpha}\mathrm{d}x' \\
&= \left(\frac{\alpha-1}{x_{min}}\right)x_{min}^{\alpha}\left[\frac{1}{-\alpha+1}x'^{-\alpha+1}\Big|_{x_{min}}^{x}\right] \\
&= (\alpha-1)x_{min}^{\alpha-1}\left[\frac{1}{-\alpha+1}x^{-\alpha+1} + \frac{1}{\alpha-1}x_{min}^{-(\alpha-1)}\right] \\
&= 1 - \left(\frac{x_{min}}{x}\right)^{-(\alpha-1)}
\end{aligned} \tag{2.16}$$

The survival function, $1 - P(x)$

One of the most useful properties of the power law distribution is that the survival function (sometimes also referred to as the upper cumulative distribution function) also follows a power law:

$$Pr(X > x) = 1 - P(x) = \left(\frac{x_{min}}{x}\right)^{-(\alpha-1)} \tag{2.17}$$

Note that the power law scaling exponent of the survival function is $\alpha - 1$, which is one less than the scaling exponent of the power law density function.

The moment generating function

The expected value of the power law distribution satisfies:

$$\begin{aligned}
E[x] &= \int_{x_{min}}^{\infty} x p(x) \mathrm{d}x \\
&= \int_{x_{min}}^{\infty} x \left(\frac{\alpha-1}{x_{min}}\right) \left(\frac{x_{min}}{x}\right)^{\alpha} \mathrm{d}x \\
&= (\alpha-1) x_{min}^{\alpha-1} \int_{x_{min}}^{\infty} x^{-\alpha+1} \mathrm{d}x \\
&= (\alpha-1) x_{min}^{\alpha-1} \left[\frac{1}{-\alpha+2} x^{-\alpha+2} \Big|_{x_{min}}^{\infty}\right] \\
&= (\alpha-1) x_{min}^{\alpha-1} \left[0 - \frac{1}{-\alpha+2} x_{min}^{-\alpha+2}\right] \\
&= \left(\frac{\alpha-1}{\alpha-2}\right) x_{min}
\end{aligned} \tag{2.18}$$

Note that this expression is only defined for $\alpha > 2$: if $\alpha \leq 2$ then the expected value of the power law distribution is infinite.

In general, the k^{th} moment of the power law distribution can be calculated as follows:

$$\begin{aligned}
E[x^k] &= \int_{x_{min}}^{\infty} x^k p(x)\mathrm{d}x \\
&= \int_{x_{min}}^{\infty} x^k \left(\frac{\alpha-1}{x_{min}}\right)\left(\frac{x_{min}}{x}\right)^{\alpha} \\
&= (\alpha-1)x_{min}^{\alpha-1}\int_{x_{min}}^{\infty} x^{-\alpha+k}\mathrm{d}x \\
&= (\alpha-1)x_{min}^{\alpha-1}\left[\frac{1}{-\alpha+k+1}x^{-\alpha+k+1}\Big|_{x_{min}}^{\infty}\right] \\
&= (\alpha-1)x_{min}^{\alpha-1}\left[0-\frac{1}{-\alpha+k+1}x_{min}^{-\alpha+k+1}\right] \\
&= \left(\frac{\alpha-1}{\alpha-k-1}\right)x_{min}^{k}
\end{aligned} \tag{2.19}$$

Expected largest value, $E[x_{max}]$

Suppose we draw n observations from a power law distribution. What value is the largest of those measurements likely to take? More precisely, what is the probability $\pi(x)\mathrm{d}x$ that the largest observation falls in the interval between x and $x+\mathrm{d}x$?

From the density function we know that the probability that observation i lies between x and $x+\mathrm{d}x$ is $p(x)$; and using the distribution function for the power law the probability that this observation is the largest of the n total observations is $P(x)^{n-1}$ (so long as $\alpha > 1$). Combining these two observations with the fact that there are n ways to choose the initial observation i yields an expression for $\pi(x)$:

$$\pi(x) = np(x)P(x)^{n-1} \tag{2.20}$$

To calculate the expected value of the largest observation, $E[x_{max}]$ we need to evaluate the following integral:

$$\begin{aligned}
E[x_{max}] &= \int_{x_{min}}^{\infty} x\pi(x)\mathrm{d}x \\
&= n\int_{x_{min}}^{\infty} xp(x)P(x)^{n-1}\mathrm{d}x \\
&= n(\alpha-1)\int_{x_{min}}^{\infty}\left(\frac{x_{min}}{x}\right)^{\alpha-1}\left[1-\left(\frac{x_{min}}{x}\right)^{\alpha-1}\right]^{n-1}\mathrm{d}x
\end{aligned}$$

Using a change of variables, $y = 1 - \left(\frac{x_{min}}{x}\right)^{\alpha-1}$, the above integral becomes:

$$\begin{aligned} E[x_{max}] =& n x_{min} \int_0 1 \frac{y^{n-1}}{(1-y)^{\frac{1}{\alpha-1}}} \mathrm{d}y \\ =& n x_{min} B\left(n, \frac{\alpha-2}{\alpha-1}\right) \end{aligned} \tag{2.21}$$

where

$$B\left(n, \frac{\alpha-2}{\alpha-1}\right)$$

is Legendre's beta-function. The beta function has the interesting property that for large values of its arguments the function follows a power law. Specifically, for large n

$$B\left(n, \frac{\alpha-2}{\alpha-1}\right) \sim n^{\frac{\alpha-2}{\alpha-1}}$$

and therefore

$$E[x_{max}] \sim n^{\frac{1}{\alpha-1}} \tag{2.22}$$

This result implies that as the sample size becomes larger the expected value for the largest observation increases.[30]

Top heavy distributions and the 80/20 rule

For any power law distribution with $\alpha > 1$ the median, x_{median} of the distribution is well-defined and satisfies the following equation:

$$\int_{x_{median}}^{\infty} p(x)\mathrm{d}x = \frac{1}{2} \int_{x_{median}}^{\infty} p(x)\mathrm{d}x$$

Working through the algebra yields the following expression for x_{median}:

$$\begin{aligned} \left(\frac{\alpha-1}{x_{min}}\right) x_{min}^{\alpha} \left[\frac{1}{-\alpha+1} x^{-\alpha+1}\Big|_{x_{median}}^{\infty}\right] =& \frac{1}{2}\left(\frac{\alpha-1}{x_{min}}\right) x_{min}^{\alpha} \left[\frac{1}{-\alpha+1} x^{-\alpha+1}\Big|_{x_{min}}^{\infty}\right] \\ \left[0 - \frac{1}{-\alpha+1} x_{median}^{-\alpha+1}\right] =& \frac{1}{2}\left[0 - \frac{1}{-\alpha+1} x_{min}^{-\alpha+1}\right] \\ x_{median}^{-\alpha+1} =& \frac{1}{2} x_{min}^{-\alpha+1} \\ x_{median} =& 2^{\frac{1}{\alpha-1}} x_{min} \end{aligned} \tag{2.23}$$

[30] Crucially, the n in this formula refers to the number of observations in the power law tail only, and not to the total sample size.

In the context of this paper, where $\hat{\alpha} > 1$ in every instance, there will be some well defined median bank size that divides the largest 50% of banks from the smallest 50% of banks. In addition, the estimates of $\hat{\alpha} \approx 2$ in which case the above formula says that the median bank should be roughly twice as large as the smallest bank (which still in the tail).

We might also like to ask what fraction of total U.S. banking sector assets, loans, liabilities, deposits, equity or employees are controlled by the largest 50% of banks? Or even more generally, what fraction of U.S. banking sector totals is controlled by banks whose size exceeds x? The fraction of U.S. banking sector totals controlled by the largest 50% of banks will satisfy:

$$\frac{\int_{x_{median}}^{\infty} xp(x)\mathrm{d}x}{\int_{x_{min}}^{\infty} xp(x)\mathrm{d}x}$$

Making use of the formulas derived above for $E[x]$ and x_{median} (and a bit more algebra!) yields the following:

$$\begin{aligned}
&= \frac{(\alpha-1)x_{min}^{\alpha-1}\left[\frac{1}{-\alpha+2}x^{-\alpha+2}\Big|_{x_{median}}^{\infty}\right]}{\left(\frac{\alpha-2}{\alpha-1}\right)x_{min}} \\
&= \frac{(\alpha-1)x_{min}^{\alpha-1}\left(\frac{1}{\alpha-2}x_{median}^{-\alpha+2}\right)}{\left(\frac{\alpha-1}{\alpha-2}\right)x_{min}} \\
&= \left(\frac{x_{min}}{x_{median}}\right)^{\alpha-2} \\
&= 2^{-\left(\frac{\alpha-2}{\alpha-1}\right)}
\end{aligned} \tag{2.24}$$

More generally, the formula for the fraction of U.S. banking sector totals controlled by banks whose size exceeds x will satisfy:

$$\frac{\int_{x}^{\infty} x'p(x')\mathrm{d}x'}{\int_{x_{min}}^{\infty} xp(x)\mathrm{d}x} = \left(\frac{x}{x_{median}}\right)^{\alpha-2} \tag{2.25}$$

Crucially, in order for these results to be valid, $E[x]$ must be well defined which, in turn, requires that $\alpha > 2$. Unfortunately, as mentioned above, the parameter estimates obtained for the scaling exponents of the U.S. bank distribution fail to satisfy this basic requirement for almost all measures of size in almost every year for which I have data. This implies that the distribution of U.S. banks is *extremely* top-heavy: the fraction of U.S. banking sector totals in the top *anything* of the size distribution (even the top 1% of banks) tends to 1! Put another way, effectively all of the assets, loans, liabilities, deposits, and equity in the U.S. banking sector are controlled by a small number of banks in the extreme tail

of the size distribution.

Scale invariance

A power law distribution is also sometimes called a scale-free distribution. In fact the power law is the only probability distribution that is scale free or scale invariant. Roughly speaking, by scale invariant I simply mean that changing the units of the variable (i.e., re-scaling) does not alter the functional form of the distribution. Technically, a distribution function $p(x)$ is scale-invariant if and only if it satisfies

$$p(bx) = g(b)p(x) \tag{2.26}$$

The proof of the claim this claim follows Newman (2005). Start by setting $x = 1$. This implies that

$$g(b) = \frac{p(b)}{p(1)}$$

and then we can re-write the scale-free condition as:

$$p(bx) = \frac{p(b)}{p(1)}p(x)$$

Since this equation must hold for any value b, we can differentiate both sides of the above equation with respect to b in order to get:

$$xp'(bx) = \frac{p'(b)}{p(1)}p(x)$$

Setting $b = 1$ yields

$$xp'(x) = \frac{p'(1)}{p(1)}p(x)$$

which is a simple first-order differential equation with solution:

$$\ln\ p(x) = \frac{p(1)}{p'(1)} \ln\ x + C \tag{2.27}$$

To find the constant, C, simply set $x = 1$ which yields $C = \ln\ p(1)$. Finally, exponentiation of both sides yields:

$$p(x) = p(1)x^{-\alpha} \tag{2.28}$$

where $\alpha = -\frac{p(1)}{p'(1)}$.

2.B Estimation results

2.B.1 Power law parameter estimates

	Assets					
Year	$\hat{\alpha}$	$\hat{x}_{min}$	D	N	n_{tail}	$\frac{n_{tail}}{N}$
1992	1.91	2.46e+08	0.01139	13973	5445	0.39
	(1.88,1.94)	(1.91e+08,3.15e+08)				
1993	1.91	2.07e+08	0.01263	13325	6054	0.45
	(1.88,1.94)	(1.78e+08,2.69e+08)				
1994	1.89	1.94e+08	0.01487	12644	5926	0.47
	(1.87,1.93)	(1.71e+08,2.21e+08)				
1995	1.88	1.88e+08	0.0153	12003	5828	0.49
	(1.86,1.91)	(1.7e+08,2.26e+08)				
1996	1.89	1.96e+08	0.01546	11480	5395	0.47
	(1.86,1.92)	(1.69e+08,2.49e+08)				
1997	1.9	1.91e+08	0.01295	10946	5065	0.46
	(1.86,1.94)	(1.71e+08,5.74e+08)				
1998	1.9	1.72e+08	0.01828	10484	5185	0.49
	(1.77,1.93)	(1.56e+08,1.28e+09)				
1999	1.9	1.83e+08	0.01798	10240	4797	0.47
	(1.78,1.93)	(1.63e+08,1.22e+09)				
2000	1.88	1.7e+08	0.01951	9920	4759	0.48
	(1.79,1.93)	(1.6e+08,8.82e+08)				
2001	1.9	2e+08	0.01959	9630	4168	0.43
	(1.81,1.94)	(1.62e+08,9.96e+08)				
2002	1.9	2.16e+08	0.01979	9369	3824	0.41
	(1.79,1.94)	(1.59e+08,1.15e+09)				
2003	1.89	2.05e+08	0.0232	9194	3873	0.42
	(1.79,1.94)	(1.57e+08,9.89e+08)				
2004	1.9	2.25e+08	0.02723	8988	3374	0.38
	(1.77,1.94)	(1.46e+08,1.07e+09)				
2005	1.89	2.12e+08	0.02574	8845	3469	0.39
	(1.75,1.94)	(1.6e+08,1.13e+09)				
2006	1.89	1.89e+08	0.02193	8691	3674	0.42
	(1.77,1.94)	(1.55e+08,1.22e+09)				
2007	1.89	1.77e+08	0.02047	8544	3669	0.43
	(1.78,1.94)	(1.53e+08,1.03e+09)				
2008	1.88	1.7e+08	0.02342	8314	3671	0.44
	(1.74,1.94)	(1.52e+08,1.36e+09)				
2009	1.91	2.03e+08	0.02012	8021	3374	0.42
	(1.86,1.97)	(1.68e+08,1.02e+09)				
2010	1.92	2.12e+08	0.01425	7667	3100	0.4
	(1.88,1.96)	(1.67e+08,2.47e+08)				
2011	1.9	1.82e+08	0.01906	7366	3351	0.45
	(1.85,1.96)	(1.52e+08,2.81e+08)				

Table 2.2: Estimates for α and x_{min} were obtained using the maximum likelihood methods advocated in Clauset et al. (2009). Reported 95% percentile confidence intervals were estimated using a non-parametric bootstrap with $B = 2500$ replications.

	Loans					
Year	$\hat{\alpha}$	$\hat{x}_{min}$	D	N	n_{tail}	$\frac{n_{tail}}{N}$
1992	1.88	1.55e+08	0.01002	13932	4386	0.31
	(1.85,1.92)	(1.12e+08,2.68e+08)				
1993	1.88	1.35e+08	0.009586	13299	4749	0.36
	(1.85,1.91)	(1.05e+08,1.78e+08)				
1994	1.88	1.26e+08	0.01197	12613	4868	0.39
	(1.85,1.91)	(1.05e+08,1.42e+08)				
1995	1.87	1.18e+08	0.01169	11975	4864	0.41
	(1.84,1.9)	(1.02e+08,1.46e+08)				
1996	1.87	1.12e+08	0.01401	11449	4811	0.42
	(1.83,1.9)	(9.81e+07,3.09e+08)				
1997	1.89	1.2e+08	0.01628	10902	4368	0.4
	(1.82,1.92)	(9.74e+07,3.92e+08)				
1998	1.89	1e+08	0.01876	10418	4734	0.45
	(1.78,1.92)	(9.45e+07,5.8e+08)				
1999	1.89	1.12e+08	0.01613	10172	4260	0.42
	(1.81,1.92)	(9.73e+07,5.22e+08)				
2000	1.89	1.19e+08	0.01894	9832	3917	0.4
	(1.81,1.93)	(9.27e+07,5.39e+08)				
2001	1.91	1.38e+08	0.01741	9547	3500	0.37
	(1.86,1.94)	(1.05e+08,1.6e+08)				
2002	1.89	1.15e+08	0.01954	9283	3945	0.42
	(1.85,1.94)	(1.01e+08,1.92e+08)				
2003	1.9	1.31e+08	0.02104	9109	3455	0.38
	(1.81,1.94)	(1.04e+08,5.67e+08)				
2004	1.9	1.29e+08	0.02245	8909	3362	0.38
	(1.8,1.94)	(1.02e+08,5.83e+08)				
2005	1.89	1.26e+08	0.02332	8770	3353	0.38
	(1.79,1.94)	(9.98e+07,6.83e+08)				
2006	1.9	1.36e+08	0.02375	8619	3112	0.36
	(1.78,1.95)	(1.1e+08,8.6e+08)				
2007	1.82	5.98e+08	0.02045	8474	730	0.086
	(1.77,1.95)	(9.95e+07,8.57e+08)				
2008	1.82	7.7e+08	0.01827	8248	597	0.072
	(1.78,1.95)	(1.05e+08,8.59e+08)				
2009	1.91	1.42e+08	0.02205	7963	3114	0.39
	(1.8,1.98)	(1.19e+08,8.35e+08)				
2010	1.92	1.33e+08	0.01755	7615	3071	0.4
	(1.83,1.97)	(1.07e+08,7.58e+08)				
2011	1.79	7.78e+08	0.02133	7319	511	0.07
	(1.75,1.95)	(9.95e+07,8.7e+08)				

Table 2.3: Estimates for α and x_{min} were obtained using the maximum likelihood methods advocated in Clauset et al. (2009). Reported 95% percentile confidence intervals were estimated using a non-parametric bootstrap with $B = 2500$ replications.

	Liabilities					
Year	$\hat{\alpha}$	$\hat{x}_{min}$	D	N	n_{tail}	$\frac{n_{tail}}{N}$
1992	1.9	2.17e+08	0.01051	13971	5395	0.39
	(1.88,1.94)	(1.65e+08,2.84e+08)				
1993	1.9	1.86e+08	0.01301	13325	5897	0.44
	(1.87,1.93)	(1.53e+08,2.35e+08)				
1994	1.89	1.76e+08	0.01517	12644	5740	0.45
	(1.86,1.92)	(1.44e+08,1.97e+08)				
1995	1.88	1.71e+08	0.01415	12003	5625	0.47
	(1.85,1.91)	(1.47e+08,1.97e+08)				
1996	1.88	1.67e+08	0.01702	11479	5432	0.47
	(1.84,1.92)	(1.41e+08,5.75e+08)				
1997	1.9	1.65e+08	0.01397	10946	5075	0.46
	(1.84,1.93)	(1.44e+08,5.84e+08)				
1998	1.8	1.14e+09	0.01712	10484	783	0.075
	(1.77,1.93)	(1.4e+08,1.15e+09)				
1999	1.89	1.55e+08	0.01851	10240	4901	0.48
	(1.78,1.92)	(1.42e+08,1.05e+09)				
2000	1.88	1.57e+08	0.02026	9918	4548	0.46
	(1.77,1.92)	(1.38e+08,8.76e+08)				
2001	1.9	1.74e+08	0.01968	9630	4184	0.43
	(1.8,1.93)	(1.43e+08,8.93e+08)				
2002	1.9	1.9e+08	0.02026	9367	3806	0.41
	(1.8,1.94)	(1.4e+08,1.03e+09)				
2003	1.89	1.79e+08	0.02315	9192	3886	0.42
	(1.81,1.94)	(1.37e+08,8.48e+08)				
2004	1.89	1.87e+08	0.026	8987	3551	0.4
	(1.79,1.94)	(1.28e+08,9.26e+08)				
2005	1.9	2.1e+08	0.0263	8845	3169	0.36
	(1.75,1.94)	(1.39e+08,1.1e+09)				
2006	1.89	1.7e+08	0.02132	8691	3642	0.42
	(1.76,1.94)	(1.44e+08,1.1e+09)				
2007	1.89	1.6e+08	0.01999	8544	3597	0.42
	(1.77,1.94)	(1.3e+08,1.03e+09)				
2008	1.91	1.83e+08	0.02199	8314	3151	0.38
	(1.76,1.94)	(1.41e+08,1.18e+09)				
2009	1.93	2.08e+08	0.01916	8021	3033	0.38
	(1.86,1.97)	(1.5e+08,3.6e+08)				
2010	1.92	1.9e+08	0.01453	7667	3103	0.4
	(1.88,1.96)	(1.48e+08,2.27e+08)				
2011	1.92	1.9e+08	0.0195	7366	2964	0.4
	(1.78,1.96)	(1.41e+08,9.97e+08)				

Table 2.4: Estimates for α and x_{min} were obtained using the maximum likelihood methods advocated in Clauset et al. (2009). Reported 95% percentile confidence intervals were estimated using a non-parametric bootstrap with $B = 2500$ replications.

	Deposits					
Year	$\hat{\alpha}$	$\hat{x}_{min}$	D	N	n_{tail}	$\frac{n_{tail}}{N}$
1992	1.95	2.47e+08	0.01105	13945	4617	0.33
	(1.91,1.97)	(1.75e+08,3.26e+08)				
1993	1.95	2.05e+08	0.01062	13317	5327	0.4
	(1.92,1.98)	(1.69e+08,2.73e+08)				
1994	1.93	1.86e+08	0.01477	12635	5513	0.44
	(1.91,1.97)	(1.6e+08,2.41e+08)				
1995	1.93	1.83e+08	0.01278	11997	5424	0.45
	(1.9,1.96)	(1.64e+08,2.3e+08)				
1996	1.94	1.97e+08	0.01367	11469	4871	0.42
	(1.9,1.97)	(1.67e+08,2.51e+08)				
1997	1.95	1.97e+08	0.01261	10934	4553	0.42
	(1.91,1.99)	(1.63e+08,5.73e+08)				
1998	1.95	1.86e+08	0.01617	10472	4545	0.43
	(1.84,2)	(1.57e+08,1.09e+09)				
1999	1.94	1.71e+08	0.01708	10229	4751	0.46
	(1.84,1.99)	(1.57e+08,1.04e+09)				
2000	1.94	1.85e+08	0.01794	9903	4197	0.42
	(1.84,1.99)	(1.58e+08,1.01e+09)				
2001	1.95	1.92e+08	0.01809	9625	4027	0.42
	(1.8,1.99)	(1.63e+08,1.16e+09)				
2002	1.95	2.14e+08	0.02193	9366	3567	0.38
	(1.8,2)	(1.6e+08,1.28e+09)				
2003	1.94	1.93e+08	0.02425	9189	3782	0.41
	(1.8,2)	(1.55e+08,1.4e+09)				
2004	1.95	1.99e+08	0.02624	8986	3481	0.39
	(1.8,2)	(1.52e+08,1.13e+09)				
2005	1.95	2.16e+08	0.02561	8843	3161	0.36
	(1.8,2)	(1.62e+08,1.07e+09)				
2006	1.94	1.84e+08	0.02232	8688	3498	0.4
	(1.81,2)	(1.64e+08,1.11e+09)				
2007	1.94	1.78e+08	0.01993	8541	3418	0.4
	(1.82,1.99)	(1.58e+08,9.9e+08)				
2008	1.95	1.85e+08	0.02201	8312	3197	0.38
	(1.8,1.99)	(1.58e+08,9.16e+08)				
2009	1.95	1.81e+08	0.0178	8020	3257	0.41
	(1.9,2.01)	(1.52e+08,5.9e+08)				
2010	1.95	1.83e+08	0.01493	7664	3105	0.41
	(1.91,1.99)	(1.51e+08,2.31e+08)				
2011	1.95	1.8e+08	0.02011	7363	2933	0.4
	(1.89,1.99)	(1.41e+08,2.34e+08)				

Table 2.5: Estimates for α and x_{min} were obtained using the maximum likelihood methods advocated in Clauset et al. (2009). Reported 95% percentile confidence intervals were estimated using a non-parametric bootstrap with $B = 2500$ replications.

	Equity					
Year	$\hat{\alpha}$	$\hat{x}_{min}$	D	N	n_{tail}	$\frac{n_{tail}}{N}$
1992	1.96	3.41e+07	0.01225	13839	5229	0.38
	(1.92,1.99)	(2.7e+07,7.32e+07)				
1993	1.95	3.31e+07	0.01359	13220	5104	0.39
	(1.91,1.98)	(2.57e+07,6.93e+07)				
1994	1.94	3.36e+07	0.01457	12604	4745	0.38
	(1.9,1.97)	(2.48e+07,3.77e+07)				
1995	1.92	2.8e+07	0.0121	11972	5405	0.45
	(1.89,1.95)	(2.41e+07,4.78e+07)				
1996	1.93	3.15e+07	0.01332	11454	4635	0.4
	(1.9,1.96)	(2.52e+07,4.1e+07)				
1997	1.93	2.67e+07	0.01257	10922	4915	0.45
	(1.9,1.97)	(2.41e+07,5.19e+07)				
1998	1.94	2.67e+07	0.01604	10463	4590	0.44
	(1.81,1.97)	(2.11e+07,1.25e+08)				
1999	1.93	2.85e+07	0.01673	10219	4096	0.4
	(1.83,1.96)	(2.11e+07,1.32e+08)				
2000	1.92	2.46e+07	0.01959	9903	4444	0.45
	(1.8,1.95)	(1.86e+07,1.25e+08)				
2001	1.9	2.2e+07	0.01904	9611	4544	0.47
	(1.86,1.95)	(1.83e+07,2.98e+07)				
2002	1.91	2.32e+07	0.01834	9354	4280	0.46
	(1.85,1.95)	(2e+07,7.27e+07)				
2003	1.91	2.33e+07	0.01906	9180	4173	0.45
	(1.87,1.94)	(1.9e+07,2.72e+07)				
2004	1.88	1.7e+07	0.02155	8976	4630	0.52
	(1.85,1.93)	(1.5e+07,2.23e+07)				
2005	1.88	1.7e+07	0.02233	8833	4516	0.51
	(1.84,1.93)	(1.5e+07,2.9e+07)				
2006	1.89	1.73e+07	0.02357	8679	4323	0.5
	(1.72,1.94)	(1.46e+07,1.78e+08)				
2007	1.89	1.68e+07	0.02195	8534	4341	0.51
	(1.74,1.94)	(1.56e+07,1.35e+08)				
2008	1.91	1.86e+07	0.02156	8298	4093	0.49
	(1.75,1.95)	(1.56e+07,1.32e+08)				
2009	1.9	1.71e+07	0.0201	7997	3884	0.49
	(1.86,1.95)	(1.4e+07,2.38e+07)				
2010	1.9	1.79e+07	0.02015	7648	3533	0.46
	(1.85,1.93)	(1.37e+07,2.03e+07)				
2011	1.88	1.67e+07	0.02222	7352	3705	0.5
	(1.84,1.93)	(1.41e+07,2.22e+07)				

Table 2.6: Estimates for α and x_{min} were obtained using the maximum likelihood methods advocated in Clauset et al. (2009). Reported 95% percentile confidence intervals were estimated using a non-parametric bootstrap with $B = 2500$ replications.

	Employees					
Year	$\hat{\alpha}$	$\hat{x}_{min}$	D	N	n_{tail}	$\frac{n_{tail}}{N}$
1992	1.99	7.55e+04	0.01403	13915	3398	0.24
	(1.95,2.02)	(4.6e+04,9.31e+04)				
1993	1.99	6.83e+04	0.01422	13264	3650	0.28
	(1.95,2.02)	(4.83e+04,1.11e+05)				
1994	1.99	6.13e+04	0.01607	12597	3974	0.32
	(1.95,2.02)	(3.97e+04,7.27e+04)				
1995	1.98	6.07e+04	0.01735	11960	3910	0.33
	(1.93,2.01)	(4.37e+04,1.37e+05)				
1996	1.99	6.29e+04	0.01559	11436	3656	0.32
	(1.93,2.02)	(4.23e+04,1.28e+05)				
1997	2	5.55e+04	0.01709	10907	3855	0.35
	(1.92,2.03)	(4.43e+04,1.82e+05)				
1998	1.91	1.85e+05	0.01939	10449	1019	0.098
	(1.87,2.03)	(4.07e+04,2.15e+05)				
1999	1.9	1.92e+05	0.01925	10206	956	0.094
	(1.85,2.03)	(4.32e+04,2.88e+05)				
2000	2	5.06e+04	0.0206	9889	3687	0.37
	(1.91,2.04)	(3.96e+04,1.89e+05)				
2001	1.99	5.04e+04	0.01819	9598	3597	0.37
	(1.93,2.03)	(4.07e+04,1.41e+05)				
2002	1.98	4.81e+04	0.01901	9335	3652	0.39
	(1.87,2.03)	(3.87e+04,2.47e+05)				
2003	1.99	5.05e+04	0.02039	9162	3490	0.38
	(1.82,2.04)	(4.02e+04,3.15e+05)				
2004	2	5.23e+04	0.01987	8958	3286	0.37
	(1.92,2.05)	(4.22e+04,1.91e+05)				
2005	2.03	6.08e+04	0.0197	8814	2862	0.32
	(1.84,2.07)	(4.71e+04,2.88e+05)				
2006	2.02	5.54e+04	0.01814	8661	3027	0.35
	(1.87,2.07)	(4.68e+04,2.67e+05)				
2007	2.02	5.23e+04	0.0181	8515	3157	0.37
	(1.91,2.07)	(4.66e+04,2.55e+05)				
2008	2.03	5.88e+04	0.01562	8284	2835	0.34
	(1.98,2.09)	(4.8e+04,8.03e+04)				
2009	2.04	6.23e+04	0.01468	7996	2681	0.34
	(1.99,2.1)	(5.21e+04,1e+05)				
2010	2.03	5.85e+04	0.01494	7639	2703	0.35
	(1.97,2.07)	(5.05e+04,1.15e+05)				
2011	2.02	5.6e+04	0.0146	7341	2729	0.37
	(1.97,2.07)	(5.1e+04,1.13e+05)				

Table 2.7: Estimates for α and x_{min} were obtained using the maximum likelihood methods advocated in Clauset et al. (2009). Reported 95% percentile confidence intervals were estimated using a non-parametric bootstrap with $B = 2500$ replications.

2.B.2 Goodness-of-fit and likelihood ratio test results

Goodness-of-fit tests

Year	asset	lnlsnet	liab	dep	eqtot	numemp
1992	0.01139**	0.01002	0.01051*	0.01105*	0.01225**	0.01403**
	(0.01092)	(0.01195)	(0.01087)	(0.0119)	(0.01108)	(0.01397)
1993	0.01263***	0.009586	0.01301***	0.01062*	0.01359***	0.01422**
	(0.0106)	(0.01168)	(0.01075)	(0.01105)	(0.01084)	(0.01367)
1994	0.01487***	0.01197**	0.01517***	0.01477***	0.01457***	0.01607***
	(0.01079)	(0.0119)	(0.01092)	(0.01105)	(0.01148)	(0.01346)
1995	0.0153***	0.01169**	0.01415***	0.01278**	0.0121**	0.01735***
	(0.01066)	(0.0117)	(0.01102)	(0.01129)	(0.01113)	(0.01358)
1996	0.01546***	0.01401***	0.01702***	0.01367***	0.01332**	0.01559**
	(0.01113)	(0.01171)	(0.0112)	(0.01156)	(0.01165)	(0.01403)
1997	0.01295**	0.01628***	0.01397***	0.01261**	0.01257**	0.01709***
	(0.01146)	(0.01242)	(0.01166)	(0.01225)	(0.01155)	(0.0135)
1998	0.01828***	0.01876***	0.01712	0.01617***	0.01604***	0.01939
	(0.0116)	(0.01202)	(0.02239)	(0.01216)	(0.01223)	(0.02238)
1999	0.01798***	0.01613***	0.01851***	0.01708***	0.01673***	0.01925
	(0.01192)	(0.0117)	(0.01171)	(0.01203)	(0.01238)	(0.02366)
2000	0.01951***	0.01894***	0.02026***	0.01794***	0.01959***	0.0206***
	(0.01198)	(0.0132)	(0.01211)	(0.01246)	(0.01207)	(0.01398)
2001	0.01959***	0.01741***	0.01968***	0.01809***	0.01904***	0.01819***
	(0.01287)	(0.01388)	(0.01287)	(0.01287)	(0.01209)	(0.01408)
2002	0.01979***	0.01954***	0.02026***	0.02193***	0.01834***	0.01901***
	(0.01303)	(0.01303)	(0.01346)	(0.01334)	(0.01278)	(0.01394)
2003	0.0232***	0.02104***	0.02315***	0.02425***	0.01906***	0.02039***
	(0.01304)	(0.01355)	(0.01308)	(0.01331)	(0.01265)	(0.01432)
2004	0.02723***	0.02245***	0.026***	0.02624***	0.02155***	0.01987***
	(0.0137)	(0.01403)	(0.01342)	(0.01395)	(0.01225)	(0.01469)
2005	0.02574***	0.02332***	0.0263***	0.02561***	0.02233***	0.0197***
	(0.01377)	(0.01399)	(0.0141)	(0.01434)	(0.01215)	(0.01544)
2006	0.02193***	0.02375***	0.02132***	0.02232***	0.02357***	0.01814***
	(0.01365)	(0.01433)	(0.01369)	(0.01378)	(0.01258)	(0.01494)
2007	0.02047***	0.02045	0.01999***	0.01993***	0.02195***	0.0181***
	(0.01335)	(0.02593)	(0.01363)	(0.01422)	(0.01253)	(0.01483)
2008	0.02342***	0.01827	0.02199***	0.02201***	0.02156***	0.01562**
	(0.01331)	(0.02639)	(0.01467)	(0.01444)	(0.01292)	(0.01538)
2009	0.02012***	0.02205***	0.01916***	0.0178***	0.0201***	0.01468*
	(0.01412)	(0.01453)	(0.01495)	(0.01431)	(0.01318)	(0.01561)
2010	0.01425*	0.01755***	0.01453**	0.01493**	0.02015***	0.01494*
	(0.01447)	(0.01465)	(0.0145)	(0.0145)	(0.01359)	(0.01571)
2011	0.01906***	0.02133	0.0195***	0.02011***	0.02222***	0.0146*
	(0.01447)	(0.02772)	(0.01467)	(0.01476)	(0.01342)	(0.01557)

Table 2.8: KS goodness-of-fit test statistics. Numbers in parentheses are the upper bound of a one-sided 95% confidence interval (the lower bound is zero in all cases). Confidence intervals were estimated using a non-parametric bootstrap with $B = 2500$ replications. Significance codes: *** $< 1\%$, ** $< 5\%$, * $< 10\%$.

	Assets										
	Exponential			Stretched Exponential			Log-normal			Truncated Power Law	
Year	Vuong	Two-sided	One-sided	Vuong	Two-sided	One-sided	Vuong	Two-sided	One-sided	LR	One-sided
1992	13.41	0	1	31	0	1	-0.5072	0.61	0.31	-4.62	0.0024
1993	13.99	0	1	2.431	0.015	0.99	-0.221	0.83	0.41	-3.492	0.0082
1994	13.97	0	1	1.945	0.052	0.97	0.1426	0.89	0.56	-2.992	0.014
1995	14.15	0	1	2.81	0.005	1	0.1874	0.85	0.57	-2.811	0.018
1996	12.77	0	1	32.58	0	1	0.7735	0.44	0.78	-1.666	0.068
1997	12.55	0	1	31.68	0	1	0.9966	0.32	0.84	-0.7065	0.23
1998	12.43	0	1	2.403	0.016	0.99	0.8569	0.39	0.8	-0.5402	0.3
1999	11.07	0	1	30.98	0	1	1.247	0.21	0.89	-0.4349	0.35
2000	11.55	0	1	3.292	0.001	1	0.8026	0.42	0.79	-0.4869	0.32
2001	10.97	0	1	3.521	0.00043	1	2.014	0.044	0.98	-0.2479	0.48
2002	10.61	0	1	28.44	0	1	1.694	0.09	0.95	-0.2172	0.51
2003	10.74	0	1	2.949	0.0032	1	1.383	0.17	0.92	-0.2187	0.51
2004	9.586	0	1	3.527	0.00042	1	1.252	0.21	0.89	-0.1259	0.62
2005	9.118	0	1	3.214	0.0013	1	1.469	0.14	0.93	-0.1527	0.58
2006	8.944	0	1	2.508	0.012	0.99	1.061	0.29	0.86	-0.1105	0.64
2007	9.021	0	1	26.06	0	1	1.322	0.19	0.91	-0.0759	0.7
2008	8.616	0	1	1.76	0.078	0.96	0.7255	0.47	0.77	-0.07625	0.7
2009	8.27	2.2e-16	1	2.077	0.038	0.98	1.128	0.26	0.87	-0.02975	0.81
2010	8.299	0	1	3.133	0.0017	1	1.384	0.17	0.92	-0.0154	0.86
2011	8.543	0	1	2.584	0.0098	1	0.8888	0.37	0.81	-0.02334	0.83

Table 2.9: Vuong (1989) likelihood ratio test results for the power law model against several alternatives. The reported test statistic (Vuong) for non-nested tests is the normalized likelihood ratio between the power law and the considered alternative. For the non-nested test the standard LR is reported. For non-nested tests both two-sided and a one-sided p-values are reported (note that the nested test is one-sided by construction).

	Loans										
	Exponential			Stretched Exponential			Log-normal			Truncated Power Law	
Year	Vuong	Two-sided	One-sided	Vuong	Two-sided	One-sided	Vuong	Two-sided	One-sided	LR	One-sided
1992	11.7	0	1	26.32	0	1	-1.256	0.21	0.1	-6.589	0.00028
1993	13.04	0	1	1.207	0.23	0.89	-0.439	0.66	0.33	-4.702	0.0022
1994	14.02	0	1	2.303	0.021	0.99	0.4063	0.68	0.66	-3.677	0.0067
1995	14.4	0	1	31.25	0	1	0.6267	0.53	0.73	-3.509	0.0081
1996	13.73	0	1	2.269	0.023	0.99	0.1822	0.86	0.57	-2.727	0.02
1997	13.33	0	1	2.761	0.0058	1	0.9603	0.34	0.83	-1.135	0.13
1998	13	0	1	2.553	0.011	0.99	0.8737	0.38	0.81	-0.7906	0.21
1999	10.84	0	1	29.21	0	1	0.8756	0.38	0.81	-0.7665	0.22
2000	11.24	0	1	29.18	0	1	0.7988	0.42	0.79	-0.4973	0.32
2001	11.41	0	1	28.06	0	1	1.783	0.075	0.96	-0.2913	0.45
2002	11.34	0	1	1.994	0.046	0.98	0.6312	0.53	0.74	-0.4088	0.37
2003	11.48	0	1	27.67	0	1	1.277	0.2	0.9	-0.2817	0.45
2004	10.71	0	1	3.359	0.00078	1	1.43	0.15	0.92	-0.2316	0.5
2005	10.1	0	1	2.61	0.009	1	0.6354	0.53	0.74	-0.2864	0.45
2006	9.452	0	1	2.155	0.031	0.98	0.8726	0.38	0.81	-0.1715	0.56
2007	7.502	6.3e-14	1	0.9333	0.35	0.82	0.2387	0.81	0.59	-0.7164	0.23
2008	6.478	9.3e-11	1	0.1913	0.85	0.58	-0.2697	0.79	0.39	-0.9388	0.17
2009	8.561	0	1	1.771	0.077	0.96	0.06066	0.95	0.52	-0.1233	0.62
2010	8.652	0	1	24.04	0	1	0.8677	0.39	0.81	-0.05547	0.74
2011	6.087	1.2e-09	1	0.5297	0.6	0.7	-0.1997	0.84	0.42	-0.8833	0.18

Table 2.10: Vuong (1989) likelihood ratio test results for the power law model against several alternatives. The reported test statistic (Vuong) for non-nested tests is the normalized likelihood ratio between the power law and the considered alternative. For the non-nested test the standard LR is reported. For non-nested tests both two-sided and a one-sided p-values are reported (note that the nested test is one-sided by construction).

	Liabilities										
	Exponential			Stretched Exponential			Log-normal			Truncated Power Law	
Year	Vuong	Two-sided	One-sided	Vuong	Two-sided	One-sided	Vuong	Two-sided	One-sided	LR	One-sided
1992	13.4	0	1	0.9621	0.34	0.83	-0.5804	0.56	0.28	-4.852	0.0018
1993	14.04	0	1	1.644	0.1	0.95	-0.2428	0.81	0.4	-3.641	0.007
1994	13.97	0	1	2.869	0.0041	1	0.3598	0.72	0.64	-2.835	0.017
1995	14.22	0	1	3.43	6e-04	1	0.6751	0.5	0.75	-2.546	0.024
1996	12.74	0	1	1.924	0.054	0.97	0.2702	0.79	0.61	-1.913	0.05
1997	12.57	0	1	2.845	0.0044	1	0.811	0.42	0.79	-0.721	0.23
1998	8.076	6.7e-16	1	5.4	6.7e-08	1	-0.9322	0.35	0.18	-3.173	0.012
1999	11.01	0	1	3.156	0.0016	1	0.9037	0.37	0.82	-0.5474	0.3
2000	11.51	0	1	30.78	0	1	1.18	0.24	0.88	-0.4377	0.35
2001	10.92	0	1	3.102	0.0019	1	1.639	0.1	0.95	-0.2822	0.45
2002	10.51	0	1	28.34	0	1	1.483	0.14	0.93	-0.2172	0.51
2003	10.62	0	1	2.808	0.005	1	1.268	0.2	0.9	-0.2036	0.52
2004	9.542	0	1	26.84	0	1	0.8733	0.38	0.81	-0.1559	0.58
2005	9	0	1	3.32	9e-04	1	1.722	0.085	0.96	-0.1226	0.62
2006	8.759	0	1	25.52	0	1	0.9071	0.36	0.82	-0.1011	0.65
2007	8.859	0	1	25.62	0	1	1.09	0.28	0.86	-0.07532	0.7
2008	8.519	0	1	3.439	0.00058	1	1.129	0.26	0.87	-0.02453	0.82
2009	8.105	4.4e-16	1	3.143	0.0017	1	1.495	0.13	0.93	-0.01361	0.87
2010	8.098	6.7e-16	1	3.154	0.0016	1	1.275	0.2	0.9	-0.01555	0.86
2011	8.339	0	1	3.353	8e-04	1	0.8946	0.37	0.81	-0.009701	0.89

Table 2.11: Vuong (1989) likelihood ratio test results for the power law model against several alternatives. The reported test statistic (Vuong) for non-nested tests is the normalized likelihood ratio between the power law and the considered alternative. For the non-nested test the standard LR is reported. For non-nested tests both two-sided and a one-sided p-values are reported (note that the nested test is one-sided by construction).

	Deposits										
	Exponential			Stretched Exponential			Log-normal			Truncated Power Law	
Year	Vuong	Two-sided	One-sided	Vuong	Two-sided	One-sided	Vuong	Two-sided	One-sided	LR	One-sided
1992	12.08	0	1	1.015	0.31	0.84	-0.5556	0.58	0.29	-4.504	0.0027
1993	12.81	0	1	1.638	0.1	0.95	-0.239	0.81	0.41	-3.254	0.011
1994	13.2	0	1	1.707	0.088	0.96	-0.3051	0.76	0.38	-2.914	0.016
1995	13.16	0	1	31.89	0	1	-0.1944	0.85	0.42	-2.787	0.018
1996	11.96	0	1	2.238	0.025	0.99	0.4465	0.66	0.67	-1.452	0.088
1997	11.96	0	1	2.452	0.014	0.99	0.8702	0.38	0.81	-0.5495	0.29
1998	11.27	0	1	2.266	0.023	0.99	0.6677	0.5	0.75	-0.3394	0.41
1999	9.877	0	1	2.475	0.013	0.99	0.5962	0.55	0.72	-0.3665	0.39
2000	10.17	0	1	2.515	0.012	0.99	1.291	0.2	0.9	-0.2224	0.5
2001	9.971	0	1	3.135	0.0017	1	1.191	0.23	0.88	-0.1493	0.58
2002	9.864	0	1	2.679	0.0074	1	1.535	0.12	0.94	-0.1268	0.61
2003	9.853	0	1	2.372	0.018	0.99	0.7537	0.45	0.77	-0.1264	0.62
2004	8.983	0	1	2.424	0.015	0.99	1.12	0.26	0.87	-0.06605	0.72
2005	8.466	0	1	23.54	0	1	0.8873	0.37	0.81	-0.06011	0.73
2006	8.43	0	1	2.292	0.022	0.99	0.7299	0.47	0.77	-0.05292	0.74
2007	8.421	0	1	24.68	0	1	1.298	0.19	0.9	-0.02566	0.82
2008	8.058	8.9e-16	1	3.217	0.0013	1	0.8709	0.38	0.81	-0.01056	0.88
2009	8.021	1.1e-15	1	23.63	0	1	0.772	0.44	0.78	-0.01636	0.86
2010	7.825	5.1e-15	1	3.042	0.0024	1	1.802	0.072	0.96	-0.009154	0.89
2011	8.222	2.2e-16	1	3.218	0.0013	1	2.26	0.024	0.99	-0.004497	0.92

Table 2.12: Vuong (1989) likelihood ratio test results for the power law model against several alternatives. The reported test statistic (Vuong) for non-nested tests is the normalized likelihood ratio between the power law and the considered alternative. For the non-nested test the standard LR is reported. For non-nested tests both two-sided and a one-sided p-values are reported (note that the nested test is one-sided by construction).

	Equity										
	Exponential			Stretched Exponential			Log-normal			Truncated Power Law	
Year	Vuong	Two-sided	One-sided	Vuong	Two-sided	One-sided	Vuong	Two-sided	One-sided	LR	One-sided
1992	6.087	1.2e-09	1	1.744	0.081	0.96	-0.111	0.91	0.46	-3.255	0.011
1993	6.087	1.2e-09	1	2.81	0.005	1	0.2897	0.77	0.61	-2.624	0.022
1994	6.087	1.2e-09	1	2.07	0.038	0.98	0.5127	0.61	0.7	-2.663	0.021
1995	6.087	1.2e-09	1	2.144	0.032	0.98	-0.171	0.86	0.43	-3.495	0.0082
1996	6.087	1.2e-09	1	2.018	0.044	0.98	0.5009	0.62	0.69	-1.817	0.057
1997	6.087	1.2e-09	1	2.556	0.011	0.99	0.8217	0.41	0.79	-0.8332	0.2
1998	6.087	1.2e-09	1	3.544	0.00039	1	1.971	0.049	0.98	-0.3741	0.39
1999	6.087	1.2e-09	1	3.61	0.00031	1	1.943	0.052	0.97	-0.3828	0.38
2000	6.087	1.2e-09	1	2.966	0.003	1	1.199	0.23	0.88	-0.362	0.39
2001	6.087	1.2e-09	1	2.254	0.024	0.99	0.9281	0.35	0.82	-0.4012	0.37
2002	6.087	1.2e-09	1	2.835	0.0046	1	1.527	0.13	0.94	-0.278	0.46
2003	6.087	1.2e-09	1	3.838	0.00012	1	1.332	0.18	0.91	-0.2181	0.51
2004	6.087	1.2e-09	1	2.854	0.0043	1	1.556	0.12	0.94	-0.1798	0.55
2005	6.087	1.2e-09	1	2.631	0.0085	1	0.6432	0.52	0.74	-0.2111	0.52
2006	6.087	1.2e-09	1	2.84	0.0045	1	0.8935	0.37	0.81	-0.119	0.63
2007	6.087	1.2e-09	1	2.356	0.018	0.99	1.133	0.26	0.87	-0.08918	0.67
2008	6.087	1.2e-09	1	3.218	0.0013	1	1.782	0.075	0.96	-0.0379	0.78
2009	6.087	1.2e-09	1	3.014	0.0026	1	1.814	0.07	0.97	-0.01907	0.85
2010	6.087	1.2e-09	1	3.504	0.00046	1	0.7673	0.44	0.78	-0.02134	0.84
2011	6.087	1.2e-09	1	2.39	0.017	0.99	1.103	0.27	0.86	-0.05592	0.74

Table 2.13: Vuong (1989) likelihood ratio test results for the power law model against several alternatives. The reported test statistic (Vuong) for non-nested tests is the normalized likelihood ratio between the power law and the considered alternative. For the non-nested test the standard LR is reported. For non-nested tests both two-sided and a one-sided p-values are reported (note that the nested test is one-sided by construction).

	Employees										
	Exponential			Stretched Exponential			Log-normal			Truncated Power Law	
Year	Vuong	Two-sided	One-sided	Vuong	Two-sided	One-sided	Vuong	Two-sided	One-sided	LR	One-sided
1992	10.25	0	1	1.287	0.2	0.9	-0.006794	0.99	0.5	-2.62	0.022
1993	11.06	0	1	2.389	0.017	0.99	0.7373	0.46	0.77	-2.029	0.044
1994	12.47	0	1	2.746	0.006	1	1.119	0.26	0.87	-1.619	0.072
1995	12.58	0	1	3.625	0.00029	1	0.769	0.44	0.78	-1.413	0.093
1996	12.09	0	1	2.996	0.0027	1	0.9672	0.33	0.83	-0.8779	0.19
1997	11.94	0	1	3.504	0.00046	1	1.37	0.17	0.91	-0.3105	0.43
1998	8.375	0	1	1.084	0.28	0.86	0.3416	0.73	0.63	-1.233	0.12
1999	7.114	1.1e-12	1	0.928	0.35	0.82	0.1658	0.87	0.57	-1.146	0.13
2000	10.11	0	1	3.433	6e-04	1	1.932	0.053	0.97	-0.06995	0.71
2001	10.04	0	1	3.144	0.0017	1	1.963	0.05	0.98	-0.0787	0.69
2002	9.987	0	1	2.678	0.0074	1	0.8847	0.38	0.81	-0.1033	0.65
2003	9.973	0	1	2.734	0.0063	1	1.065	0.29	0.86	-0.05286	0.75
2004	8.739	0	1	2.846	0.0044	1	1.992	0.046	0.98	-0.01854	0.85
2005	8.235	2.2e-16	1	3.718	2e-04	1	2.982	0.0029	1	-0.001679	0.95
2006	8.19	2.2e-16	1	2.731	0.0063	1	1.526	0.13	0.94	-0.002922	0.94
2007	8.219	2.2e-16	1	2.623	0.0087	1	1.797	0.072	0.96	-0.002394	0.94
2008	7.986	1.3e-15	1	2.949	0.0032	1	1.821	0.069	0.97	0.0001372	1
2009	7.777	7.5e-15	1	2.86	0.0042	1	0.6553	0.51	0.74	-4.472e-05	0.99
2010	7.132	9.9e-13	1	2.566	0.01	0.99	1.416	0.16	0.92	-0.0003178	0.98
2011	7.074	1.5e-12	1	2.239	0.025	0.99	1.206	0.23	0.89	-0.0001462	0.99

Table 2.14: Vuong (1989) likelihood ratio test results for the power law model against several alternatives. The reported test statistic (Vuong) for non-nested tests is the normalized likelihood ratio between the power law and the considered alternative. For the non-nested test the standard LR is reported. For non-nested tests both two-sided and a one-sided p-values are reported (note that the nested test is one-sided by construction).

Chapter 3

The wealth of cities and diminishing increasing returns

In this chapter I look at the relationship between output and city size. A well known result from the urban economics literature is that a monopolistically competitive market structure combined with internal increasing returns to scale can be used to generate log-linear relations between total and per capita output and population. I extend this theoretical framework to allow for a variable elasticity of substitution between factors of production in a manner similar to Zhelobodko et al. (2012). Using data on Metropolitan Statistical Areas (MSAs) in the U.S. I find evidence that supports what Zhelobodko et al. (2012) refer to as "increasing relative love for variety (RLV)." Increasing RLV generates pro-competitive effects as market size increases which means that IRS, whilst important for small to medium sized cities, are exhausted as cities become large. This has important policy implications as it suggests that focusing intervention on creating scale for small populations is potentially much more valuable than further investments to increase market size in the largest population centers.[1]

3.1 Introduction

In a series of recent articles, Bettencourt et al. (2007), Bettencourt and West (2010), Bettencourt et al. (2010), and Bettencourt (2013), have suggested that important socio-economic indicators display "super-linear" scaling (i.e., exhibit increasing returns to scale)

[1]This chapter is based on joint work with Dr. David Comerford. In addition to my individual contributions to our ongoing joint work, this chapter represents a unique and significant contribution to our research agenda.

with city size while indicators of city resource use display "sub-linear" scaling (i.e., exhibit decreasing returns to scale).[2] Both types of scaling can be captured by an elegant power law (i.e., log-linear) relation between a particular indicator or measure, X, and city size, L

$$X \propto L^b \tag{3.1}$$

where city size is typically measured by total population. Super-linear scaling (i.e., increasing returns to scale) requires $b > 1$ while sub-linear scaling (i.e., decreasing returns to scale) requires $b < 1$.

In this chapter I show how such scaling relationships can be supported as equilibrium outcomes within a theoretical framework, widely used in both urban economics and economic geography, that assumes markets are monopolistically competitive and that firms producing differentiated intermediate input goods face fixed costs of production. I then extend this benchmark model, which assumes a constant elasticity of substitution between intermediate inputs used in production, to allow for a variable elasticity of substitution between factors of production in a manner similar to Behrens and Murata (2007) and Zhelobodko et al. (2012). The extension demonstrates that, when the elasticity of substitution is no longer a fixed constant, returns to scale depend greatly on the size of a city. In particular, while output, income, wages, and productivity of small to medium sized cities exhibit increasing returns to scale, in larger cities these important socio-economic indicators tend to exhibit constant returns to scale.

I estimate the structural parameters of both our models using data on output, per capita output, and population from $N = 366$ U.S. Metropolitan Statistical Areas (MSAs) in 2010. My findings suggest that the elasticity of substitution is indeed increasing with the number of available inputs and, although I am not able to statistically reject our benchmark model in favor of our extension where the elasticity of substitution varies with the number of available inputs, our extension does have significantly higher predictive power than a model with constant elasticity of substitution both in and out of sample. The results have important policy implications as they suggest that focusing intervention on creating scale and economic integration for small and medium sized cities is potentially much more valuable than further investment to increase the size and integration of the largest population centers.

[2] Bettencourt et al. (2007) and Bettencourt et al. (2010) document purported "super-linear" scaling relations between socio-economic indicators such as gross metropolitan product (GMP), income, wages, number of patents filed, total employment in R&D, etc.; and "sub-linear" scaling between measures of resource usage such as the number of gasoline stations, total road surface, total length of electrical cables, etc. Bettencourt and West (2010) sketches the outlines of a formal model of city scaling, the details of which are worked out in Bettencourt (2013).

At this point it is worth emphasizing what I am not attempting to do in this paper. My current work is not an attempt to explain the existence or emergence of cities, nor does it seek to explain the distribution of people across cities.[3] My theoretical objective in this chapter is more modest. I wish to take the distribution of people across cities as given and seek to explore the implications of a variable elasticity of substitution for the functional form of the returns to scale between output, income, wages, and productivity. I view my analysis as a necessary precursor to a richer model.

The objective of the empirical analysis in this chapter is equally modest. Although my model is capable of generating sharp predictions about how the magnitude of the agglomeration effects changes with city size, the objective of my empirical work is simply to asses whether or not the available data support the functional relation between per capita output and total population predicted by my model. My empirical results are not intended to be interpreted as direct measures of agglomeration effects.

With these caveats in mind, the remainder of the chapter proceeds as follows. In the next section I review some of the related literature. In section 3.3, I develop the theoretical framework and discuss its contribution. In section 3.4, I discuss the data used in the analysis and assess the relative performance of two statistical models of the relationship between per capita output and city size which motivated us to consider models with a varying elasticity of substitution before discussing the structural estimation results. Section 3.5 offers some conclusions and discusses what I think to be promising directions for future research.

3.2 Related literature

My benchmark model descends directly from the seminal work on monopolistic competition of Spence (1976) and Dixit and Stiglitz (1977). Following Abdel-Rahman (1988), Fujita (1988), Rivera-Batiz (1988), Krugman (1991), Ciccone and Hall (1996), and Fujita et al. (1999), I focus on the role of product differentiation and fixed costs of production as the drivers of agglomeration economies (i.e., increasing returns to scale). Like Rivera-Batiz (1988) and Ciccone and Hall (1996) product differentiation in my model occurs on the

[3] It is difficult to explain the existence of large concentrations of people in a particular area without allowing people to choose the area in which to locate, and in my model, location is not a choice variable. Clearly people do move between cities, however movement between cities takes time and money and thus one justification for taking population as exogenous is that population levels are slow to respond to differentials in per capita income and productivity. Glaeser and Gyourko (2005) argue that durable housing may cause population adjustments in response to productivity shocks to be spread over several decades. In his detailed study of the "American dust bowl," Hornbeck (2012) argues that the large exogenous shock experienced by much of the American midwest in the 1930's (i.e., the widespread erosion of top-soil by extensive and repeated dust storms) was absorbed over several decades by mostly permanent migration.

supply side of the economy: firms facing fixed costs of production produce differentiated intermediate inputs which are then used in the production of a homogenous consumption good.[4] In such models the monopoly power of firms producing the differentiated intermediate inputs.[5][6]

The inspiration for the general model lies in the recent work of Behrens and Murata (2007) and Zhelobodko et al. (2012) on monopolistic competition with variable elasticity of substitution. The major theoretical contribution of this chapter is to extend the models of Rivera-Batiz (1988) and Ciccone and Hall (1996) to allow the elasticity of substitution between intermediate inputs used in the production of consumption goods to vary across cities depending on the number of available intermediate inputs. The variation of the elasticity of substitution with the number of available intermediate inputs creates a link between the size of a city and the market power of firms producing intermediate inputs in that city via the markup that these firms charge for the use of their products in the production of consumption goods. I show that if firms producing consumption goods find it easier to substitute between any two intermediate inputs when the number of available inputs is larger (i.e., if the elasticity of substitution is an increasing function of the number of available intermediate inputs), then firms producing intermediate inputs will be larger (in terms of both output and employment) and the markups charged by these firms will be lower in larger cities.[7] As the size of a city increases, the competitive pressure generated by falling markups erodes the market power of firms selling intermediate inputs which is the root source of increasing returns to city size. In the limit of large cities, the competitive pressure becomes so great that firms producing intermediate inputs are forced to equate

[4]The models of Abdel-Rahman (1988) and Fujita (1988) assume that product differentiation occurs on the demand side of the economy: households derive utility from consuming an aggregate consumption bundle consisting of differentiated consumption goods.

[5]The essence of standard models models of economic agglomeration with monopolistic competition, product differentiation, and fixed costs of production is summarized in a slightly different manner by Ciccone and Hall (1996): "when local markets are more active, a larger number of producers of the differentiated intermediate inputs break even. The production of final goods is more productive when a greater variety of intermediate inputs is available."

[6]A major strand of the urban economics literature to which my model is somewhat more distantly related are the spatial equilibrium models pioneered by Mills (1967), Rosen (1979), and Roback (1982). Models of spatial equilibrium allow for free movement of consumers across cities and then assume a kind of "no-arbitrage" condition which requires that the welfare, at least for the marginal consumer/migrant, is equalized across space (i.e., cities). In such spatial equilibrium models, output, income, and population of a city are jointly determined by prices (i.e., housing prices, wages, etc.) and productivity. See Glaeser (2008) and Glaeser and Gottlieb (2009) for a more modern treatment of the spatial equilibrium approach. My current model, contrary to the spirit of the spatial equilibrium approach, treats the population of a city as exogenous and fixed. In section 3.5 I briefly discuss some extensions of the model that would explicitly incorporate a notion of spatial equilibrium.

[7]Both of these predictions find empirical support. Syverson (2007) provides evidence that firms operating in larger cities charge lower markups. Campbell and Hopenhayn (2005) characterize the effects of market size on the size distribution of firms using data on retail trade establishments across 225 U.S. cities. In general they find that larger firms tend to exist in larger cities.

their price with their marginal costs of production and the production of consumption goods exhibits constant, rather than increasing, returns to city size.

3.3 Model

I begin by developing the benchmark model, which descends from Dixit and Stiglitz (1977) and Spence (1976) via Abdel-Rahman (1988), Fujita (1988), Rivera-Batiz (1988) and Ciccone and Hall (1996) where the elasticity of substitution between factors of production, σ, is constant. The benchmark model predicts a log-linear (i.e., power law scaling) relations between city size and total (and per capita) income. Such a relation generates unbounded growth in per capita income with city size consistent with increasing returns to scale.

Following Zhelobodko et al. (2012), I then develop a more general model in which the elasticity of substitution is itself an equilibrium outcome of the model that varies across cities depending on their size. The general model predicts an equilibrium relationship between city size and total (and per capita) output that is consistent with the limiting behavior of the logistic scaling model discussed in the preceding section.

3.3.1 Benchmark model with constant σ

I model the world as consisting of N distinct economic agglomerations which I will refer to as "cities" for simplicity. The production side of the economy in city $i = 1, \ldots, N$ consists of final goods producers and intermediate goods producers.

Final goods producers

All final goods producers in city i create a homogenous consumption good from a different number (formally a different measure) of intermediate goods n_i via the following constant returns to scale (CRTS), constant elasticity of substitution (CES) production function:

$$Y_i = \left[\int_0^{n_i} q_j^{\frac{\sigma-1}{\sigma}} dj \right]^{\frac{\sigma-1}{\sigma}} \tag{3.2}$$

where the parameter $1 < \sigma \leq \infty$ is the elasticity of substitution between the n_i different intermediate goods used in production.

Firms producing the final consumption good are assumed to choose their demand for the

various intermediate goods q_j in order to maximize profits.

$$\max_{\{q_j\}_{j=0}^{n_i}} \Pi_i = \left[\int_0^{n_i} q_j^{\frac{\sigma-1}{\sigma}} dj \right]^{\frac{\sigma}{\sigma-1}} - \int_0^{n_i} p_j q_j \tag{3.3}$$

Therefore the demand for each intermediate good j must satisfy the following first-order necessary condition:

$$\frac{\partial \Pi_i}{\partial q_j} \equiv q_j^{\frac{\sigma-1}{\sigma}-1} \left[\int_0^{n_i} q_j^{\frac{\sigma-1}{\sigma}} dj \right]^{\frac{\sigma}{\sigma-1}-1} - p_j = 0 \tag{3.4}$$

which yields the following standard demand function for good j:

$$q_j = p_j^{-\sigma} Y_i. \tag{3.5}$$

The demand for each intermediate good j of the final goods producers in city i varies directly with total output (which is also total revenue since the final consumption good is the numeraire) and inversely as the σ of the price of good j. The easier it is to substitute between inputs to production (i.e., the larger is σ), the lower will be the demand for any particular intermediate good.

Intermediate goods producers

Firms producing intermediate good j take labor L_j and use it to produce q_j units of the intermediate good via the following production function

$$q_j = \phi(L_j - f) \tag{3.6}$$

where $\phi > 0$ is physical productivity and $f > 0$ is fixed costs of production (both of which as assumed constant across cities).

I assume monopolistic competition in the production of intermediate goods: firms producing good j take the demand for good j from firms producing the final consumption good as given and set their price in order to maximize profits subject to the additional constraint imposed by their production technology:

$$\max_{p_j} \Pi_i = p_j q_j - w_i L_j \tag{3.7}$$

subject to

$$q_j = p_j^{-\sigma} Y_i \tag{3.8}$$

$$q_j = \phi(L_j - f) \tag{3.9}$$

Substituting these constraints into the objective function yields

$$\max_{p_j} \ \Pi_i = p_j^{1-\sigma} Y_i - w_i \left(\frac{1}{\phi} p_j^{-\sigma} Y_i + f \right). \tag{3.10}$$

The optimal price must satisfy the following first-order necessary condition:

$$\frac{\partial \Pi_i}{\partial p_j} \equiv -(\sigma - 1) p_j^{-\sigma} Y_i + \frac{w_i}{\phi} \sigma p_j^{-\sigma-1} Y_i = 0 \tag{3.11}$$

which implies that the optimal price for good j is

$$p_j = \frac{w_i}{\phi} \frac{\sigma}{\sigma - 1}. \tag{3.12}$$

Note that the optimal price for good j depends only on exogenous parameters of the model (and the wage, w_i). Firms producing in city i have marginal costs that are proportional to the wage in city i and inversely proportional to productivity:

$$MC \equiv \frac{w_i}{\phi} \tag{3.13}$$

thus I can interpret

$$\mu \equiv \frac{\sigma}{\sigma - 1} \geq 1 \tag{3.14}$$

as the markup over marginal costs charged by the producers of good j. Note that the markup is strictly decreasing in σ:

$$\frac{\partial \mu}{\partial \sigma} = -\left(\frac{1}{\sigma - 1} \right)^2 < 0 \tag{3.15}$$

and that

$$\lim_{\sigma \to \infty} \mu = 1. \tag{3.16}$$

The larger is the elasticity of substitution between intermediate inputs in production, the lower is the markup. In the limit as the intermediate inputs become perfect substitutes in the production of the final consumption good, the markup disappears and firms producing good j are forced to set price equal to marginal cost.

I now make the standard assumption of free entry in the production of intermediate good

j. The assumption of free entry drives profits of firms producing good j to zero.

$$\Pi_i = p_j q_j - w_i L_j = 0 \tag{3.17}$$

Substituting for L_i using equation 3.6 and then substituting for p_i using equation 3.12 yields the following expressions for q_j, the quantity of good j produced, and L_j, the quantity of labor required to produce q_j units of good j. Combining these expression with the equation for p_j completely describes the behavior of firms producing good j.

$$p_j = \mu \frac{w_i}{\phi} \tag{3.18}$$

$$q_j = \left(\frac{1}{\mu - 1}\right)\phi f \tag{3.19}$$

$$L_j = \left(\frac{\mu}{\mu - 1}\right) f \tag{3.20}$$

Note that none of these expression depends directly on j which implies that firms producing intermediate goods are symmetric.

Households

The behavior of households in each city is trivial. I assume linear preferences over the homogenous generic consumption good.

$$U_i = C_i \tag{3.21}$$

Households in city i supply labor to the intermediate goods producers in return for a wage w_i. Household budget constraint is

$$C_i = w_i L_i \tag{3.22}$$

I assume that households choose there demand for the consumption good in order to maximize U_i subject to the budget constraint. The assumption of linear preferences implies that optimal behavior for households is to simply consume their income.

Equilibrium

Labor market clearing requires that

$$L_i = \int_0^{n_i} L_j dj = n_i \left(\frac{\mu}{\mu - 1}\right) f. \tag{3.23}$$

Solving the above equation for n_i yields

$$n_i = n(L_i) = \left(\frac{\mu - 1}{\mu}\right)\frac{1}{f}L_i \tag{3.24}$$

The number (formally measure), n_i, of intermediate goods produced in city i is an increasing function of the size of the city (as measured by population) and is a decreasing function of both the elasticity of substitution, σ, and fixed costs, f.

At this point I am in a position to characterize both equilibrium output/income, Y_i, equilibrium per capita output/income, $y_i = \frac{Y_i}{L_i}$, and the equilibrium productivity of the final goods sector (i.e., quantity of the final output good produced using given amounts of the various intermediate goods/inputs).

$$Y_i = \phi f\left(\frac{1}{\mu - 1}\right)\left(\left(\frac{\mu - 1}{\mu}\right)\frac{1}{f}L_i\right)^{\mu} \tag{3.25}$$

$$y_i = \phi\left(\frac{1}{\mu}\right)\left(\left(\frac{\mu - 1}{\mu}\right)\frac{1}{f}L_i\right)^{\mu-1} \tag{3.26}$$

$$prod_i = \left(\left(\frac{\mu - 1}{\mu}\right)\frac{1}{f}L_i\right)^{\mu-1} \tag{3.27}$$

To determine the equilibrium wage in city i note that the goods market clearing condition requires that

$$Y_i = C_i = w_i L_i. \tag{3.28}$$

This condition just requires that households in city i consume all of the final goods produced by firms in city i. Goods market clearing implies that the wage equal per capita output, y_i.

$$w_i = \frac{Y_i}{L_i} = y_i. \tag{3.29}$$

In keeping with standard results from seminal work in urban economics and economic geography,[8] equations 3.25, 3.26, 3.27, and 3.29 define power law (i.e., log-linear) scaling relations between output, per capita output, productivity, and wages, respectively, and city size. The strength of these various power law scaling relations depends on the elasticity of output, Y, with respect to city size, L, which in the benchmark model, is equal to the markup charged by firms producing intermediate inputs, μ.

$$\frac{\partial \ln Y}{\partial \ln L} = \mu \tag{3.30}$$

[8] See, for example, Abdel-Rahman (1988), Fujita (1988), Rivera-Batiz (1988), Ciccone and Hall (1996). For detailed surveys of similar results from economic geography see Krugman (1996) and Fujita et al. (1999).

In the benchmark model, indeed in most all models where increasing returns are generated by product variety, the degree of increasing returns is intimately connected to the degree of market power held by firms producing the intermediate inputs. In the next section, I extend the benchmark model by creating an indirect link between the size of a city and the markup by explicitly assuming a link between the size of a city and the elasticity of substitution between intermediate inputs. Under certain conditions, I show that this link eliminates the market power of firms producing intermediate inputs which in turn destroys the increasing returns to scale in the production of final consumption goods.

3.3.2 Extended model with variable $\sigma(n_i)$

I follow in the spirit of Zhelobodko et al. (2012) and suppose that the elasticity of substitution between any two inputs to production in city i varies with the total number of available inputs, n_i: $\sigma(n_i)$. Specifically, I define

$$\rho(n_i) = \frac{\sigma(n_i) - 1}{\sigma(n_i)} \iff \sigma(n_i) = \frac{1}{1 - \rho(n_i)} \tag{3.31}$$

and suppose that ρ is a linear function of n_i:

$$\rho(n_i) = \beta_0 + \beta_1 n_i. \tag{3.32}$$

Recall that I require intermediate goods to be gross substitutes in production: $1 < \sigma(n_i) \leq \infty$ which requires $0 < \rho(n_i) < 1$ for all n_i.

Our assumption nests the benchmark model as a special case where $\beta_1 = 0$ and implies that the elasticity of substitution between intermediate inputs it a convex function that is either increasing (decreasing) in the number of available inputs, n_i, if $\beta_1 > (<) 0$. Using the language of Zhelobodko et al. (2012), $\beta_1 > (<) 0$ implies increasing (decreasing) "relative love of variety (RLV)."

$$\sigma(n_i) = \frac{1}{1 - \beta_0 - \beta_1 n_i} \tag{3.33}$$

$$\sigma'(n_i) = \frac{\beta_1}{(1 - \beta_0 - \beta_1 n_i)^2} \tag{3.34}$$

$$\sigma''(n_i) = \frac{2\beta_1^2}{(1 - \beta_0 - \beta_1 n_i)^3} \tag{3.35}$$

From this point forward, much of the analysis from the previous section goes through with only minor modification.

Final goods producers

Producers of the final consumption continue to use a CRTS, CES production technology and choose their respective demands for the various intermediate inputs in order to maximize profits. The first-order necessary condition for good j implies the following demand function for good j.

$$\begin{aligned} q_j =& p_j^{-\sigma(n_i)} Y_i \\ =& p_j^{-\left(\frac{1}{1-\beta_0-\beta_1 n_i}\right)} Y_i \end{aligned} \tag{3.36}$$

Intermediate goods producers

Firms producing intermediate goods continue to operate in a monopolistically competitive environment and choose a price for good j in order to maximize their profits subject to the constraints imposed by equations 3.6 and 3.36. The first-order necessary condition for the optimal price p_j implies that firms set their price equal to some markup over marginal costs.

$$\begin{aligned} p_j =& \frac{\sigma(n_i)}{\sigma(n_i)-1}\frac{w_i}{\phi} \\ =& \frac{1}{\beta_0+\beta_1 n_i}\frac{w_i}{\phi} \\ =& \mu(n_i)\frac{w_i}{\phi}. \end{aligned} \tag{3.37}$$

The requirement that $1 < \sigma(n_i) \leq \infty$ implies that $1 \leq \mu(n_i) < \infty$ for all n_i. Note that the markup varies with the number of available intermediate inputs. In particular, if the elasticity of substitution between intermediate inputs is an increasing function of the number of available inputs (i.e., $\beta_1 > 0$), then the markup is a decreasing function of the number of available inputs

$$\mu'(n_i) = -\frac{\beta_1}{(\beta_0+\beta_1 n_i)^2} < 0. \tag{3.38}$$

This is an example of the type of "pro-competitive" effects generated by increasing relative love for variety. With $\beta_1 > 0$, an increase in the number of available intermediate inputs increases the elasticity of substitution between inputs in the production of consumption goods. The rise in the elasticity of substitution pushes down the markup and lowers the price that firms producing input j can charge.

Free entry in the production of intermediate goods drives the profits of firms producing

these goods to zero in equilibrium which, in turn, implies that quantity of intermediate input j produced in equilibrium is

$$\begin{aligned} q_j =&(\sigma(n_i)-1)\phi f \\ =&\left(\frac{1}{\mu(n_i)-1}\right)\phi f \\ =&\left(\frac{\beta_0+\beta_1 n_i}{1-\beta_0-\beta_1 n_i}\right)\phi f. \end{aligned} \tag{3.39}$$

The amount of labor required to produce q_j of intermediate input j in equilibrium is

$$\begin{aligned} L_j =&\sigma(n_i) f \\ =&\left(\frac{\mu(n_i)}{\mu(n_i)-1}\right) f \\ =&\left(\frac{1}{1-\beta_0-\beta_1 n_i}\right) f. \end{aligned} \tag{3.40}$$

Note that if $\sigma'(n_i) = \beta_1 > 0$, then both the output of intermediate input j and the labor used its production are increasing functions of number of available inputs, n_i.

$$\frac{\partial q_j}{\partial n_i} = \frac{\beta_1}{(1-\beta_0-\beta_1 n_i)^2} > 0 \tag{3.41}$$

$$\frac{\partial L_j}{\partial n_i} = \frac{\beta_1}{(1-\beta_0-\beta_1 n_i)^2} > 0 \tag{3.42}$$

This is another example of the "pro-competitive effects" generated by increasing relative love for variety: the larger the number of available intermediate inputs (i.e., the larger the market), the larger are firms (in terms of either the size of their labor force, L_j, or the amount of output, q_j).

Equilibrium

Since the behavior of households in the general model does not change, I move straight to characterizing the model equilibrium. Imposing labor market clearing requires that

$$\begin{aligned} L_i = \int_0^{n_i} L_j dj =&n_i \sigma(n_i) f \\ =&n_i \left(\frac{\mu(n_i)}{\mu(n_i)-1}\right) f \\ =&n_i \left(\frac{1}{1-\beta_0-\beta_1 n_i}\right) f. \end{aligned} \tag{3.43}$$

Equilibrium level of product differentiation: As in the benchmark model, the labor market clearing condition implies an explicit relation between the size of a city, L_i, and the variety of intermediate inputs, n_i.[9]

$$n(L_i) = \frac{L_i(1-\beta_0)}{f+\beta_1 L_i} \tag{3.46}$$

Elasticity of substitution and the markup: With a functional form for $n(L_i)$ in hand, I can express the elasticity of substitution between intermediate inputs, σ, and the markup, μ, as functions of city size, L_i. After substituting the result for $n(L_i)$ into equation 3.31 I find that σ is a linear function of city size.

$$\sigma(L_i) = \frac{1}{1-\beta_0} + \frac{\beta_1}{1-\beta_0}\frac{L_i}{f} \tag{3.47}$$

Note that, in the limit of large cities, intermediate inputs become perfect substitutes in the production of final consumption goods.

$$\lim_{L_i \to \infty} \sigma(L_i) = \infty \tag{3.48}$$

Using the expression for $\sigma(L_i)$, I can derive the following expression for μ.

$$\mu(L_i) = \frac{f+\beta_1 L_i}{\beta_0 f + \beta_1 L_i} \tag{3.49}$$

Note that in order for $1 \leq \mu(L_i) < \infty$ I require that $0 < \beta_0 \leq 1$. Following an application of L'hôpital's rule, I find that, in the limit of large cities, the markup tends towards unity.

$$\lim_{L_i \to \infty} \mu(L_i) = 1 \tag{3.50}$$

If $\beta > (<)\ 0$, then firms producing intermediate inputs in larger markets (i.e., large cities) should have lower (higher) markups. Empirical evidence provided by Syverson (2007) suggests that firms operating in larger cities have lower markups, which, in the model,

[9] In the most general framework, the labor market clearing condition defines the variety of intermediate inputs only implicitly as a function of city size. Given some value for L_i I can define $n_i \equiv n(L_i)$ to be the value n_i that solves the following non-linear equation.

$$R(L_i, n_i) \equiv \frac{L_i}{f} - n_i\left(\frac{\mu(n_i)}{\mu(n_i)-1}\right) = 0 \tag{3.44}$$

From the implicit function theorem, I know that the function $n_i \equiv n(L_i)$ exists so long as

$$\frac{\partial R}{\partial n_i} \equiv \frac{1}{\mu(n_i)-1}\left[\left(\frac{\mu'(n_i)}{\mu(n_i)-1}\right)n_i - \mu(n_i)\right] \neq 0. \tag{3.45}$$

requires $\beta_1 > 0$.

Intermediate goods producers: Recall that in the benchmark model there is no link between the size of firms producing in a given city and the size of that city i. This runs against empirical evidence which suggests that firms in larger markets tend to be larger.[10] In the general model the variation in the markup with city size described by equation 3.49 creates a link between the size of a city and the size (as measured by either output or labor force) of firms producing intermediate goods in that city.

$$\begin{aligned} q_j =& \frac{\phi f \beta_0}{1-\beta_0} + \frac{\phi \beta_1}{1-\beta_0} L_i \\ =& c_0 + c_1 L_i \end{aligned} \tag{3.51}$$

$$\begin{aligned} L_j =& \frac{f}{1-\beta_0} + \frac{\beta_1}{1-\beta_0} L_i \\ =& c_2 + c_3 L_i \end{aligned} \tag{3.52}$$

Note that c_1 and c_3 represent the marginal effect on firm size in response to a change in the size of a city. Equations 3.51 and 3.52 tell us that if $\beta_1 > (<)\ 0$ (alternatively, $c_1, c_3 > (<)\ 0$), then the size of firms producing intermediate goods is an increasing (decreasing) function of city size.

Output, productivity, and wages: I are now ready to characterize equilibrium output, productivity, and wages as functions of the city size, L_i, and structural parameters of the model. I begin with total output, Y_i.

$$\begin{aligned} Y_i =& (c_0 + c_1 L_i) \left(\frac{L_i}{c_2 + c_3 L_i} \right)^{\frac{c_1 c_2 + c_1 c_3 L_i}{c_0 c_3 + c_1 c_3 L_i}} \\ =& q_j \left(\frac{L_j}{L_i} \right)^{-\epsilon_{q_j, L_i}} \end{aligned} \tag{3.53}$$

where ϵ_{q_j, L_i} is the elasticity of the output of firms producing input j with respect to city size. Equation 3.53 says that total output produced in city i varies directly with the total output of firms producing intermediate good j and inversely as the ϵ_{q_j, L_i} with the fraction of total labor used by firms producing intermediate good j.

Per capita output in city i, y_i, varies directly with the productivity of firms producing intermediate good j and inversely as the $1 - \epsilon_{q_j, L_i}$ with the fraction of total labor used by

[10] Campbell and Hopenhayn (2005) characterize the effects of market size on the size distribution of firms using data on retail trade establishments across 225 U.S. cities. In general they find that larger firms to to exist in larger cities.

firms producing intermediate good j.

$$\begin{aligned} y_i &= \left(\frac{c_0 + c_1 L_i}{c_2 + c_3 L_i}\right)\left(\frac{L_i}{c_2 + c_3 L_i}\right)^{\frac{c_1 c_2 - c_0 c_3}{c_0 c_3 + c_1 c_3 L_i}} \\ &= \left(\frac{q_j}{L_j}\right)\left(\frac{L_j}{L_i}\right)^{1-\epsilon_{q_j,L_i}} \end{aligned} \tag{3.54}$$

The productivity of firms producing the final consumption good in city i is the amount of output produced per unit of inputs

$$\begin{aligned} prod_i \equiv \frac{Y_i}{\int_0^{n(L_i)} q_j dj} &= \left(\frac{L_i}{c_2 + c_3 L_i}\right)^{\frac{c_1 c_2 - c_0 c_3}{c_0 c_3 + c_1 c_3 L_i}} \\ &= \left(\frac{L_j}{L_i}\right)^{1-\epsilon_{q_j,L_i}} \end{aligned} \tag{3.55}$$

Finally, as was the case in the benchmark model, the goods market clearing implies that the wage equals per capita output, y_i.

$$w_i = \frac{Y_i}{L_i} = y_i \tag{3.56}$$

I are interested in the limiting behavior of per capita output.

Claim: As the intermediate inputs used in the production of the final consumption good become perfect substitutes, firms producing these inputs are forced to set their output price equal to their marginal costs of production:

$$\lim_{L_i \to \infty} \sigma(L_i) = \infty \implies \lim_{L_i \to \infty} \mu(L_i) = 1. \tag{3.57}$$

This loss of market power by intermediate goods producers, in turn, eliminates the increasing returns to scale in the production of final consumption goods:

$$\lim_{\sigma(L_i) \to \infty} y_i = \lim_{\mu(L_i) \to 1} y_i = \phi. \tag{3.58}$$

Proof: See appendix 3.A. ■

In the benchmark model with constant σ there was unbounded growth in per capita income

with city size: cities exhibit increasing returns to scale at all scales. However, in the more general model with a variable elasticity of substitution (i.e., $\sigma(L_i)$) I have bounded growth of per capita income with city size: while there are increasing returns to scale for small to medium sized cities, in the limit of large cities there is only constant returns to scale.

3.4 Empirics

3.4.1 Data

I follow the common approach in the literature and analyze data on population, income, and output of Metropolitan Statistical Areas (MSAs) available from the Bureau of Economic Analysis (BEA). MSAs are defined for each urban agglomeration with at least 50,000 people and attempts to capture the overall size of the agglomeration by merging administratively defined entities (i.e., cities, counties, places, towns, etc) in the US, based on social or economic ties.[11] Figure 3.1 displays scatter plots of real total and per capita Gross Metropolitan Product (GMP) versus population for the $N = 366$ Metropolitan Statistical Areas.

From the scatter plots its is clear that the relationship between per capita output and city size is rather noisy. Distinguishing between different models is going to be difficult. Nonetheless I will give it a go!

The analysis of Bettencourt et al. (2007) and Bettencourt et al. (2010) focuses on scaling relations between aggregate (i.e., extensive) quantities and city size as measured by total population. This approach is heavily criticized on statistical grounds by Shalizi (2014b).[12] The approach also makes very little economic sense. As such I choose to focus

[11] MSAs, as defined by the Office of Management and Budget (OMB), are the standard unit of analysis for much of urban economics. The OMB defines a Core Based Statistical Area (CBSA) as "a statistical geographic entity consisting of the county or counties associated with at least one core (urbanized area or urban cluster) of at least 10,000 population, plus adjacent counties having a high degree of social and economic integration with the core as measured through commuting ties with the counties containing the core."

Micropolitan and Metropolitan Statistical Areas are the two categories of CBSAs. A Micropolitan Statistical Area is a CBSA with "at least one urban cluster that has a population of at least 10,000, but less than 50,000. The Micropolitan Statistical Area comprises the central county or counties containing the core, plus adjacent outlying counties having a high degree of social and economic integration with the central county or counties as measured through commuting." A Metropolitan Statistical Area is a CBSA with "at least one urbanized area that has a population of at least 50,000. The Metropolitan Statistical Area comprises the central county or counties containing the core, plus adjacent outlying counties having a high degree of social and economic integration with the central county or counties as measured through commuting."

Ideally I would have liked to include Micropolitan Statistical Area data in the analysis, unfortunately the BEA only publish GMP data for Metropolitan Statistical Areas.

[12] For example, Shalizi (2014b) argues that the high values of R^2 reported by Bettencourt et al. (2007) are

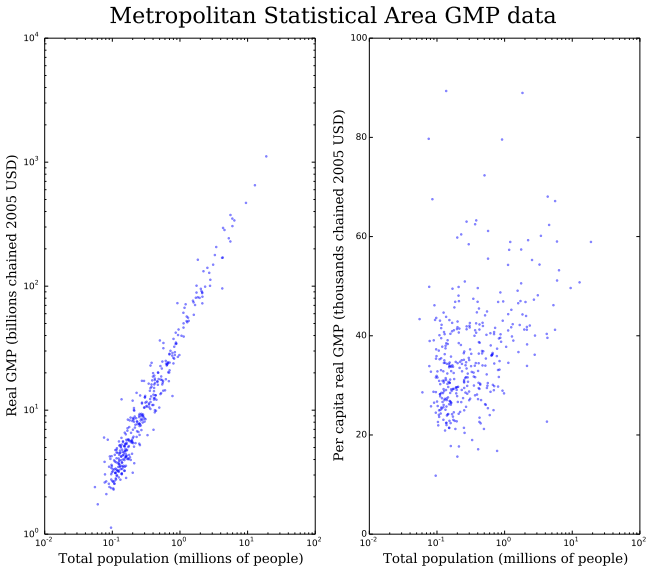

Figure 3.1: From the scatter plots its is clear that the relationship between per capita output and city size is rather noisy. Distinguishing between different models is going to be difficult.

on modeling the relationship between per capita GMP and city size as measured by total population.

3.4.2 Methodology

I start by comparing two statistical models of the relationship between per capita GMP and total population that have wildly different implications for the behavior of per capita income in the limit of large cities: a power law (i.e, log-linear) scaling model and a logistic scaling model.

---

largely an artifact of data aggregation. Suppose, for example, that per capita GMP is actually statistically independent of city size. Under this hypothesis $\ln Y = \ln y + \ln L$ is the sum of two independent random variables and the variance of $\ln Y$ is the sum of the variances of log per capita GMP and log population. The R^2 of a regression of $\ln Y$ on $\ln L$ is

$$R^2 = \frac{var(\ln L)}{var(\ln y) + var(\ln L)}. \tag{3.59}$$

If the variance of $\ln L$ is large relative to the variance of $\ln y$, then R^2 will be close to unity! Indeed, performing this calculation on the data set yields $R^2 = 0.93$. Thus, given the data, one would expect to find an R^2 of roughly 0.93 when regressing $\ln Y$ on $\ln L$ even when per capita income and population are completely independent!

	Population, L	GMP, Y	Per capita GMP, y
count	366	366	366
mean	0.71	31.69	36.00
std	1.58	86.66	11.25
min	0.06	1.12	11.79
25%	0.14	4.33	28.51
50%	0.25	8.40	34.04
75%	0.56	20.10	41.30
max	18.90	1113.38	89.35

Table 3.1: Descriptive statistics for Metropolitan Statistical Area data. Population, L, is measured in millions of people, real GMP, Y, is measured in billions of chained 2005 USD, and per capita real GMP, y, is measured in thousands of chained 2005 USD.

Power law scaling

Bettencourt et al. (2007) report that GMP scales as a power of population: $Y \sim cL^b$. A scaling relation between GMP and population mechanically implies a scaling relation between per capita GMP, $y = \frac{Y}{L}$ and population of the form $y \sim cL^{b-1}$ which can be connected to the data via the regression equation:

$$\ln y_i = \gamma_0 + \gamma_1 \ln L_i + \epsilon_i. \tag{3.60}$$

where $\gamma_0 = \ln c$ and $\gamma_1 = b - 1$ and ϵ is assumed to be independent, mean-zero disturbance. I estimate equation 3.60 using OLS.[13] Note that if $\gamma_1 > 0$, then the power law scaling model predicts that growth in per capita GMP with total population is unbounded.

Logistic scaling

Following Shalizi (2014b) I also consider a logistic scaling relation between per capita GMP, y, and population, L. Logistic scaling implies that:

$$\ln y \sim d_1 + d_2 \left(\frac{1}{1 + e^{-(d_3 + d_4 L)}} \right). \tag{3.61}$$

Note that while the power law scaling model implies that per capita GMP growth is unbounded in total population, the logistic scaling model predicts that per capita GMP

[13] Classic OLS standard errors will be too narrow in the presence of heteroskedasticity. Although analysis of OLS residuals failed to indicate the presence of substantial heteroskedasticity, I re-estimated equation 3.60 with White (1980) standard errors. The resulting heteroskedasticity consistent standard errors and associated confidence intervals did not differ substantially from the classic standard errors and confidence intervals reported below.

converges to a constant in the limit of large cities.

$$\lim_{L\to\infty} \ln y = d_1 + d_2 \tag{3.62}$$

The logistic scaling model can be connected to the data via the following non-linear regression equation

$$\begin{aligned} \ln y_i =& d_1 + d_2 \left(\frac{1}{1 + e^{-(d_3 + d_4 L)}} \right) + \epsilon_i \\ =& F(d_1, d_2, d_3, d_4; L_i) + \epsilon_i \end{aligned} \tag{3.63}$$

where ϵ is assumed to be independent, mean-zero disturbance. I estimate equation 3.63 using non-linear least squares as follows. First, I define the following residual function

$$R(\theta; y_i, L_i) = \ln y_i - F(\theta; L_i) \tag{3.64}$$

for each of the $i = 1, \ldots, N$ cities where θ is a vector containing the four structural parameters of the model $d_1,\ d_2, d_3, d_4$. I then choose the vector of parameters which minimizes the sum of squared deviations between the model predicted per capita GMP and the observed data on per capita GMP.

$$\hat{\theta}^* = \arg\min_{\theta}\ \frac{1}{2} \sum_{i=1}^{N} R(\theta; y_i, L_i)^2 \tag{3.65}$$

Under the assumption that the disturbances ϵ_i (and therefore the residuals $R(\theta; y_i, L_i)$), are independently and identically distributed, the non-linear least squares estimator $\hat{\theta}^*$ is consistent, asymptotically efficient, and asymptotically normal.[14] In general solving non-linear optimization problems like 3.65 is non-trivial. In order to solve for $\hat{\theta}^*$ in equation 3.65 I use the Levenberg-Marquardt (LM) algorithm which has been developed to exploit the structure of non-linear least squares problems.[15]

The uncertainty of the parameter estimates is inversely related to the amount of curvature

[14] For those interested in the gory details, see either Amemiya (1985), Davidson and MacKinnon (1993), or Hayashi (2000).

[15] The Levenberg-Marquardt (LM) algorithm, which was originally developed by Levenberg (1944) and Marquardt (1963), adaptively combines two general approaches to non-linear optimization: the gradient descent and Gauss-Newton methods. When the current solution is "far" from the optimal solution, the LM algorithm behaves like a gradient descent method: the rate of convergence is slow, but ultimate convergence to a local minimum is guaranteed. When the current solution is "close" to the optimal solution, the LM algorithm behaves like a Gauss-Newton method: the rate of convergence is fast, and because I am already close to the optimal solution it is unlikely that Gauss-Newton will fail to converge. See Judd (1998) and Nocedal and Wright (2006) for more details on gradient descent and Gauss-Newton methods as well as the mathematical details of the Levenberg-Marquardt (LM) algorithm.

in the objective function in the neighborhood of the optimal parameter vector, $\hat{\theta}^*$. If the objective function has a lot of curvature around $\hat{\theta}^*$, then the parameter estimates will be quite precise and the standard errors will be small; if, on the other hand, the objective function is very flat in the neighborhood of $\hat{\theta}^*$, then the parameter estimates will be very imprecise and the standard errors will be large. Information about the curvature of the objective function around $\hat{\theta}^*$ is encoded by $H(\hat{\theta}^*; y, L)$, the Hessian of the objective function evaluated at $\hat{\theta}^*$.

Formally, I compute the standard errors for the parameter estimates in two steps. First, I construct an estimate of the residual variance $\hat{\sigma}^2$ as follows

$$\hat{\sigma}^2 = \frac{1}{N-k} \sum_{i=1}^{N} R(\hat{\theta}^*; y_i, L_i)^2 \tag{3.66}$$

where N is the total number of observations and k is the number of estimated parameters and $\hat{\theta}^*$ is the optimal parameter vector. I then estimate the variance-covariance matrix Σ as

$$\hat{\Sigma} = \hat{\sigma}^2 H^{-1}(\hat{\theta}^*; y, L) \tag{3.67}$$

where H^{-1} is the inverse Hessian of the objective function defined in equation 3.65. The standard errors for the $k = 4$ parameters can be found by taking the square root of the diagonal elements of $\hat{\Sigma}$.

Structural scaling

My structural model predicts that per capita output in city i, $\hat{y}_i$, is given by the following non-linear function of the population of city i, L_i, and the four structural parameters $\beta_0, \beta_1, \phi, f$:

$$\ln\ \hat{y}_i = \ln\left[\phi\left(\frac{\beta_0 f + \beta_1 L_i}{f + \beta_1 L_i}\right)\right] + \frac{f(1-\beta_0)}{\beta_0 f + \beta_1 L_i} \ln\left(\frac{L_i(1-\beta_0)}{f + \beta_1 L_i}\right) \tag{3.68}$$

where I have substituted for $c_0 = \frac{\phi f \beta_0}{1-\beta_0}$, $c_1 = \frac{\phi \beta_1}{1-\beta_0}$, $c_2 = \frac{f}{1-\beta_0}$, and $c_3 = \frac{\beta_1}{1-\beta_0}$ using equations 3.51 and 3.52.

The model imposes the following parameter restrictions: $0 < \beta_0 < 1, -\infty < \beta_1 < \infty, 0 < \phi, 0 < f$. To incorporate these restrictions into the estimation procedure I define the

following change of variables:

$$\begin{aligned}\beta_0 =& \frac{1}{1+e^{-\ln\left(\frac{\beta_0}{1-\beta_0}\right)}}\\ \beta_1 =& \frac{\beta}{f} f\\ \phi =& e^{\ln\phi} = e^{\gamma_2}\\ f =& e^{-\ln\frac{1}{f}} = e^{\gamma_3}.\end{aligned}$$

This change of variables, in turn, leads to new structural parameters

$$\begin{aligned}\gamma_0 =& -\ln\frac{\beta_0}{1-\beta_0}\\ \gamma_1 =& \frac{\beta_1}{f}\\ \gamma_2 =& \ln\phi\\ \gamma_3 =& -\ln\frac{1}{f}\end{aligned}$$

which, no matter the estimation results for $\gamma_0, \gamma_1, \gamma_2, \gamma_3$, guarantee that the values of the original structural parameters satisfy the restrictions imposed by the model.

Re-writing equation 3.68 in terms of the new structural parameters $\gamma_0, \gamma_1, \gamma_2, \gamma_3$ yields

$$\ln\ \hat{y}_i = A(\gamma_0, \gamma_1, \gamma_2; L_i) + B(\gamma_0, \gamma_1; L_i)C(\gamma_0, \gamma_1, \gamma_3; L_i) \tag{3.69}$$

where

$$\begin{aligned}A(\gamma_0, \gamma_1, \gamma_2; L_i) =& \ln\left[\left(\frac{e^{\gamma_2}}{1+e^{\gamma_0}}\right)\left(\frac{1+\gamma_1 L_i+\gamma_1 L_i e^{\gamma_0}}{1+\gamma_1 L_i}\right)\right]\\ B(\gamma_0, \gamma_1; L_i) =& \left(\frac{e^{\gamma_0}}{1+\gamma_1 L_i+\gamma_1 L_i e^{\gamma_0}}\right)\\ C(\gamma_0, \gamma_1, \gamma_3; L_i) =& \ln\left[\left(\frac{e^{\gamma_0-\gamma_3}}{1+e^{\gamma_0}}\right)\left(\frac{L_i}{1+\gamma_1 L_i}\right)\right]\end{aligned}$$

The model can be connected to the data via the following non-linear regression equation.

$$\begin{aligned}\ln y_i =& A(\gamma_0, \gamma_1, \gamma_2; L_i) + B(\gamma_0, \gamma_1; L_i)C(\gamma_0, \gamma_1, \gamma_3; L_i) + \epsilon_i\\ =& F(\gamma_0, \gamma_1, \gamma_2, \gamma_3; L_i) + \epsilon_i\end{aligned} \tag{3.70}$$

For each of the $i = 1, \ldots, N$ cities, I define the following residual function.

$$R(\theta; y_i, L_i) = \ln\ y_i - F(\theta; L_i) \tag{3.71}$$

where θ is a vector of the four structural parameters $\gamma_0, \gamma_1, \gamma_2, \gamma_3$. Given observed data on per capita incomes, y, and city population, L, I choose θ to minimize the sum of squared deviations between the model's predictions and the observed values.

$$\hat{\theta}^* = \arg\min_{\theta}\ \frac{1}{2}\sum_{i=1}^{N} R(\theta; y_i, L_i)^2 \tag{3.72}$$

Again, under the assumption that the disturbances ϵ_i (and therefore the residuals $R(\theta; y_i, L_i)$), are independently and identically distributed, the non-linear least squares estimator $\hat{\theta}^*$ is consistent, asymptotically efficient, and asymptotically normal.

Confidence intervals are computing using the same two step procedure used to obtain the confidence intervals for the estimated parameters of the logistic scaling model discussed above. First, I construct an estimate of the residual variance $\hat{\sigma}^2$ as follows

$$\hat{\sigma}^2 = \frac{1}{N-k}\sum_{i=1}^{N} R(\hat{\theta}^*; y_i, L_i)^2 \tag{3.73}$$

where N is the total number of observations and k is the number of estimated parameters and $\hat{\theta}^*$ is the optimal parameter vector. I then estimate the variance-covariance matrix Σ as

$$\hat{\Sigma} = \hat{\sigma}^2 H^{-1}(\hat{\theta}^*; y, L) \tag{3.74}$$

where H^{-1} is the inverse Hessian of the objective function defined in equation 3.72. The standard errors for the $k = 4$ parameters can be found by taking the square root of the diagonal elements of $\hat{\Sigma}$.

3.4.3 Results

Power law scaling

The easiest way to test the super linear power law scaling hypothesis put forward by Bettencourt et al. (2007) is to simply estimate equation 3.60 using OLS and check that $\gamma_1 > 0$. Regression results for equation 3.60 are reported in table 3.2. Consistent with super linear power law scaling $\hat{\gamma}_1 \approx 0.12$ is positive and significant. A 1% increase in the population of a city results in a 0.12% increase in per capita GMP (or equivalently, a 1.12%

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
const	3.6734	0.020	181.250	0.000	3.634 3.713
population	0.1174	0.013	9.003	0.000	0.092 0.143

Table 3.2: OLS estimation results for the power law scaling model.

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
d_1	3.8788	0.0533	72.7843	0.0000	3.7743 3.9832
d_2	-0.5127	0.1802	-2.8456	0.0047	-0.8568 -0.1596
d_3	0.7736	0.2762	2.801	0.0054	0.2322 1.3149
d_4	-0.4103	0.2660	-1.5427	0.1237	-0.9315 0.1110

Table 3.3: Estimation results for the logistic scaling model.

increase in GMP).[16] The root mean square error (RMSE) of the power law scaling model for predicting per capita GMP is \$10.22 thousand chained 2005 USD per person. The R^2 is 0.182 indicating that the fitted values retain roughly 18% of the variation in the log of per capita GMP.

Logistic scaling

Table 3.3 displays the coefficients, asymptotic standard errors, and 95% confidence intervals obtained from estimating equation 3.63 using non-linear least squares. The RMSE of the logistic model for predicting per capita GMP is \$10.12 thousand chained 2005 USD.

Given that the logistic model has $k = 4$ parameters it is not surprising that the model has a lower RMSE than the power law model with $k = 2$ parameters. I use a non-parametric bootstrap procedure to estimate the distribution of the differences in RMSE between the power law and logistic scaling models under the null hypothesis that the power law scaling model with parameters as reported in table 3.2 is the "true" model.[17] The observed

[16] Bettencourt et al. (2007) and Bettencourt et al. (2010) conveniently ignore all possible sources of endogeneity and interpret γ as the elasticity of per capita income with respect to city size (i.e., a 1% increase in the total population of a city leads to a γ_1% increase in per capital income).

[17] A detailed outline of my non-parametric bootstrap procedure can be found in appendix 3.B.

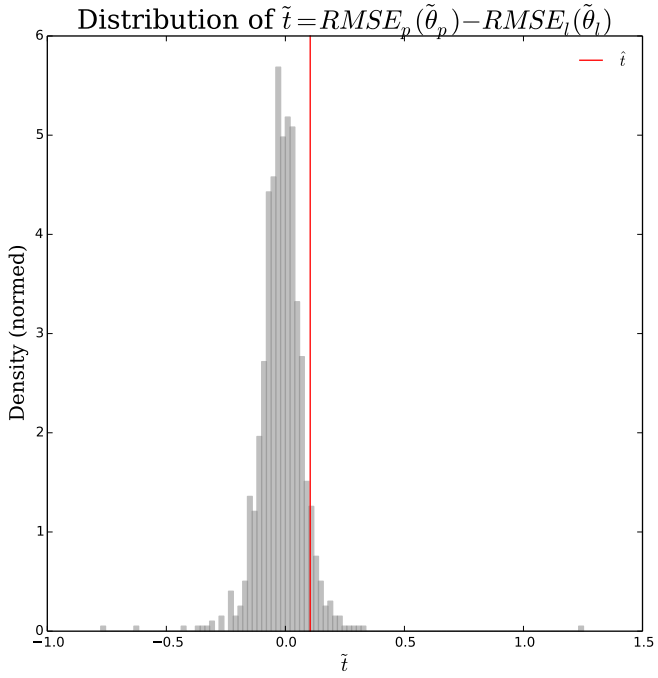

Figure 3.2: Difference in RMSE between logistic and power law scaling model is statistically significant indicating that the logistic model is preferred to the simpler power law model.

difference in RMSEs is

$$\hat{t} = RMSE_p(\hat{\theta}_p) - RMSE_l(\hat{\theta}_l) \approx 0.104 \tag{3.75}$$

where $RMSE_p(\hat{\theta}_p)$ and $RMSE_l(\hat{\theta}_l)$ are the root mean squared errors of the power law and logistic scaling models computing using the estimated parameter vectors $\hat{\theta}_p$ and $\hat{\theta}_l$. Figure 3.2 displays the distribution of $\hat{t}$. The p-value is 0.07 indicating that only 7% of the time did my simulation generate a difference in RMSE larger that the observed difference. I conclude that it is reasonably unlikely that the logistic scaling model out performs the power law scaling model "by chance."

Structural scaling

I estimate equation 3.72 using non-linear least squares and report the results in table 3.4. Reported confidence intervals were computed using the two step procedure described in

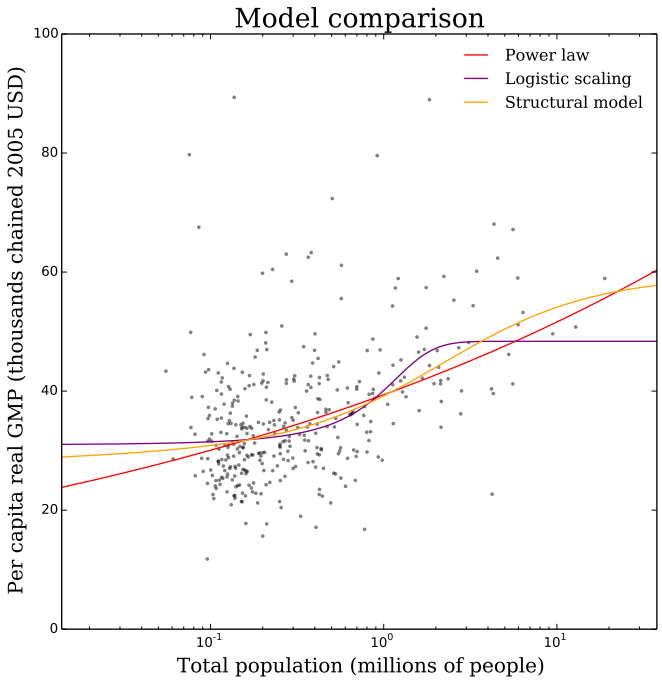

Figure 3.3: Regression curves for the power law, logistic, and structural scaling models.

the previous section.[18] Figure 3.3 plots the regression curves for the power law, logistic, and structural scaling models for comparison. Note that, like the logistic scaling model, the structural scaling model predicts that increasing returns to scale are exhausted in the limit of large cities.

Note that the parameters γ_0 and γ_3 are not identified. This lack of identification is clearly indicated by the enormous confidence intervals for the estimated values of these parameters Fortunately, the exact value of these parameters are not material for the analysis. The key estimated parameters of interest are $\hat{\gamma}_1$ and γ_2. The model predicts that per capita income and per capita output converge to a constant ϕ in the limit of large cities. Recall that my change of variables set $\phi = \ln \gamma_2$. Thus my estimation results indicate that per capita output converges to $e^{\hat{\gamma}_2} \approx \$59{,}000$. From a theoretical standpoint, the crucial parameter is γ_1 which captures the variability of the elasticity of substitution with city size. Since my general model nests the power law scaling model as a special case where $\beta_1 = \gamma_1 = 0$, I can

[18] I also computed confidence intervals using several standard non-parametric bootstrapping routines. The resulting non-parametric bootstrap confidence intervals were similar to those based on the Hessian of the objective function and lead to identical inferences. The results have been omitted for brevity. Seminal references for bootstrapping are Efron (1979) and Efron and Efron (1982). Davison (1997) is a excellent reference. Shalizi (2014a) covers the basics.

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
$\hat{\gamma}_0$	-5.9744	40.6401	-0.1470	0.8832	(-85.6274, 73.6786)
$\hat{\gamma}_1$	0.6692	0.8776	0.7626	0.4462	(-1.0508, 2.3893)
$\hat{\gamma}_2$	4.0801	0.1580	25.8188	0.0000	(3.7704, 4.3898)
$\hat{\gamma}_3$	262.41	1.0716e4	2.4489e-2	0.9805	(-2.0740e4, 2.1264e4)

Table 3.4: Non-linear least squares results for the structural scaling model.

use a simple t-test for the significance of $\hat{\gamma}_1$ as a test of the null hypothesis of power law scaling against the structural scaling model with $\gamma_1 > 0$. Recall that $\gamma_1 = 0$ implies that σ is constant and independent of city size, whereas $\gamma_1 > 0$ implies that σ is an increasing function of city size. While the estimation results indicate that $\hat{\gamma}_1$ is indeed positive, the associated 95% confidence interval includes zero and as a result I fail to reject the power law null hypothesis. Despite the fact that $\hat{\gamma}_1$ is not statistically significant, a model with $\hat{\gamma}_1 > 0$ predicts that firms operating in larger markets should be larger (in terms of both employment and output) and charge lower markups. Thus a model with $\hat{\gamma}_1 > 0$ is at least consistent with empirical results reported in Syverson (2007) and Campbell and Hopenhayn (2005).

In-sample performance

The RMSE for the structural model is \$10.14 thousand chained 2005 USD per person which is less than the RMSE for the power law scaling model reported in section 3.4. Given that the structural model has $k = 4$ parameters it is not surprising that the model has a lower RMSE than the power law model with $k = 2$ parameters. I use a non-parametric bootstrap procedure to estimate the distribution of the differences in RMSE between the power law and the structural scaling models under the null hypothesis that the power law scaling model with parameters as reported in table 3.2 is the "true" model.[19] The observed difference in RMSEs is

$$\hat{t} = RMSE_p(\hat{\theta}_p) - RMSE_s(\hat{\theta}_s) \approx 0.08 \tag{3.76}$$

[19] A detailed outline of the non-parametric bootstrap procedure can be found in appendix 3.B.

where $RMSE_p(\hat{\theta}_p)$ and $RMSE_s(\hat{\theta}_s)$ are the root mean squared errors of the power law and structural scaling models computing using the estimated parameter vectors $\hat{\theta}_p$ and $\hat{\theta}_s$. The difference between the RMSEs for the two models is statistically significant, but only marginally so: only 8% of the time did the simulations generate a difference in RMSE larger that the observed difference.

Out-of-sample performance

The RMSE measure of in-sample fit. However, as shown in figure 3.3, the predictions of the structural scaling model differs most aggressively from those of the power law and logistic scaling models for very small and very large cities. Thus what I really care about is out-of-sample forecasting ability of the model. To assess the out-of-sample predictive performance of the structural model relative to the power law and logistic scaling models I implement two types of cross-validation (CV): k-fold CV and "leave-one-out" CV. Cross-validation (CV) is a standard procedure for assessing out-of-sample performance of competing models that is widely used in statistics and machine learning.[20]

In general, CV proceeds as follows. Pick some small integer k and divide the data at random into k equally sized pieces called "folds". Call the first fold the "testing data" and then fit each of the competing models using the remaining $k-1$ folds as "training data" and evaluate each of their predictions using the testing data. Now make the second fold the testing data and the remaining $k-1$ folds the training data. Fit the models to the training data and evaluate their predictions on the testing data. Repeat this process until each of the k folds has been used as the testing data. The average predictive performance of a model across the testing sets is referred to as the k-fold cross-validation estimate of the generalization error and is an unbiased estimate of the model's out-of-sample prediction error. "Leave-one-out" cross-validation can be thought of as a special case of k-fold cross-validation with $k = N$, where N is the total number of observations in the data. Whether using k-fold or "leave-one-out" CV, the preferred model is the one with the smallest estimated generalization error as measure by the root mean squared error (averaged across the k folds).[21][22]

[20] For additional economic applications of cross-validation see Racine (1997), Racine (2000), and Hansen and Racine (2012).

[21] For those interested in a more detailed treatment of the theory behind the various flavors of CV, see Arlot and Celisse (2010). For those interested in a more practical and intuitive approach to the key ideas behind CV, see Shalizi (2014a).

[22] Those more accustomed to using standard model selection criteria, such as the Akaike Information Criterion (AIC), emphasized in the econometrics literature may take comfort from the work of Claeskens and Hjort (2008) who prove that, as sample size set gets large, AIC is equivalent to "leave-one-out" CV. Furthermore, as discussed in Shalizi (2014a), Claeskens and Hjort (2008) also demonstrate formally that the "domain of validity" for AIC is a strict subset of the domain of validity of cross-validation and that therefore the AIC is best viewed as an asymptotic approximation the more accurate "leave-one-out" CV procedure.

Fold, k	Power law model	Logistic model	Structural model
1	20.02	9.90	9.88
2	12.13	8.13	8.10
3	13.02	11.65	11.67
4	33.86	12.15	12.18
5	15.34	7.96	7.99
6	13.11	10.91	10.74
$\overline{RMSE}$	17.95	10.11	10.09

Table 3.5: k-fold CV results indicate that the structural scaling model generalizes better out-of-sample compared with both the logistic scaling model and the power law scaling model.

Table 3.5 reports the 6-fold cross validation results.[23] For completeness, I report the RMSE for each of the six folds, however the key statistic is $\overline{RMSE}$, the average RMSE across folds the $k = 6$ folds. The relative magnitudes of $\overline{RMSE}$ for the power law, logistic, and structural scaling models provide a relative ranking of these models based on how well each predict out-of-sample. Both the logistic scaling model and the structural model clearly out-perform the power law scaling model favored by Bettencourt (2013). Somewhat surprisingly given the substantial structure imposed on the relation between city size and per capita output by the structural model, the structural model has a slightly lower cross-validation RMSE.

3.5 Conclusions

The simple power law scaling relations advocated by by Bettencourt et al. (2007), Bettencourt and West (2010), and Bettencourt et al. (2010) predict that the returns to scale for various socio-economic indicators and measures of city resource use are independent of city size. The major contribution of this paper is to show that, although such scaling relations can be supported as equilibrium outcomes in a standard economic framework featuring monopolistic competition, product differentiation, and fixed costs of production, a simple extension of the standard framework that incorporates a variable elasticity of substitution between intermediate inputs generates competitive forces that are sufficient to eliminate increasing returns to scale in output, income, wages, and productivity in the limit of large cities.

The model predicts that if the elasticity of substitution is an increasing function of the number of available intermediate inputs, then firms producing intermediate inputs will

[23] The "leave-one-out" cross-validation results are similar and are omitted for brevity.

be larger (in terms of both output and employment) and the markups charged by these firms will be lower in larger cities. As the size of a city increases, the competitive pressure generated by falling markups erodes the source of increasing returns to city size. In the limit of large cities, the competitive pressure becomes so great that firms producing intermediate inputs are forced to equate their price with their marginal costs of production and the production of consumption goods exhibits constant, rather than increasing, returns to city size. Unfortunately, despite the simplicity of the mechanism driving the main results direct empirical support for the model's predicted scaling relation between city size and per capita income is weak. There is simply insufficient data to sharply distinguish between the highly non-linear structural scaling relation and the log-linear or power law scaling relation.

I feel that the current theoretical framework has at least two major shortcomings that need to be addressed in future work. First, the model fails to incorporate any notion of spatial equilibrium. As emphasized by Glaeser and Gottlieb (2009) analyzing the relationship between per capita output (or income) and population in isolation makes little sense if per capita output/income and population of cities are jointly determined in equilibrium along with prices and wages as is typically the case in spatial equilibrium models descending from Mills (1967), Rosen (1979) and Roback (1982) and epitomized by the more recent work of Glaeser (2008) and Glaeser and Gottlieb (2009). The model also fails to incorporate any notion of trade between cities that is the focal point of the economic geography approach to agglomeration economies epitomized by Krugman (1996) and Fujita et al. (1999). Extending the model to allow for the free flow of people and goods between cities is the next step in The research.

Appendix

3.A Mathematical appendix

Claim: As the intermediate inputs used in the production of the final consumption good become perfect substitutes, firms producing these inputs are forced to set their output price equal to their marginal costs of production:

$$\lim_{\sigma \to \infty} \mu = 1. \tag{3.77}$$

This loss of market power by intermediate goods producers, in turn, eliminates the increasing returns to scale in the production of final consumption goods:

$$\lim_{\sigma \to \infty} y_i = \lim_{\mu \to 1} y_i = \phi. \tag{3.78}$$

Proof: A simple application of L'hôpital's rule is sufficient to demonstrate that firms producing intermediate goods are forced to set their price equal to marginal cost if inputs are perfect substitutes in production.

$$\lim_{\sigma \to \infty} \mu = \lim_{\sigma \to \infty} \frac{\sigma}{\sigma - 1} = 1$$

To see that this loss of market power by intermediate goods producers eliminates the increasing returns to scale in the production of final goods use a standard of variables to

re-write the limit as follows.

$$\begin{aligned}\lim_{\mu\to 1} y_i &= \lim_{\mu\to 1} \phi\left(\frac{1}{\mu}\right)\left(\left(\frac{\mu-1}{\mu}\right)\frac{1}{f}L_i\right)^{\mu-1} \\ &= \lim_{\mu\to 1} \phi\left(\frac{1}{\mu}\right) \lim_{\mu\to 1} e^{\frac{\ln\left(\left(\frac{\mu-1}{\mu}\right)\frac{1}{f}L_i\right)}{\frac{1}{\mu-1}}} \\ &= \lim_{\mu\to 1} A \lim_{\mu\to 1} B\end{aligned}$$

The change of variables has allowed us to rewrite the limit as a product of two limits A and B. The limiting behavior of A is trivial to calculate:

$$\lim_{\mu\to 1} A = \phi \lim_{\mu\to 1} \frac{1}{\mu} = \phi. \tag{3.79}$$

To complete the proof I need to show that

$$\lim_{\mu\to 1} B = 1. \tag{3.80}$$

Using the composition law for limits I can write

$$\lim_{\mu\to 1} B = e^{\lim_{\mu\to 1} \frac{\ln\left(\left(\frac{\mu-1}{\mu}\right)\frac{1}{f}L_i\right)}{\frac{1}{\mu-1}}}. \tag{3.81}$$

Now applying L'hôpital's rule yields:

$$\begin{aligned}&\lim_{\mu\to 1} \frac{\ln\left(\left(\frac{\mu-1}{\mu}\right)\frac{1}{f}L_i\right)}{\frac{1}{\mu-1}} \\ &\lim_{\mu\to 1} \frac{\frac{1}{\mu(\mu-1)}}{-\left(\frac{1}{\mu-1}\right)^2} \\ &\lim_{\mu\to 1} -\frac{\mu-1}{\mu} = 0.\end{aligned} \tag{3.82}$$

Therefore

$$\lim_{\mu\to 1} B = \lim_{\mu\to 1} e^0 = 1 \tag{3.83}$$

as required. ■

Claim: In the limit of large cities, per capita output converges to a constant.

$$\lim_{L_i \to \infty} y_i = \frac{c_1}{c_3} = \phi \tag{3.84}$$

Proof: A useful first step to deriving the limiting behavior of per capita output is to derive the limiting behavior of productivity.

$$\begin{aligned} \lim_{L_i \to \infty} prod_i &= \lim_{L_i \to \infty} e^{\frac{\ln\left(\frac{L_i}{c_2+c_3 L_i}\right)}{\frac{c_0 c_3 + c_1 c_3 L_i}{c_1 c_2 - c_0 c_3}}} \\ &= e^{\lim_{L_i \to \infty} \frac{\ln\left(\frac{L_i}{c_2+c_3 L_i}\right)}{\frac{c_0 c_3 + c_1 c_3 L_i}{c_1 c_2 - c_0 c_3}}} \end{aligned}$$

Another application of the L'hôpital's rule allows us to compute the limit in the exponent as follows.

$$\frac{\ln\left(\frac{L_i}{c_2+c_3 L_i}\right)}{\frac{c_0 c_3 + c_1 c_3 L_i}{c_1 c_2 - c_0 c_3}} = \lim_{L_i \to \infty} \frac{\frac{c_2+c_3 L_i}{L_i}\left[\frac{c_2}{(c_2+c_3 L_i)^2}\right]}{\frac{c_1 c_3}{c_1 c_2 - c_0 c_3}} = 0 \tag{3.85}$$

Therefore I know that $\lim_{L_i \to \infty} prod_i = \lim_{L_i \to \infty} e^0 = 1$. It is now straightforward to show that in the limit of large cities per capita output converges to a constant.

$$\begin{aligned} \lim_{L_i \to \infty} y_i &= \lim_{L_i \to \infty} \left(\frac{c_0 + c_1 L_i}{c_2 + c_3 L_i}\right)\left(\frac{L_i}{c_2 + c_3 L_i}\right)^{\frac{c_1 c_2 - c_0 c_3}{c_0 c_3 + c_1 c_3 L_i}} \\ &= \lim_{L_i \to \infty} \left(\frac{c_0 + c_1 L_i}{c_2 + c_3 L_i}\right) \lim_{L_i \to \infty} prod_i \\ &= \frac{c_1}{c_3} = \phi \end{aligned} \tag{3.86}$$

■

3.B Technical appendix

In this section I sketch the simulation procedure for comparing the $RMSE$ of the power law, logistic, and structural scaling models.

1. Get some data on population and per capita incomes: $(L_1, y_1), \ldots, (L_N, y_N)$.
2. Fit the power law (i.e., log-linear) scaling model to the data to get parameter vector $\hat{\theta}_p$ and an estimate of the in-sample mean squared error: $MSE_p(\hat{\theta}_p)$.
3. Fit the logistic scaling and structural scaling models to the data to get parameter vectors $\hat{\theta}_l$ and $\hat{\theta}_s$ as well as estimates of the in-sample mean squared error for both models: $RMSE_l(\hat{\theta}_l)$ and $RMSE_s(\hat{\theta}_s)$.
4. Calculate the following test statistics:

$$\hat{t}_1 = MSE_p(\hat{\theta}_p) - MSE_l(\hat{\theta}_l),\ \hat{t}_2 = MSE_s(\hat{\theta}_p) - MSE_s(\hat{\theta}_s) \tag{3.87}$$

   Note that $\hat{t}_1$ and $\hat{t}_2$ are just the difference between the mean squared errors of the power law scaling model (i.e., the null hypothesis) and the two different alternatives (i.e., the logistic and structural scaling models).
5. Simulate from the null model to get synthetic data: $(L'_1, y'_1), \ldots, (L'_N, y'_N)$
6. Fit the power law null model to the synthetic data to to get parameter vector $\tilde{\theta}_p$ and an estimate of the in-sample mean squared error: $MSE_p(\tilde{\theta}_p)$.
7. Fit the logistic scaling and structural scaling models to the synthetic data to get parameter vectors $\tilde{\theta}_l$ and $\tilde{\theta}_s$ as well as estimates of the in-sample mean squared error for both models: $MSE_l(\tilde{\theta}_l)$ and $MSE_s(\tilde{\theta}_s)$.
8. Calculate the following test statistics:

$$\tilde{T}_1 = MSE_p(\tilde{\theta}_p) - MSE_p(\tilde{\theta}_l),\ \tilde{T}_2 = MSE_p(\tilde{\theta}_p) - MSE_s(\tilde{\theta}_s) \tag{3.88}$$

9. Repeat steps 5-8 a large number of times, N, to get an estimate of the distribution of t_1 and t_2 under the null hypothesis.
10. Compute the p-value as

$$p = \frac{1 + \#\{\tilde{T} > \hat{t}\}}{1 + N} \tag{3.89}$$

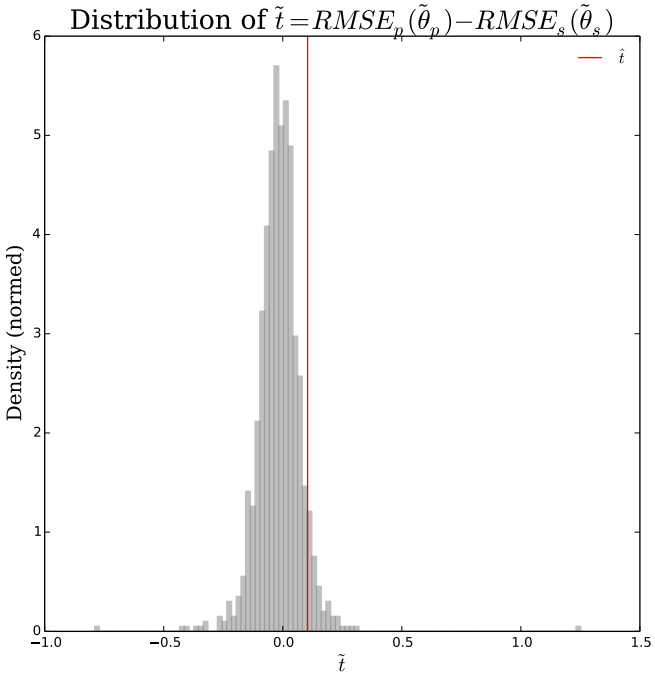

Figure 3.4: Difference in RMSE between the structural scaling model and power law scaling model is only marginally significant (p-value of 0.09).

Bibliography

Abdel-Rahman, H. M. (1988). Product differentiation, monopolistic competition and city size. *Regional Science and Urban Economics 18*(1), 69–86.

Amemiya, T. (1985). *Advanced econometrics*. Harvard university press.

Arlot, S. and A. Celisse (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys 4*, 40–79.

Barro, R. J. and X. Sala-i Martin (2003). *Economic growth*. The MIT Press.

Behnel, S., R. Bradshaw, C. Citro, L. Dalcin, D. S. Seljebotn, and K. Smith (2011). Cython: The best of both worlds. *Computing in Science & Engineering 13*(2), 31–39.

Behrens, K. and Y. Murata (2007). General equilibrium models of monopolistic competition: a new approach. *Journal of Economic Theory 136*(1), 776–787.

Beirlant, J. (2004). *Statistics of extremes: theory and applications*, Volume 558. John Wiley & Sons Inc.

Benito, E. (2008). Size, growth and bank dynamics.

Berger, A., A. Kashyap, J. Scalise, M. Gertler, and B. Friedman (1995). The transformation of the us banking industry: What a long, strange trip it's been. *Brookings papers on economic activity 1995*(2), 55–218.

Bettencourt, L. and G. West (2010). A unified theory of urban living. *Nature 467*(7318), 912–913.

Bettencourt, L. M. (2013). The origins of scaling in cities. *Science 340*(6139), 1438–1441.

Bettencourt, L. M., J. Lobo, D. Helbing, C. Kühnert, and G. B. West (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences 104*(17), 7301–7306.

Bettencourt, L. M., J. Lobo, D. Strumsky, and G. B. West (2010). Urban scaling and its

deviations: Revealing the structure of wealth, innovation and crime across cities. *PloS one 5*(11), e13541.

Bogacki, P. and L. F. Shampine (1989). A 3 (2) pair of runge-kutta formulas. *Applied Mathematics Letters 2*(4), 321–325.

Bremus, F., C. Buch, K. Russ, and M. Schnitzer (2013, May). Big banks and macroeconomic outcomes: Theory and cross-country evidence of granularity. Working Paper 19093, National Bureau of Economic Research.

Brock, W. A. and A. G. Malliaris (1989). *Differential equations, stability and chaos in dynamic economics*. North-Holland Amsterdam.

Brown, P. N., G. D. Byrne, and A. C. Hindmarsh (1989). Vode: A variable-coefficient ode solver. *SIAM journal on scientific and statistical computing 10*(5), 1038–1051.

Butcher, J. C. (2008). *Numerical methods for ordinary differential equations*. John Wiley & Sons.

Campbell, J. R. and H. A. Hopenhayn (2005). Market size matters. *The Journal of Industrial Economics 53*(1), 1–25.

Cash, J. R. and A. H. Karp (1990). A variable order runge-kutta method for initial value problems with rapidly varying right-hand sides. *ACM Transactions on Mathematical Software (TOMS) 16*(3), 201–222.

Cass, D. (1965). Optimum growth in an aggregative model of capital accumulation. *The Review of Economic Studies 32*(3), 233–240.

Champernowne, D. (1953). A model of income distribution. *The Economic Journal 63*(250), 318–351.

Chiang, A. and K. Wainwright (2005). *Fundamental methods of mathematical economics* (4th ed.). McGraw-Hill.

Ciccone, A. and R. E. Hall (1996). Productivity and the density of economic activity. *The American Economic Review 86*(1), 54–70.

Claeskens, G. and N. L. Hjort (2008). *Model selection and model averaging*, Volume 330. Cambridge University Press Cambridge.

Clauset, A., C. Shalizi, and M. Newman (2009). Power-law distributions in empirical data. *SIAM Review 51*(4), 661–703.

Clauset, A., M. Young, and K. Gleditsch (2007). On the frequency of severe terrorist events. *Journal of Conflict Resolution 51*(1), 58–87.

Davidson, R. and J. G. MacKinnon (1993). *Estimation and inference in econometrics*. Oxford University Press.

Davison, A. C. (1997). *Bootstrap methods and their application*, Volume 1. Cambridge university press.

Dixit, A. K. and J. E. Stiglitz (1977). Monopolistic competition and optimum product diversity. *The American Economic Review 67*(3), 297–308.

Dormand, J. R. and P. J. Prince (1980). A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics 6*(1), 19–26.

Eeckhout, J. and P. Kircher (2010). Sorting and decentralized price competition. *Econometrica 78*(2), 539–574.

Eeckhout, J. and P. Kircher (2012). Assortative matching with large firms: Span of control over more versus better workers.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics 7*(1), 1–26.

Efron, B. and B. Efron (1982). *The jackknife, the bootstrap and other resampling plans*, Volume 38. SIAM.

Embrechts, P., C. Kluppelberg, and T. Mikosch (1997). *Modelling extremal events for insurance and finance*, Volume 33. Springer Verlag.

Ennis, H. (2001). On the size distribution of banks. *Economic Quarterly-Federal Reserve Bank of Richmond 87*(4), 1–26.

Feenstra, R. C., R. Inklaar, and M. Timmer (2013). The next generation of the penn world table available for download at www.ggdc.net/pwt.

Fehlberg, E. (1968). Classical fifth-, sixth-, seventh-, and eighth-order runge-kutta formulas with stepsize control. Technical Report NASA TR R 287, National Aeronautics and Space Administration (NASA).

Fujita, M. (1988). A monopolistic competition model of spatial agglomeration: Differentiated product approach. *Regional Science and Urban Economics 18*(1), 87–124.

Fujita, M., P. R. Krugman, and A. J. Venables (1999). *The spatial economy: cities, regions and international trade*, Volume 213. Wiley Online Library.

Gabaix, X. (1999). Zipf's law for cities: an explanation. *The Quarterly Journal of Economics 114*(3), 739–767.

Gabaix, X. (2008). Power laws in economics and finance. Technical report, National Bureau of Economic Research.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica 79*(3), 733–772.

Gabaix, X. and R. Ibragimov (2011). Rank- 1/2: a simple way to improve the ols estimation of tail exponents. *Journal of Business and Economic Statistics 29*(1), 24–39.

Gear, C. W. (1971). *Numerical initial value problems in ordinary differential equations*. Prentice Hall PTR.

Glaeser, E. L. (2008). *Cities, agglomeration, and spatial equilibrium (The Lindhaul Lectures)*. Oxford University Press.

Glaeser, E. L. and J. D. Gottlieb (2009). The wealth of cities: Agglomeration economies and spatial equilibrium in the united states. *Journal of Economic Literature 47*(4), 983–1028.

Glaeser, E. L. and J. Gyourko (2005). Urban decline and durable housing. *Journal of Political Economy 113*(2), 345–375.

Goddard, J., P. Molyneux, and J. O. Wilson (2004). Dynamics of growth and profitability in banking. *Journal of Money, Credit and Banking*, 1069–1090.

Hairer, E., S. P. Nørsett, and G. Wanner (1993). *Solving ordinary differential equations* (2nd ed.), Volume I. Springer.

Hall, P. (1982). On some simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37–42.

Hall, R. E. and C. I. Jones (1999). Why do some countries produce so much more output per worker than others? *The quarterly journal of economics 114*(1), 83–116.

Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics 167*(1), 38–46.

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 1163–1174.

Hindmarsh, A. and K. Radhakrishnan (1993). Description and use of lsode, the livermore solver for ordinary differential equations. *Lawrence Livermore National Laboratory Report UCRL-ID-113855*.

Hornbeck, R. (2012). The enduring impact of the american dust bowl: Short and long-run adjustments to environmental catastrophe. *American Economic Review 102*(4), 1477–1507.

Hubbard, T. P., R. Kirkegaard, and H. J. Paarsch (2011). Using economic theory to guide numerical analysis: solving for equilibria in models of asymmetric first-price auctions. *Computational Economics*, 1–26.

Hubbard, T. P., T. Li, and H. J. Paarsch (2012). Semiparametric estimation in models of first-price, sealed-bid auctions with affiliation. *Journal of Econometrics 168*(1), 4–16.

Hubbard, T. P. and H. J. Paarsch (2009). Investigating bid preferences at low-price, sealed-bid auctions with endogenous participation. *International Journal of Industrial Organization 27*(1), 1–14.

Iserles, A. (2009). *A first course in the numerical analysis of differential equations*, Volume 44. Cambridge University Press.

Janicki, H. and E. S. Prescott (2006). Changes in the size distribution of us banks: 1960-2005. *Economic Quarterly-Federal Reserve Bank of Richmond 92*(4), 291.

Jones, K. and T. Critchfield (2005). Consolidation in the us banking industry: Is the long, strange trip about to end? *Strange Trip About to End*.

Judd, K. L. (1998). *Numerical methods in economics*. The MIT press.

Kamien, M. I. and N. L. Schwartz (2012). *Dynamic optimization: the calculus of variations and optimal control in economics and management*. DoverPublications. com.

Koopmans, T. C. (1965). On the concept of optimal economic growth". *The Econometric Approach to Development Planning*.

Kotz, S. and S. Nadarajah (2000). *Extreme value distributions: theory and applications*. World Scientific Publishing Company.

Krugman, P. R. (1991). Increasing returns and economic geography. *Journal of Political Economyolitical Economy 99*(3), 483–499.

Krugman, P. R. (1996). *The self-organizing economy*. Blackwell Publishers Cambridge, Massachusetts.

Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quarterly of applied mathematics 2*, 164–168.

Mandelbrot, B. (1963). The variation of certain speculative prices. *The journal of business 36*(4), 394–419.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics 11*(2), 431–441.

Mason, D. (1982). Laws of large numbers for sums of extreme values. *The Annals of Probability*, 754–764.

Merson, R. (1957). An operational method for the study of integration processes. In *Proc. Symp. Data Processing*, pp. 1–25.

Mills, E. S. (1967). An aggregative model of resource allocation in a metropolitan area. *The American Economic Review 57*(2), 197–210.

Mortensen, D. T. and C. A. Pissarides (1994). Job creation and job destruction in the theory of unemployment. *The review of economic studies 61*(3), 397–415.

Newman, M. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics 46*(5), 323–351.

Nocedal, J. and S. Wright (2006). *Numerical optimization* (2nd ed. ed.), Volume XXII of *Springer Series in Operations Research and Financial Engineering*. Springer, New York.

Pareto, V. (1896). Cours d\'economie politique.

Peterson, P. (2009). F2py: a tool for connecting fortran and python programs. *International Journal of Computational Science and Engineering 4*(4), 296–305.

Pissarides, C. A. (1985). Short-run equilibrium dynamics of unemployment vacancies, and real wages. *American Economic Review 75*(4), 676–90.

Pontryagin, L. (1959). Optimal control processes. *Usp. Mat. Nauk 14*(3).

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (2009). *Numerical recipes in C+: the art of scientific computing*, Volume 994. Cambridge University Press Cambridge.

Racine, J. (1997). Feasible cross-validatory model selection for general stationary processes. *Journal of Applied Econometrics 12*(2), 169–179.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data:¡ i¿ hv¡/i¿-block cross-validation. *Journal of Econometrics 99*(1), 39–61.

Ramsey, F. P. (1928). A mathematical theory of saving. *The Economic Journal 38*(152), 543–559.

Resnick, S. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling*, Volume 10. Springer Verlag.

Rivera-Batiz, F. L. (1988). Increasing returns, monopolistic competition, and agglomeration economies in consumption and production. *Regional Science and Urban Economics 18*(1), 125–153.

Roback, J. (1982). Wages, rents, and the quality of life. *The Journal of Political Economy 90*(6), 1257–1278.

Rosen, S. (1979). Wage-based indexes of urban quality of life. In P. Mieszkowski and M. Straszheim (Eds.), *Current issues in urban economics*, Volume 3, pp. 74–104. Baltimore and London: Johns Hopkins University Press.

Sargent, T. and J. Stachurski (2013). *Quantitative Economics*.

Sato, R. (1963). Fiscal policy in a neo-classical growth model: An analysis of time required for equilibrating adjustment. *The Review of Economic Studies 30*(1), 16–23.

Shalizi, C. (2014a). *Advanced data analysis from an elementary point of view*. Cambridge University Press.

Shalizi, C. R. (2014b). Scaling and hierarchy in urban economies. *PLoS ONE*.

Shampine, L. F. (1986). Some practical runge-kutta formulas. *Mathematics of Computation 46*(173), 135–150.

Simon, H. (1955). On a class of skew distribution functions. *Biometrika*, 425–440.

Smith, W. T. (2006). A closed form solution to the ramsey model. *Contributions in Macroeconomics 6*(1).

Solow, R. M. (1956). A contribution to the theory of economic growth. *The quarterly journal of economics 70*(1), 65–94.

Spence, M. (1974). Competitive and optimal responses to signals: An analysis of efficiency and distribution. *Journal of Economic Theory 7*(3), 296–332.

Spence, M. (1976). Product selection, fixed costs, and monopolistic competition. *The Review of Economic Studies 43*(2), 217–235.

Stachurski, J. (2009). *Economic dynamics: theory and computation*. MIT Press.

Steindl, J. (1965). *Random processes and the growth of firms: A study of the Pareto law*. Griffin London.

Syverson, C. (2007). Prices, spatial competition and heterogeneous producers: An empirical test*. *The Journal of Industrial Economics 55*(2), 197–222.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society 57*, 307–333.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817–838.

Wilks, S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics 9*(1), 60–62.

Zhang, W.-B. (2005). *Differential equations, bifurcations, and chaos in economics*, Volume 68. World Scientific.

Zhelobodko, E., S. Kokovin, M. Parenti, and J.-F. Thisse (2012). Monopolistic competition: Beyond the constant elasticity of substitution. *Econometrica 80*(6), 2765–2784.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Addison-Wesley Press, Cambridge, Massachusetts.