2-13-2014

A Survey of Lack-of-fit Tests Based on Sums of Ordered Residuals

Mohammad Hattab

Mohammad Wasef Hattab

Candidate

Mathematics and Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Ronald Christensen, Chairperson

Gabriel Huerta

Erik Erhardt

Huining Kang

A Survey of Lack-of-fit Tests Based on Sums of Ordered Residuals

by

Mohammad Wasef Hattab

B.S., Mathematics, University of Jordan, 2003
M.S., Mathematics, University of Jordan, 2005
M.S., Statistics, University of New Mexico, 2010

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Statistics

The University of New Mexico
Albuquerque, New Mexico

December, 2013

Dedication

To Dad and Mom

To my Sisters and Brothers

Acknowledgments

First and foremost, I am very thankful to Allah the most gracious and the most merciful.

I would like to express my love and gratitude to my parents (Wasef and Lamia) for everything they gave to me and are still giving to me. My PhD would have been unattainable without their love and encouragement. Also, my deepest thanks to my sisters (Hala and Hiba) for always taking care of me, and to my brothers (Samer, Osama, and Anas) for always being there for me.

I am sincerely grateful to my advisor, Dr. Ronald Christensen, for his excellent classes, continuing support and valuable guidance throughout my PhD journey. I could not have obtained my PhD without his support.

I would like to thank Dr. Erik Erhardt for his continuous encouragement and for putting his faith in me, my friend Dr. Yong Lin for his help and advice in the dissertation process and Dr. Curtis Storlie for being so supportive to me at the beginning of my study.

I would like to express special thanks to Dr. Mufid Azzam, a faculty member at the University of Jordan, who introduced me to the world of statistics and inspired me to pursue a PhD in statistics.

I would like to take this opportunity to express my sincere appreciation to all my instructors at the University of Jordan and the University of New Mexico.

I am thankful to the Graduate committee in the Department of Mathematics and Statistics at the University of New Mexico for giving me the opportunity to pursue my studies. Also I would like to thank Donna George, Roxanne Littlefield, Gail Mercer and Ana Lombard for their advice.

Finally, I would like to thank people of Albuquerque who made me feel welcome in their lovely city, especially Mike and Laurel Edenburn and Khalid Ifzarene.

A Survey of Lack-of-fit Tests Based on Sums of Ordered Residuals

by

Mohammad Wasef Hattab

B.S., Mathematics, University of Jordan, 2003

M.S., Mathematics, University of Jordan, 2005

M.S., Statistics, University of New Mexico, 2010

Ph.D, Statistics, University of New Mexico, 2013

Abstract

Christensen and Lin (2014), henceforth C-L, suggested two lack-of-fit tests to assess the adequacy of a linear model based on partial sums of residuals. In particular, their tests evaluated the adequacy of the mean function. Their tests relied on asymptotic results without requiring small sample normality. We extend this research by proposing additional tests based on partial sums of residuals. The asymptotic distribution for each test statistic is found so that the P value can be efficiently approximated. To assess their strengths and weaknesses, the C-L tests and the new tests are compared in different scenarios by simulation. We propose new tests based on partial sums of absolute residuals. Previous partial sums of residuals test have used signed residuals whose values when summed can cancel each other out. The use of absolute residuals ,which requires small sample normality, allows detection of lack of fit that was previously not possible with partial sums of residuals.

KEY WORDS: Lack-of-fit tests; Partial sums of residuals; Monte Carlo Simulations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The assumed model

Consider the linear model

$$\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{e}_n, \quad E(\boldsymbol{e}_n) = \boldsymbol{0}, \qquad Cov(\boldsymbol{e}_n) = \sigma^2\boldsymbol{I}_n, \tag{1.1}$$

where $\boldsymbol{Y}_n$ is a $n \times 1$ vector of observable random values, $\boldsymbol{X}_n$ is an $n \times p$ known model matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, $\boldsymbol{I}_n$ is an identity matrix of size n, σ^2 is some unknown parameter and $\boldsymbol{e}_n$ is an $n \times 1$ vector of independent and unobservable errors.

1.2 Outline of the dissertation

Christensen and Lin (2014), henceforth C-L, were interested in assessing the validity of the mean function specification and proposed two lack-of-fit tests based on partial sums of residuals. We present additional test statistics based on partial sums of residuals and partial sums of absolute residuals. We derive their asymptotic distributions, explore their small sample behavior and evaluate their effectiveness. We introduce an effective approximation to P values through simulations.

This dissertation is organized as follows. Chapter 2 reviews C-L's work and presents new test statistics. Chapter 3 gives the asymptotic distributions of these statistics and assumptions required to achieve convergence in distribution. Chapter 4 introduces ordering methods and a consistent estimator of σ. Approximation of P values for each test through Monte Carlo simulations is examined in Chapter 5. Several examples and various simulations with power comparisons along with recommendations are given in Chapter 6. The proofs of the asymptotic results are given in the appendix.

Chapter 2

Lack of fit tests

2.1 C-L

C-L, were interested in assessing the validity of the mean function $E(\boldsymbol{Y_n}) = \boldsymbol{X_n\beta}$ and proposed two lack-of-fit tests based on partial sums of residuals. They examined the following statistics

$$T_n = \frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\hat{\sigma}_n} \right|$$

and

$$Q_n = a_{\tilde{n}} \max_{1 \le m \le \tilde{n}} \frac{1}{\sqrt{m}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\hat{\sigma}_n} \right| - b_{\tilde{n}}.$$

where y_i and $\boldsymbol{x}_i^T$ are the ith rows of $\boldsymbol{Y_n}$ and $\boldsymbol{X_n}$, respectively, $\hat{\boldsymbol{\beta}}_n$ is the least square estimate (LSE) of $\boldsymbol{\beta}$, $\tilde{n} = \lceil n/(\log\log n)^{1+\delta} \rceil$, $a_{\tilde{n}} = \sqrt{2\log\log\tilde{n}}$, $b_{\tilde{n}} = (a_{\tilde{n}})^2 + \log a_{\tilde{n}} - \log(\sqrt{2\pi})$, and $\hat{\sigma}_n$ is a consistent estimate of σ. As in Fan and Huang (2001) and Christensen and Sun (2010), the range of maximization $\tilde{n}$ is chosen for the asymptotic results to work. In particular, T_n requires $\delta > 0$ whereas Q_n requires $\delta > 1$. Only

$\tilde{n}$ residuals out of n are used in the statistics. These $\tilde{n}$ residuals must be ordered and the ordering affects the tests ability to reveal the lack of fit. Note the slow convergence of $\tilde{n}$ to infinity compared to the sample size n. This restriction to $\tilde{n}$ is not needed if $\boldsymbol{\beta}$ and σ are known. C-L proved that T_n and Q_n converge in distribution to random variables T and Q respectively, where

$$Pr[T < t] = \frac{4}{\pi} \sum_{m=0}^{\infty} \frac{(-1)^m}{2m+1} \exp(-(2m+1)^2\pi^2/8t^2) \quad \text{for} \quad t > 0 \tag{2.1}$$

and

$$Pr[Q < t] = \exp\left[-\exp(-t)\right] \tag{2.2}$$

The limiting distribution of Q_n is a standard Gumbel distribution.

T_n and Q_n both examine the maximum of absolute values of partial sums of residuals. T_n divides each partial sum by the square root of $\tilde{n}$ whereas in Q_n each is divided by the square root of its number of terms. High observed test statistics suggest lack of fit in model (1.1). C-L noticed that Q_n is more sensitive than T_n in detecting lack of fit that occurs in the first few residuals. T_n outperforms Q_n if the lack of fit occurs at relatively higher ordered observations. Of course, neither of these tests will detect lack of fit if the $\tilde{n}$ observations in the partial sums are fitted well. Small partial sums lead to small test statistics. This shows the importance of the ordering of the data.

2.2 C-H

Table 2.1 lists T_n and Q_n along with four new test statistics. The tests depend on one of two partial sum statistics,

$$S_m = \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\hat{\sigma}_n}$$

and

$$K_m = \sum_{i=1}^{m} \left| \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\hat{\sigma}_n} \right|$$

Table 2.1 also lists the normalization that leads to the asymptotic distribution of the statistic, the 95th percentile of the asymptotic distribution and the equation number where the asymptotic distribution is given. The asymptotic distributions for the new statistics are introduced in Chapter 3.

Table 2.1: Test Statistics

Label	Statistic	95% Percentile	Asym. Dist.
T_n	$\frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \lvert S_m \rvert$	2.241	(2.1)
Q_n	$a_{\tilde{n}} \max_{1 \le m \le \tilde{n}} \frac{1}{\sqrt{m}} \lvert S_m \rvert - b_{\tilde{n}}$	2.970	(2.2)
W_n	$\frac{1}{\sqrt{\tilde{n}^3}} \sum_{m=1}^{\tilde{n}} \lvert S_m \rvert$	1.139	(3.1)
V_n	$\frac{1}{\tilde{n}^2} \sum_{m=1}^{\tilde{n}} (S_m)^2$	1.656	(3.2)
R_n	$\frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \left(\frac{K_m}{\sqrt{1 - \frac{2}{\pi}}} - m\sqrt{\frac{2}{\pi - 2}} \right)$	1.959	(3.3)
H_n	$a_{\tilde{n}} \max_{1 \le m \le \tilde{n}} \left(\frac{K_m}{\sqrt{m(1 - \frac{2}{\pi})}} - \sqrt{\frac{2m}{\pi - 2}} \right) - b_{\tilde{n}} + \log 2$	2.970	(2.2)

Just as in T_n and Q_n, the range of our partial sum statistics is limited by $\tilde{n}$. For the asymptotic results to work, T_n, W_n, V_n and R_n require $\delta > 0$ whereas $\delta > 1$ for Q_n and H_n. Note that Q_n and H_n have the same limiting distribution. The new lack of fit test statistics can be classified into two groups:

- The first group includes statistics labeled W_n and V_n. These examine sums of functions of S_m as opposed to taking maximums as in T_n and Q_n. We found that both T_n and Q_n are less capable in detecting a lack of fit that happens at middle or higher ordered observations or is not concentrated over series of successive residuals but rather they are scattered among the observations. When we would like to evaluate the collective effect of the residuals, W_n and V_n are recommended. These new tests may detect lack of fit when it occurs over a specific segment of data or in different locations.

- We also propose test statistics built upon the absolute residuals rather than the signed residuals. Using the signed residuals in partial sum statistics can reduce power. Residuals with similar magnitudes but different signs can cancel each other when taking the partial sums, which leads to large P values. Using the absolute residuals considered in R_n and H_n examines the maximum of partial sums of absolute residuals whereas T_n and Q_n evaluate the maximum of absolute values of partial sums of residuals. Unlike the other tests, the asymptotic distribution of R_n and H_n depends on an assumption of small sample normality.

To execute any of these tests, the observations must be totally ordered according to some criteria. This ordering, usually performed on the basis of some function of the predictor variables, is crucial to the effectiveness of the statistics. Ordering is discussed in Section 4.2. The asymptotic distributions for each of the test statistics above are found under the null model (1.1). The rationale behind these tests are the same. High values of test statistics indicate that the proposed linear model is not adequate and should be revised. These test statistics tend to get large if the standardized residuals are large in magnitude which is in turn a sign of lack of fit.

The P value for any test statistic, for example W_n, is defined as $Pr[W_n \geq x_n]$ where x_n is the observed value of the test statistic. The test statistics suffer from slow convergence leading to poor asymptotic approximation of the P values for small, moderate, and even somewhat large samples. An effective Monte Carlo simulation is introduced to approximate the P values. The simulation method is presented in Chapter 5.

Chapter 3

Asymptotic distributions

To establish the asymptotic distributions of our test statistics we assume throughout that:

(a) $\frac{1}{n}\boldsymbol{X}_n^T\boldsymbol{X}_n$ converges (in probability) to $\boldsymbol{A}$, where $\boldsymbol{A}$ is some positive definite matrix.

(b) $\hat{\sigma}_n = \sigma + O_p(1/\sqrt{n})$.

If known values of σ and β replace $\hat{\sigma}_n$ and $\hat{\boldsymbol{\beta}}_n$ in our test statistics, the asymptotic distributions are direct applications of results in Erdos and Kac (1945) or Darling and Erdos (1956) with $\tilde{n} = n$. In practice, both σ and β are unknown. The assumptions are used to establish that the asymptotic distributions hold when the parameters are estimated. Assumption (a) implies that the least squares estimate $\hat{\boldsymbol{\beta}}_n$ converges in probability to $\boldsymbol{\beta}$. Complications arise in the asymptotics because unlike the errors in model (1.1), the residuals are not independent. To deal with this dependency, we need to restrict the range of the partial sums to $\tilde{n} = \lceil n/(\log\log n)^{1+\delta}\rceil$. If $n - \tilde{n}$ residuals are excluded from the sums, it is important to include the residuals that are most likely to display lack of fit. The ordering of the data plays an important role in detecting lack of fit although it does not affect the asymptotic distributions. Ordering is discussed in Section 4.2.

3.1 W_n and V_n

Theorem 1. If assumptions (a) and (b) are satisfied,

$$W_n \xrightarrow{\mathcal{L}} W$$

where W has a known distribution,

$$Pr[W \le w] = \int_0^w \sqrt{\frac{3}{\pi}} u^{-1} \sum_{j=1}^{\infty} C_j \exp(-v_j) v_j^{\frac{2}{3}} U(\frac{1}{6}, \frac{4}{3}, v_j)\, du \quad , \quad w > 0 \qquad (3.1)$$

where

$$U(\frac{1}{6}, \frac{4}{3}, x) = \frac{1}{\Gamma(\frac{1}{6})} \int_0^\infty \exp(-tx) t^{\frac{-5}{6}} (1+t)^{\frac{1}{6}}\, dt,$$
$$v_j = \frac{2(a'_j)^3}{27u^2},$$
$$C_j = \frac{1 + 3\int_0^{a'_j} Ai(-r)\, dr}{3a'_j Ai(-a'_j)},$$
$$Ai(z) = \frac{1}{\pi} \int_0^\infty \cos\left(\frac{t^3}{3} + tz\right) dt$$

$U(\frac{1}{6}, \frac{4}{3}, x)$ is a confluent hypergeometric function, $Ai(z)$ is the Airy integral, $z = -a'_j$, $j = 1, 2, \cdots$, are the zeros of $Ai'(z)$, arranged so that $0 < a'_1 < a'_2 < \ldots < a'_j < \ldots$, and Γ is the gamma function.

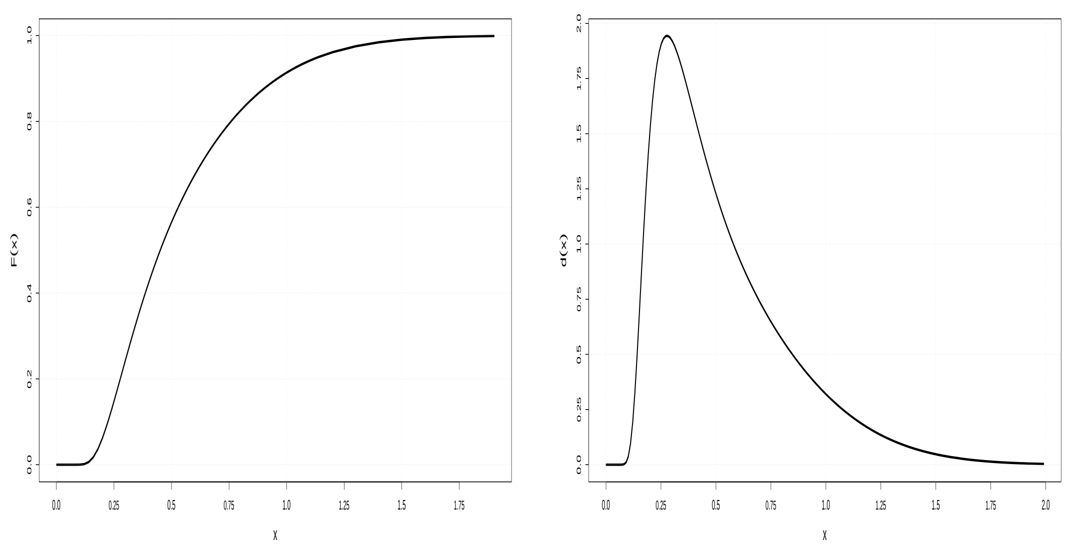
The proof is given in Appendix A. Essentially, the proof depends on using Erdos and Kac (1945) part (4) who gave the Laplace transform formula of the distribution function of W which is of course not useful for applications. That formula, not given here, is very complicated. Takács (1993) worked out the cumulative distribution function (cdf) of W given in (3.1).

The cdf of W involves an infinite sum. For practical computations, using $j = 7$ as an upper limit for the sum gives reasonable approximations to the distribution function.

Table 3.1 lists basic properties of W. Its cdf and pdf are produced in Figure 3.1.

Table 3.1: Basic Properties of W

2.5%	5%	25%	50%	Mean	Var	75%	95%	97.5%
0.169	0.192	0.303	0.451	0.532	0.092	0.688	1.139	1.300



(a) The cdf of W (b) The pdf of W

Figure 3.1: The cdf and the pdf of W

Theorem 2. If assumptions (a) and (b) are satisfied,

$$V_n \xrightarrow{\mathcal{L}} V$$

with

$$Pr[V < v] = \sqrt{2}\sum_{j=0}^{\infty}(-1)^j \delta_j \operatorname{erfc}\left(\frac{4j+1}{2\sqrt{2v}}\right) \quad , \quad v > 0 \tag{3.2}$$

where $\delta_j = \Gamma(j+\frac{1}{2})/\sqrt{\pi}j!$, Γ is the gamma function, and erfc is the complementary error function defined by $\operatorname{erfc(x)} = \frac{2}{\sqrt{\pi}}\int_x^{\infty}\exp(-t^2/2)\,dt$.
The proof is given in Appendix A. The cdf is well approximated if the infinite sum is taken from $j = 0$ to $j = 7$. This form of the cdf was given by Cameron and Martin (1944) and much simpler than the form used by Erdos and Kac part (3) that states

$$Pr[V < v] = \frac{\pi^{-3/2}}{4}\int_0^{v/2}\int_0^{\pi/2} u^{-3/2}(\cos t)^{-1/2}\theta'\left(\frac{t}{2}, \exp(-\frac{1}{4}u)\right)\,dt\,du \quad , \quad v > 0$$

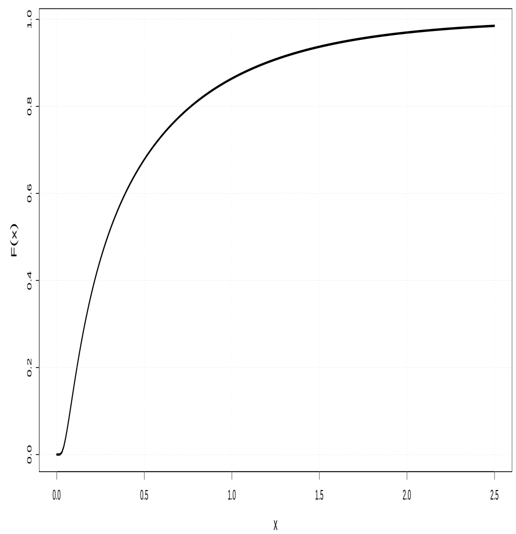
and

$$\theta(z,q) = 2\sum_{n=0}^{\infty}(-1)^n q^{(2n+1)^2/4}\sin(2n+1)z, \quad \theta' = \frac{\partial}{\partial z}\theta$$

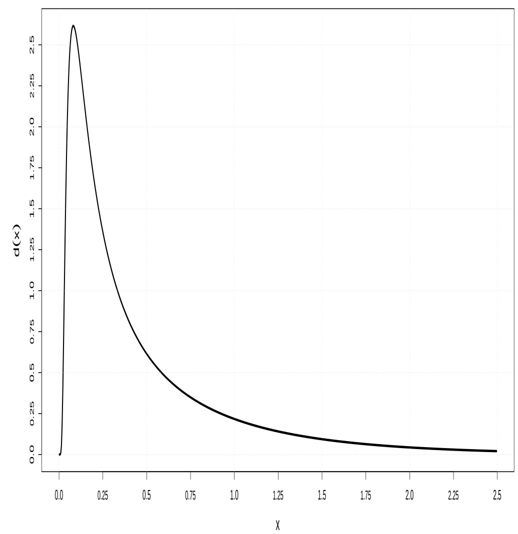
Table 3.2 lists basic properties of V. Its cdf and pdf are given in Figure 3.2.

Table 3.2: Basic Properties of V

2.5%	5%	25%	50%	Mean	Var	75%	95%	97.5%
0.044	0.057	0.137	0.290	0.500	0.333	0.638	1.656	2.135



(a) The cdf of V



(b) The pdf of V

Figure 3.2: The cdf and the pdf of V

3.2 Absolute Residuals

The earlier lack of fit test statistics are based on the partial sums of ordinary residuals. Those test statistics indicate the inadequacy of the model when the first residuals in the ordering are dominated by a series of either relatively high positive residuals or low negative residuals. Their ability to detect lack-of-fit is hurt by either an ordering for the data that has the first observations being from a region where the model fits well or the first residuals consist of both high and low residuals that cancel each other out when taking the partial sums. In the latter case, we suggest using the absolute value of the residuals rather than the signed residuals.

The next example shows the need to use absolute residuals rather than the signed residuals in some cases. The example is only for illustrative purposes since the lack of fit is clear when the data are plotted and tests are not needed.

EXAMPLE 3.1. Figure 3.3 shows simulated data along with the fitted regression line. The lack of fit is clear and severe and yet none of the test statistics based on the partial sums of the residuals came close to detecting it. The p-values are: T_n, 0.68; Q_n, 0.20; V_n, 0.40; and W_n, 0.78. The tests based on the partial sums of the absolute residuals report P values close to 0. First notice that the covariate x is distributed symmetrically around its center 7. The ordering method orders the data starting from those farthest from the center, so both sides will be represented almost equally in the first $\tilde{n}$ observations. These two sides carry residuals that have similar magnitudes but with different signs, so they cancel each other when taking the partial sums. This leads to small test statistics and hence large P values explaining the inability of the tests based on the partial sums of the residuals to detect the lack of fit. For the tests based on the partial sums of the absolute residuals, each absolute residual contributed positively to the partial sums producing - correctly - a large statistic and hence small P values.

To normalize the absolute residuals in R_n and H_n, we assume normal errors. Note that small sample normality was not required for any of the earlier tests. Note also we need to assume a certain distribution, not necessarily normal, to standardize the absolute residuals correctly. If a non-normal distribution is assumed such as a t distribution, then they should be standardized accordingly yielding different statistics from R_n and H_n.

The mean and the standard deviation of the absolute value of a standard normal random variable, Z, are $\sqrt{\frac{2}{\pi}}$ and $\sqrt{1-\frac{2}{\pi}}$. It follows that, $\frac{|Z|}{\sqrt{1-\frac{2}{\pi}}} - \sqrt{\frac{2}{\pi-2}}$ has mean 0 and variance 1.

The next theorem introduces the statistics R_n and H_n that are based on absolute residuals.

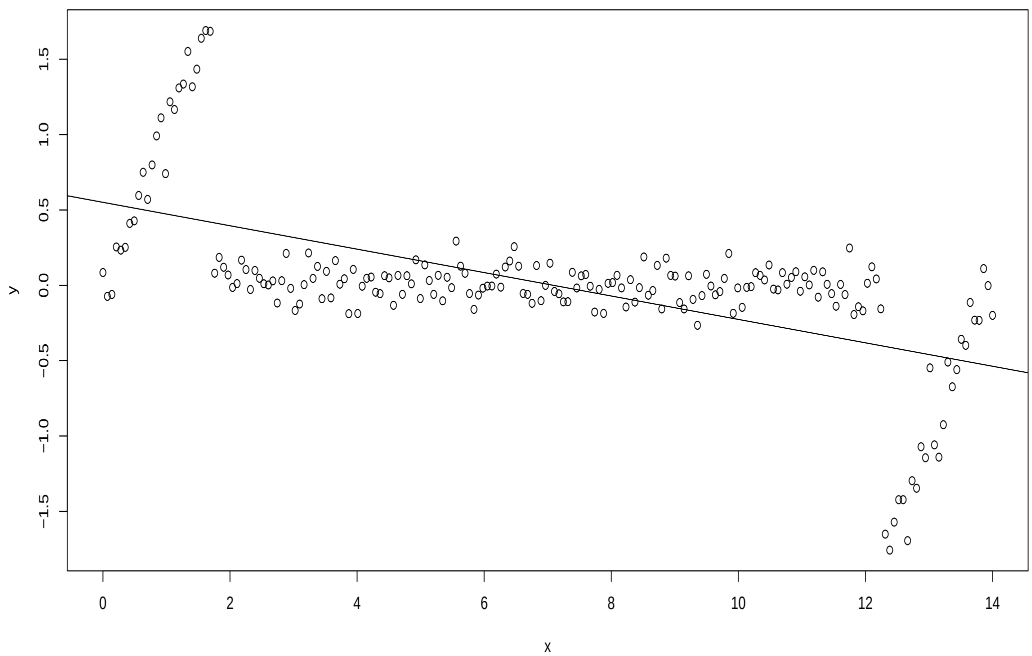


Figure 3.3: Example 3.1; Simulated data with the fitted regression line

Theorem 3. If $\boldsymbol{Y_n} \sim N(\boldsymbol{X}_n\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$ and assumptions (a) and (b) are satisfied, and R_n and H_n are defined in Table 2.1, then

$$R_n \xrightarrow{\mathcal{L}} U$$

where U has the half-normal distribution

$$Pr[U < u] = \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \int_0^u \exp(-\frac{x^2}{2})\, dx \quad \text{for} \quad u > 0 \tag{3.3}$$

and

$$H_n \xrightarrow{\mathcal{L}} Q$$

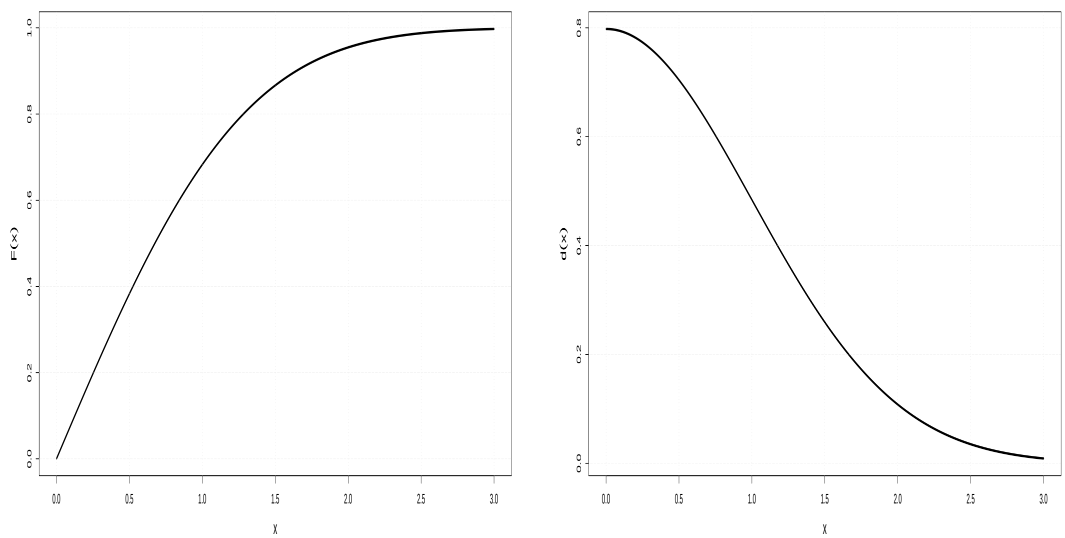
where $a_{\tilde{n}}$ and $b_{\tilde{n}}$ are defined as for Q_n, and Q has the distribution in equation (2.2) which is a standard Gumbel distribution.

See Appendix A for the proof. The main difference between R_n and H_n can be understood in light of the difference between T_n and Q_n. H_n is more sensitive than R_n in detecting lack of fit that occurs in the first few absolute residuals. R_n outperforms H_n if the lack of fit occurs at relatively higher ordered observations. As more terms are included in R_n and H_n, the partial sums of absolute residuals are getting subtracted by a larger number, $m\sqrt{\frac{2}{\pi-2}}$. Thus, when either positive or negative residuals dominate the first residuals in the ordering, R_n and H_n may have less power than the earlier tests.

Basic properties of U, Q and T are given in Tables 3.3, 3.4 and 3.5. The cdf and the pdf of each distribution are produced in Figures 3.4, 3.5 and 3.6.

Table 3.3: Basic Properties of U

2.5%	5%	25%	50%	Mean	Var	75%	95%	97.5%
0.031	0.063	0.318	0.674	0.798	0.363	1.150	1.956	2.241

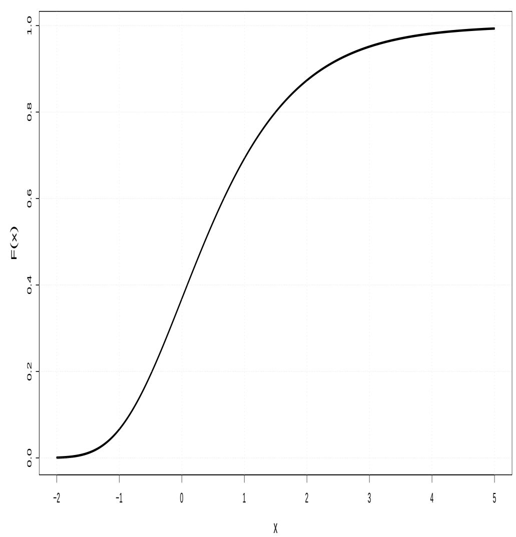


(a) The cdf of U (b) The pdf of U

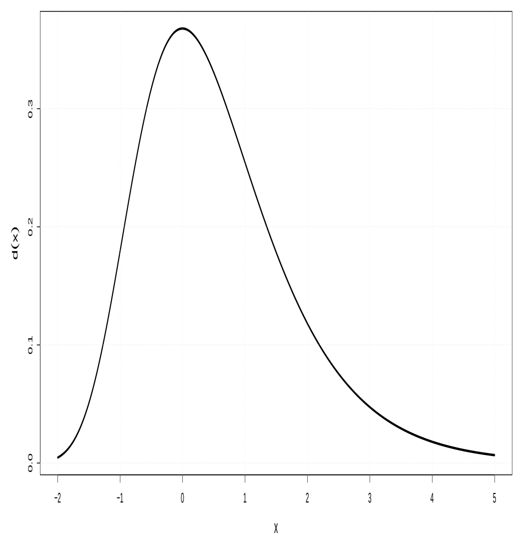
Figure 3.4: The cdf and the pdf of U

Table 3.4: Basic Properties of Q

2.5%	5%	25%	50%	Mean	Var	75%	95%	97.5%
−1.305	−1.097	−0.327	0.367	0.577	1.645	1.246	2.970	3.676



(a) The cdf of Q

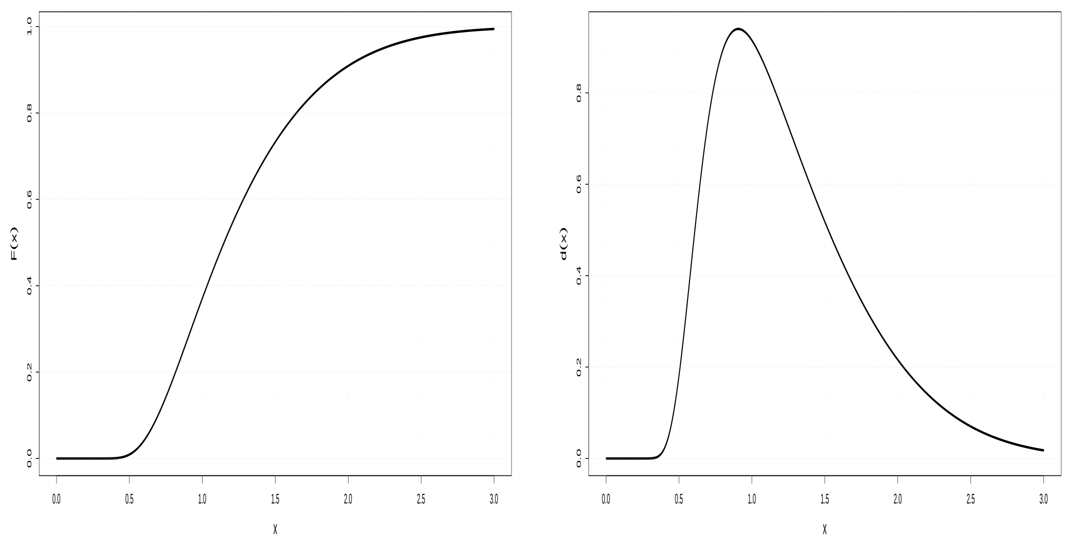


(b) The pdf of Q

Figure 3.5: The cdf and the pdf of Q

Table 3.5: Basic Properties of T

2.5%	5%	25%	50%	Mean	Var	75%	95%	97.5%
0.560	0.618	0.870	1.150	1.253	0.261	1.534	2.241	2.498



(a) The cdf of T (b) The pdf of T

Figure 3.6: The cdf and the pdf of T

3.3 Additional Statistics

If model (1.1) has the tendency to under-estimate or over-estimate the $\tilde{n}$ observations then it is possible to build more powerful tests than T_n and Q_n by using the maximum of the partial sums rather than the maximum of the absolute value of partial sums of the residuals. For example, if the model produces mostly positive residuals that are large in magnitude and negative residuals that are small in magnitude for the low ordered observations then there is no need to take the absolute value of the partial sums as most - if not all - of the high magnitude partial sums have positive signs. Taking this information into account, tests statistics that are based on partial sums rather than the absolute value of partial sums are recommended such as Z_n and G_n presented in Table 3.3. Note that the difference in abilities for detecting lack of fit between Z_n and G_n can be understood in the light of the differences between Q_n and T_n.

Limiting distributions for these statistics are given in Table 3.6. The proofs are given in Appendix A.

Table 3.6: Other Test Statistics

Label	Statistic	95% Percentile	Asym. Dist.
Z_n	$\frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} S_m$	1.959	(3.3)
M_n	$\frac{1}{\sqrt{\tilde{n}}} \left\vert \min_{1 \le m \le \tilde{n}} S_m \right\vert$	1.959	(3.3)
G_n	$a_{\tilde{n}} \max_{1 \le m \le \tilde{n}} \frac{1}{\sqrt{\tilde{m}}} S_m - b_{\tilde{n}} + \log 2$	2.970	(2.2)

It is obvious that $T_n = \max(Z_n, M_n)$. If the model suggests under-estimated predictions or high positive residuals for low ordered observations then T_n and Z_n will have the same observed value but the latter has much lower 95% percentile. It is also noted the 97.5% percentile for Z_n and M_n is 2.24 which is exactly T_n's 95% percentile.

Since their usefulness is restricted to special scenarios just described, these additional statistics are not included in power comparisons and not pursued further.

Chapter 4

Estimating σ^2 and Ordering

4.1 Estimating σ^2

Any estimate of σ that satisfies assumption (b) can be used in all test statistics. In particular, the mean squared error(MSE) of model (1.1) works. The problem is that under an alternative the MSE often tends to get inflated. In turn, test statistics get deflated leading to a power reduction. Instead, following Christensen and Sun (2010), C-S, we use the MSE of a more general model that contains model (1.1). After the data have been ordered let $\boldsymbol{\Gamma_k} = \begin{bmatrix} \boldsymbol{v_1} & \boldsymbol{v_2} & \ldots & \boldsymbol{v_k} \end{bmatrix}$ where

$$\boldsymbol{v}_{2q} = \begin{bmatrix} \cos\left(2\pi q \frac{1}{n}\right) & \cos\left(2\pi q \frac{2}{n}\right) & \ldots & \cos\left(2\pi q \frac{n}{n}\right) \end{bmatrix}^T,$$

$$\boldsymbol{v}_{2q+1} = \begin{bmatrix} \sin\left(2\pi q \frac{1}{n}\right) & \sin\left(2\pi q \frac{2}{n}\right) & \ldots & \sin\left(2\pi q \frac{n}{n}\right) \end{bmatrix}^T,$$

and $\boldsymbol{v}_1 = \begin{bmatrix} 1 & 1 & \ldots & 1 \end{bmatrix}^T$. To avoid redundancy, $\boldsymbol{v}_1$ is dropped if model (1.1) contains the intercept term. We estimate σ^2 by the MSE of the extended model on the ordered data

$$\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{\Gamma}_k\boldsymbol{\gamma}_k + \boldsymbol{e}$$

where $\boldsymbol{\gamma}_k$ is a $k \times 1$ vector of unknown parameters. The estimate of σ^2 is given by

$$\hat{\sigma}_n^2 = \boldsymbol{Y}_n^T(\boldsymbol{I}_n - \boldsymbol{M}_{\boldsymbol{X}_n,\boldsymbol{\Gamma}_k})\boldsymbol{Y}_n \Big/ [n - r(\boldsymbol{X}_n, \boldsymbol{\Gamma}_k)], \tag{4.1}$$

where $\boldsymbol{M}_{\boldsymbol{X}_n,\boldsymbol{\Gamma}_k}$ represents the perpendicular projection operator onto the column space of the matrix $[\boldsymbol{X}_n : \boldsymbol{\Gamma}_k]$ and $r(\cdot)$ is the rank of a matrix. This estimate satisfies assumption (b) if $\frac{k}{n}$ converges to c as $n \to \infty$ where $0 \leq c < 1$ as shown by C-S. In particular any k smaller than $\tilde{n}$ is acceptable. C-S suggested using $k = \lceil n/10(\log\log n)^2 \rceil$. This is the same estimator used by C-L who noted that tests using this estimator often achieve higher power than when using the MSE of model (1.1).

4.2 Ordering

Although the ordering of the data does not affect asymptotic distributions under the null model (1.1), it plays a highly influential role in detecting lack of fit using partial sum of residuals. If the data are poorly ordered, one may not be able to reject a poor model. A good ordering increases power and a bad one decreases it. We would like to use the $\tilde{n}$ observations that are most likely to show lack of fit when it exists.

We follow C-L's suggestion to order the observations according to a modified version of Mahablanobis distance starting from the farthest points. Specifically, the modified squared distance for the ith observation, $\boldsymbol{x}_i$, is $d_i = (\tilde{\boldsymbol{x}}_i - \boldsymbol{\eta})^T\boldsymbol{S}^{-1}(\tilde{\boldsymbol{x}}_i - \boldsymbol{\eta})$ where $\boldsymbol{x}_i^T = \begin{bmatrix} 1 & \tilde{\boldsymbol{x}}_i^T \end{bmatrix}$ and the vector η contains the midrange of each covariate, and $\boldsymbol{S}^{-1}$ is the inverse of the covariance matrix of $\tilde{\boldsymbol{X}}$. Whenever the lack of fit is expected to come from a subset of covariates, the ordering of the data could be restricted to them, with η and S defined as the midranges and the covariance matrix of only the suspected covariates. C-L found this ordering method is preferable to Mahablanobis distance when the data come from skewed or irregular distributions.

C-L did not suggest, nor do we, that the aforementioned ordering method is perfect. The data can be ordered in any way, as long as it does not depend on $\boldsymbol{Y}_n$. C-L found their method to be effective and so have we. One might need to implement several ordering methods to reveal lack of fit. For example, C-L's method can be reversed so the data are ordered from points nearest to the center to farthest. Mahablanobis distance, the standard Euclidean distance from the center of the data or even choosing observations randomly might be adopted. This flexibility in choosing the ordering method stems from the fact that the null model asymptotic results do not depend on the particular ordering. In situations where the lack of fit is suspected to come from one predictor only, one might merely order the data according to that variable ascending or descending since occasionally the lack of fit increases as the predictor increases or decreases.

There are no fool-proof methods for detecting lack of fit. There is no way to know the structure of the lack of fit. And for any method of detecting lack of fit, one can define a lack of fit that the method will miss. Even Fisher's famous lack of fit test based on exact replicates will miss any lack of fit that exists within the replicates e.g. a time trend within the replicates.

In a nutshell, there is no perfect ordering method for all possible scenarios. What works well in one situation might fail utterly in another. Again, the ordering method must be chosen without reference to the fit of the model.

Chapter 5

Monte Carlo Simulations

A large value for one of the proposed statistics provides evidence for lack of fit. The strength of the evidence is assessed by the P value. For example, the P value associated with an observed value of V_n is $P = Pr[V_n \geq v_n]$. This quantity can be approximated by $Pr[V \geq v_n]$ where V is the limiting distribution of V_n. The quality of this approximation depends greatly on the sample size n.

Unfortunately, partial sum statistics with estimated parameters converge very slowly to their asymptotic distributions. In the case of small to moderate sample sizes, the distributions of partial sum statistics do not much resemble their limiting distributions which leads to imprecise approximation of the P value. We suggest using Monte Carol simulations to approximate P values instead. We have found that Monte Carlo simulations lead to more accurate approximations of P values. For more details on Monte Carlo computation of P values see Hart (1997) and MacKinnon (2002).

Assume the data in model (1.1) are ordered. To approximate the P value using simulation, of say T_n, for the data in hand $\boldsymbol{Y_n}$ and $\boldsymbol{X_n}$:

1. Fit model (1.1) to the ordered data. Calculate T_n and call it T_{obs}.

2. Simulate data

$$\boldsymbol{Y}^* = \boldsymbol{X_n}\boldsymbol{\beta} + \boldsymbol{e}^*, \quad \boldsymbol{e}^* \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

   where $\boldsymbol{I}_n$ is the identity matrix of size n. We refer to this as the assumed data distribution. The choice of $\boldsymbol{\beta}$ and σ^2 is irrelevant as explained below, therefore, $\boldsymbol{\beta}$ and σ are chosen as the vector $\mathbf{0}$ and 1 respectively.

3. Regress $\boldsymbol{Y}^*$ on $\boldsymbol{X_n}$. Find the residuals.

4. Using the residuals in 3, compute T_n and call it t_1.

5. Repeat steps 2-4 B times to obtain: $t_1, t_2 \ldots t_B$. We now have an empirical distribution of T_n. For accurate results, we take $B = 19999$. For practical computations, $B = 2999$ works well.

6. The P-value is approximated as the proportion of times that t_j is greater than T_{obs}.

The procedure was described assuming the error vector follows a normal distribution. While this is a standard assumption in linear model, the simulation only requires that the components of the error vector $\boldsymbol{e_n}$ be iid with mean 0 and some unknown scale parameter; MacKinnon (2002). Simulation from the normal is not necessary except for test statistics that require normality. These tests are R_n and H_n which are based on absolute residuals. Assuming a certain distribution is necessary to standardize the absolute residuals correctly. If a non-normal distribution is assumed such as t distribution, then they should be standardized accordingly yielding different statistics from R_n and H_n. They may even fail to converge if the correct standardization is not taken into the account. Alternatively, Hart and Mackinnon recommend re-sampling (bootstrapping) from the residuals of the fitted model if one is unwilling to make certain distributional assumptions about the error distribution.

To test the sensitivity of partial sum statistics to the normality assumption for the data, two simulation studies are performed. The first study uses one predictor, $p = 2$ with the intercept, and the second study uses five predictors, $p = 6$. For each statistic three data distributions are used: the standard normal, a t distribution with 6 degrees of freedom (heavy tails) and Uniform$[-1,1]$ (short tails). For each statistic, comparisons are made on the 95% percentile of the three empirical distributions across n. Little or no difference between the percentiles indicates robustness to distributional assumptions. Results are also shown for R_n and H_n without changing the standardization implied by the assumption of normality. The results are shown in Figures 5.1 through 5.12 along with a horizontal line that indicates the percentile of the asymptotic distribution.

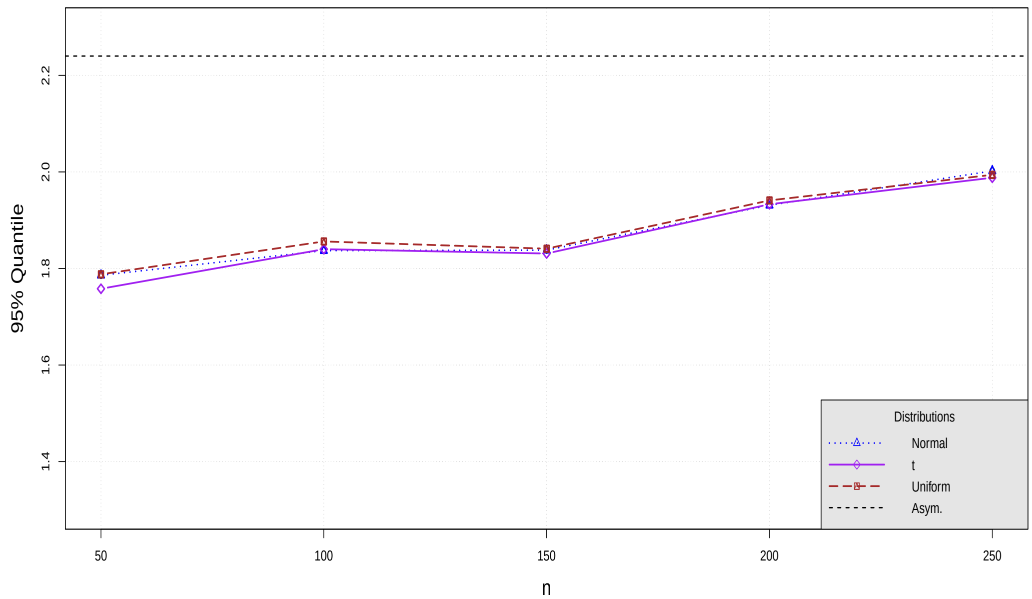


Figure 5.1: 95% Quantile Comparison for T_n; $p = 2$

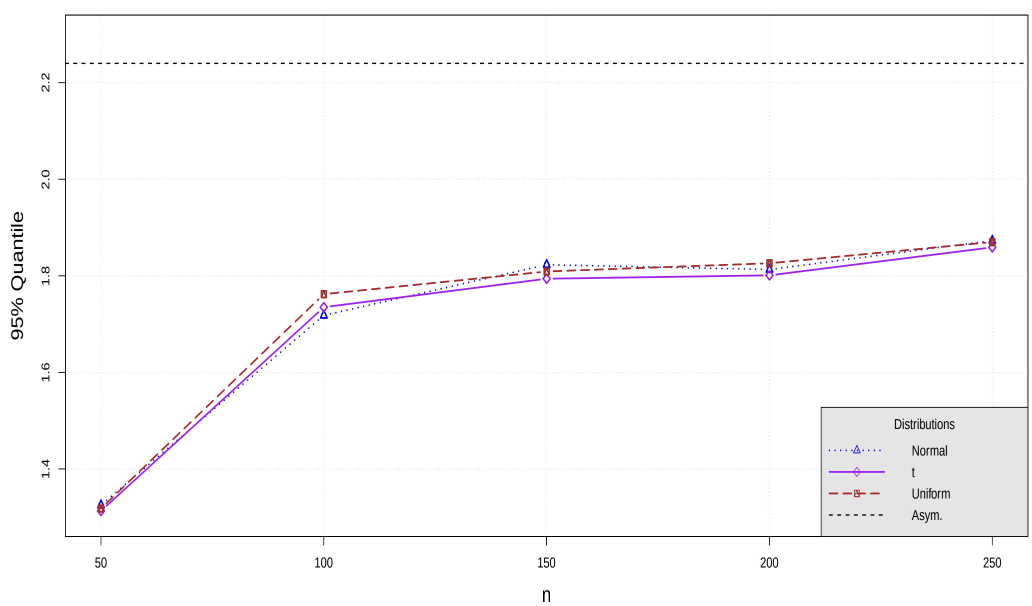


Figure 5.2: 95% Quantile Comparison for T_n; $p = 6$

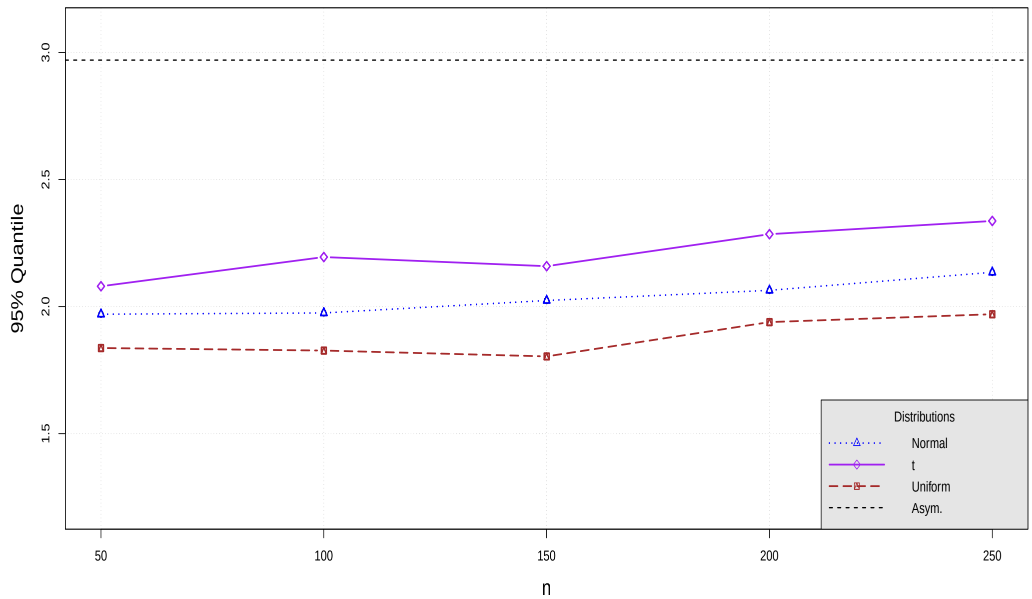


Figure 5.3: 95% Quantile Comparison for Q_n; $p = 2$

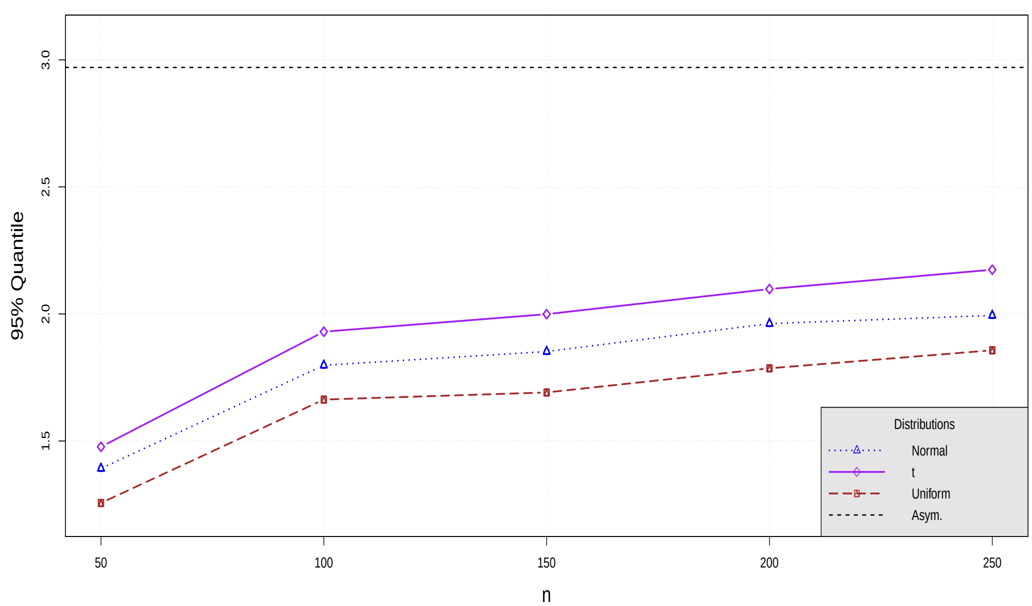


Figure 5.4: 95% Quantile Comparison for Q_n; $p = 6$

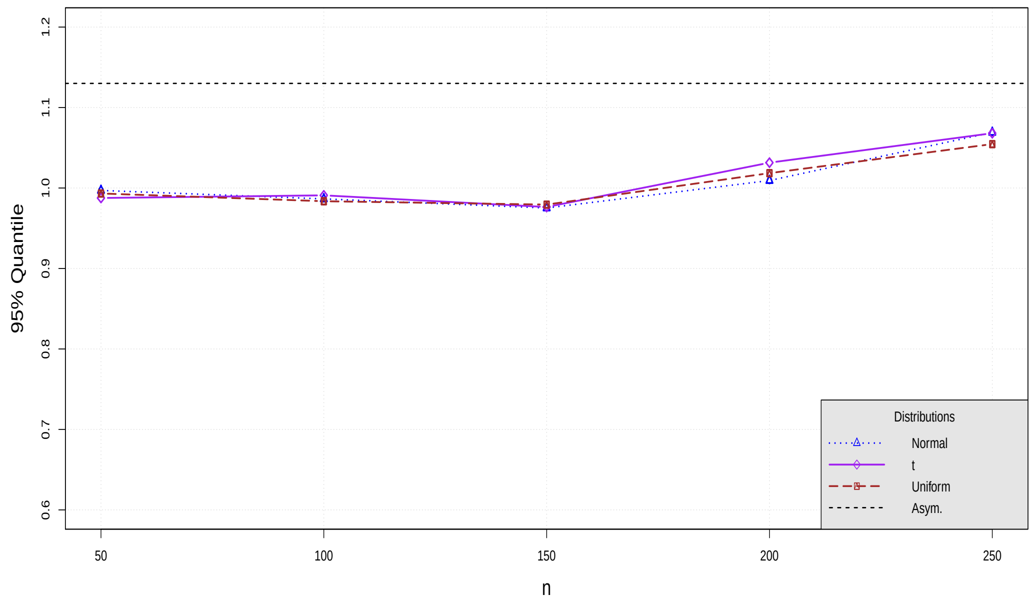


Figure 5.5: 95% Quantile Comparison for W_n; $p = 2$

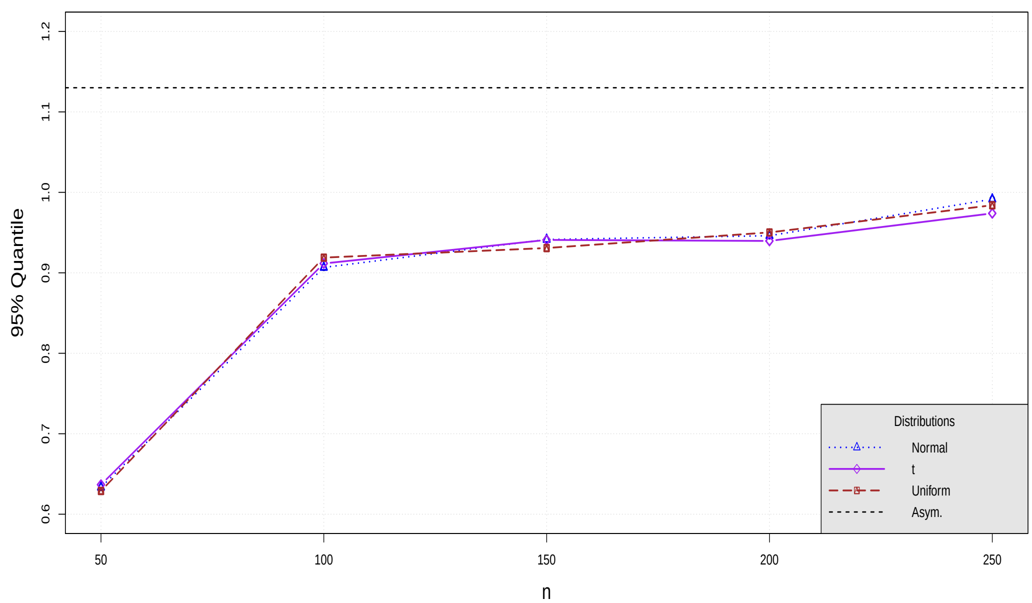


Figure 5.6: 95% Quantile Comparison for W_n; $p = 6$

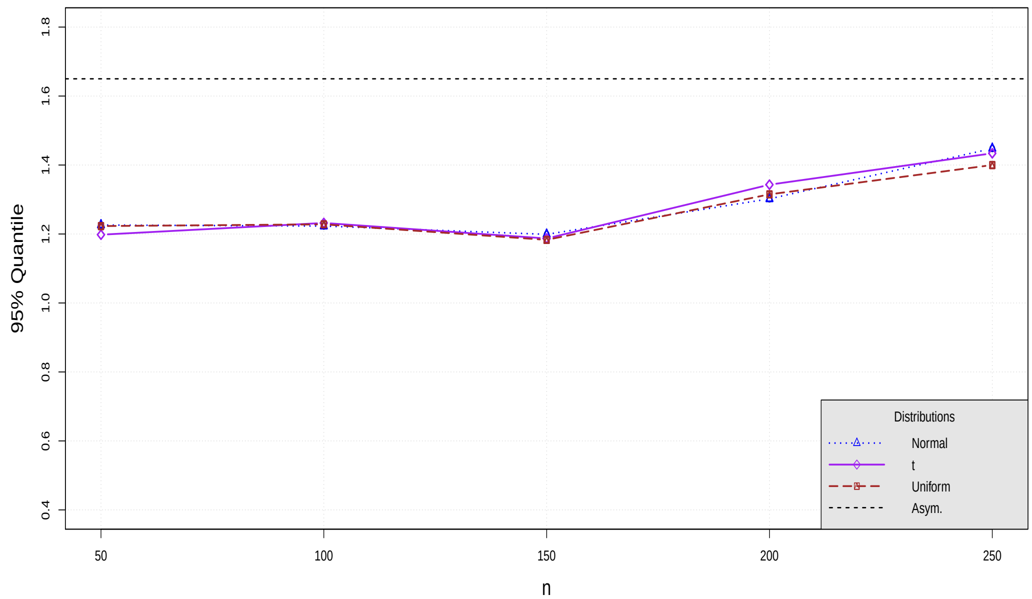


Figure 5.7: 95% Quantile Comparison for V_n; $p = 2$

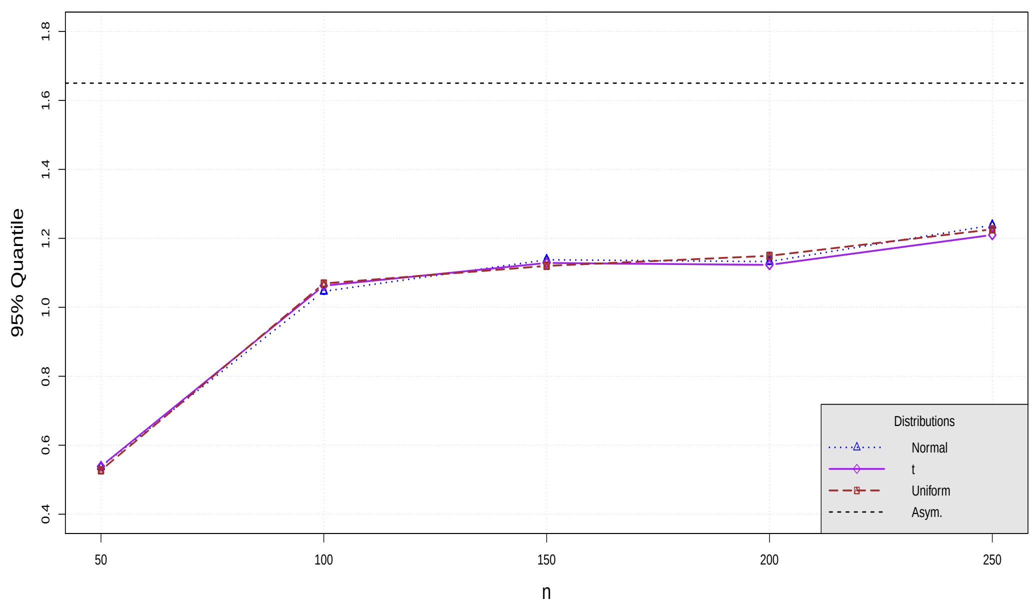


Figure 5.8: 95% Quantile Comparison for V_n; $p = 6$

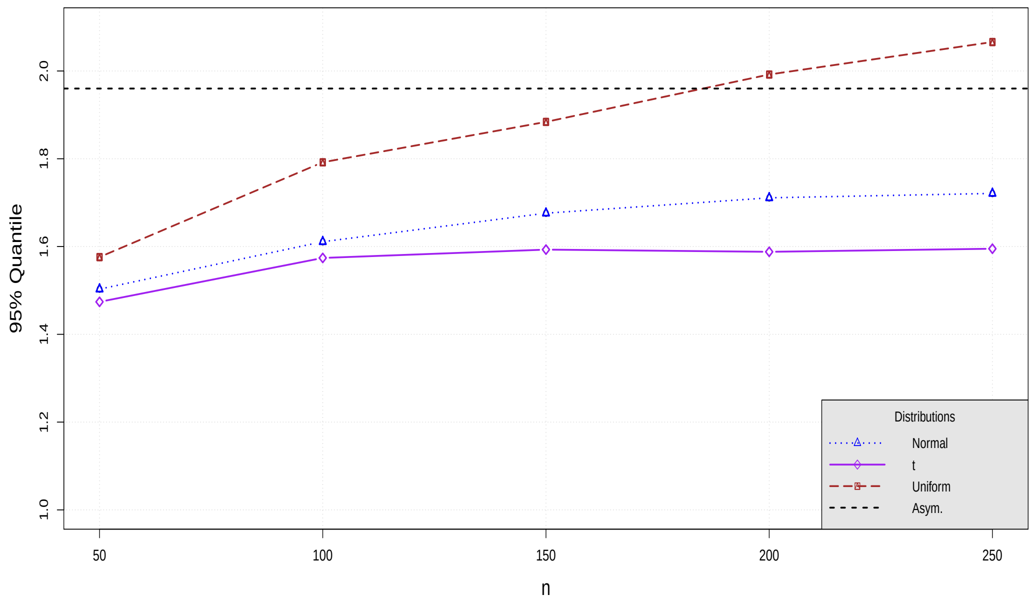


Figure 5.9: 95% Quantile Comparison for R_n; $p = 2$

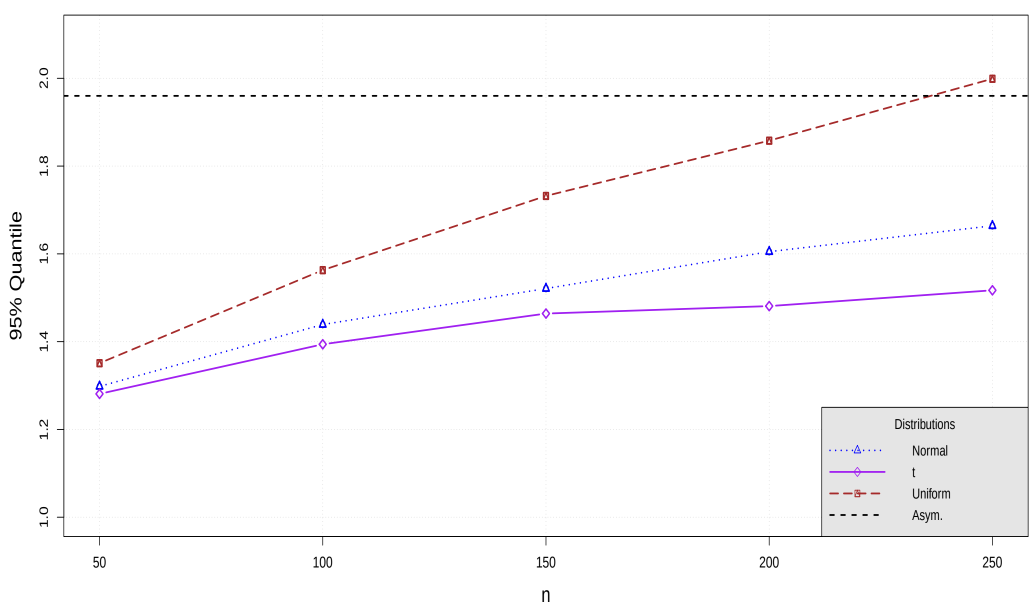


Figure 5.10: 95% Quantile Comparison for R_n; $p = 6$

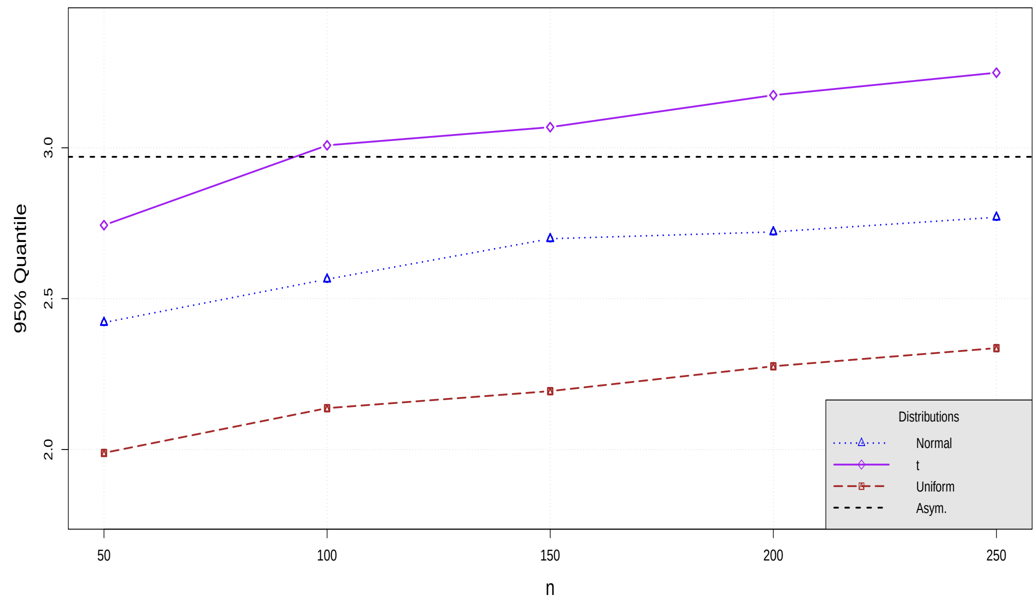


Figure 5.11: 95% Quantile Comparison for H_n; $p = 2$

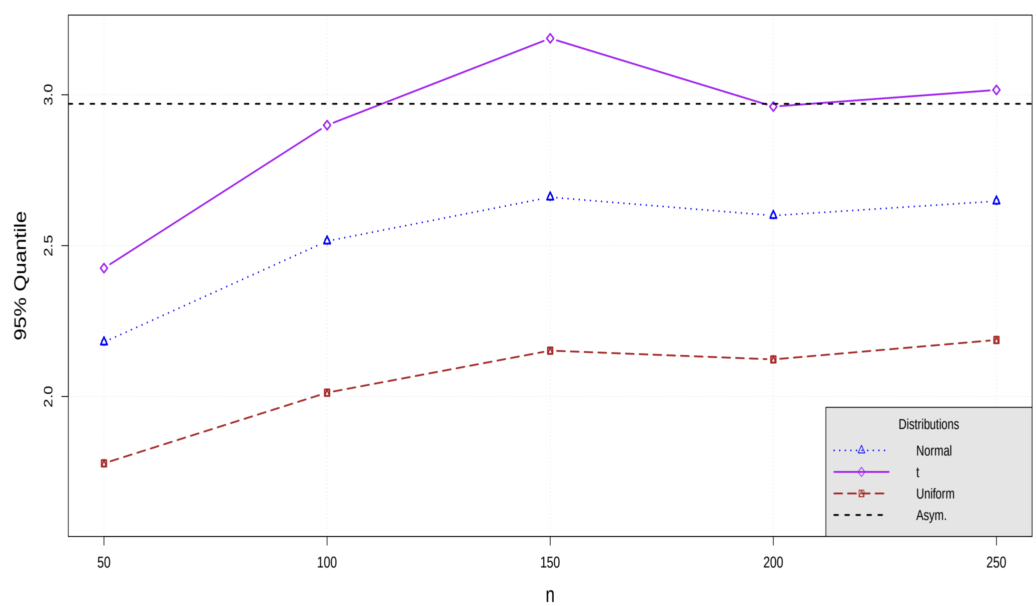


Figure 5.12: 95% Quantile Comparison for H_n; $p = 6$

First, with few exceptions, and none involving normal data, when using the asymptotic quantiles the tests are less likely to reject the null hypothesis. The sizes are lower than the nominal level showing the need to adopt simulations for P value computations rather than depending on the asymptotic percentiles.

It is worth noting the robustness of T_n, W_n and V_n. Their percentiles are little affected by changing the distribution of the error vector. The differences between the three percentiles are indistinguishable regardless of the sample size, n. Whereas Q_n is moderately affected if the distribution of the error vector is altered producing three different percentiles although its asymptotic distribution does not require normality.

Figures 13 through 18 allow us to evaluate the difference in rejection regions that assume normality when the correct distribution is not. Figures 13 and 14 display that the size of the Q_n-test does not much differ from the nominal level if normality is assumed mistakenly. The sizes are constantly higher than 0.05 rising up to 0.068 if the parent distribution is t and constantly lower than 0.05 falling down to 0.033 if the parent distribution is uniform. As expected, R_n and H_n are severely affected if the distribution of the data is not normal. It is less affected if the distribution is t as compared to uniform. We expect that they both - R_n and H_n - become less affected as the degrees of freedom of t increases.

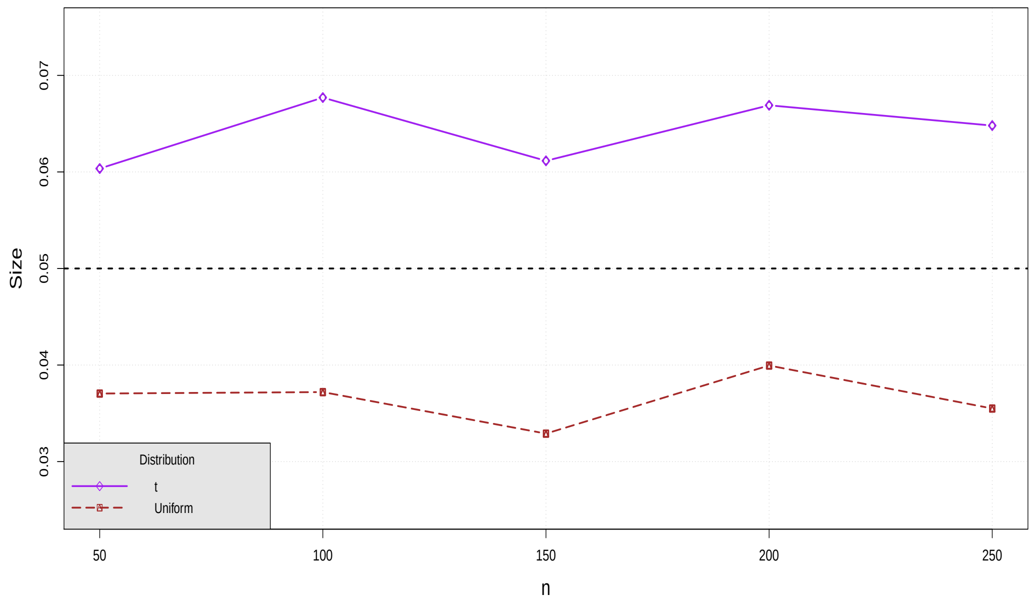


Figure 5.13: Size comparison when rejecting based on 0.05 level from simulated normals for Q_n; $p = 2$

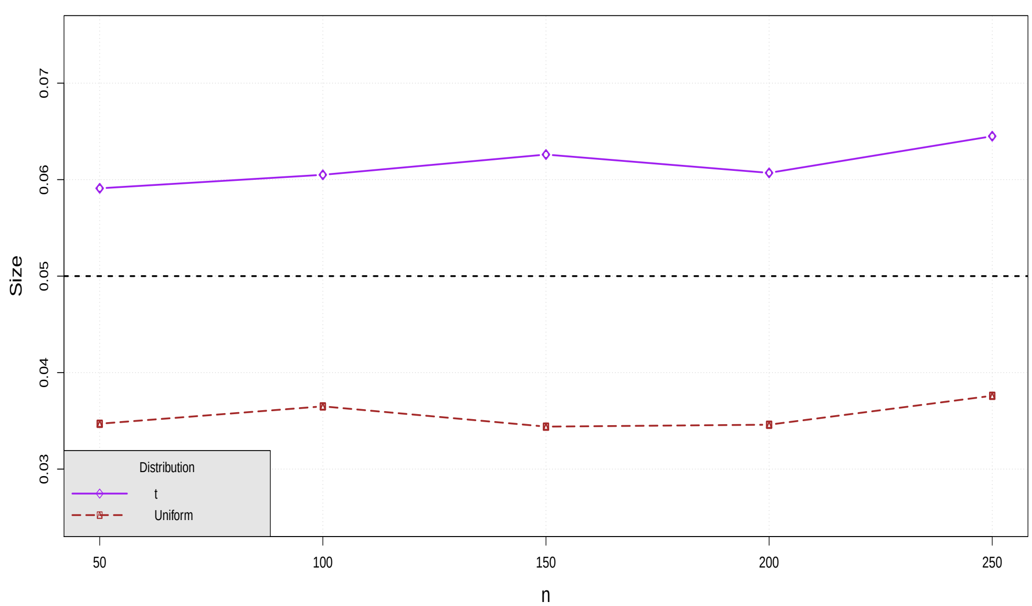


Figure 5.14: Size comparison when rejecting based on 0.05 level from simulated normals for Q_n; $p = 6$

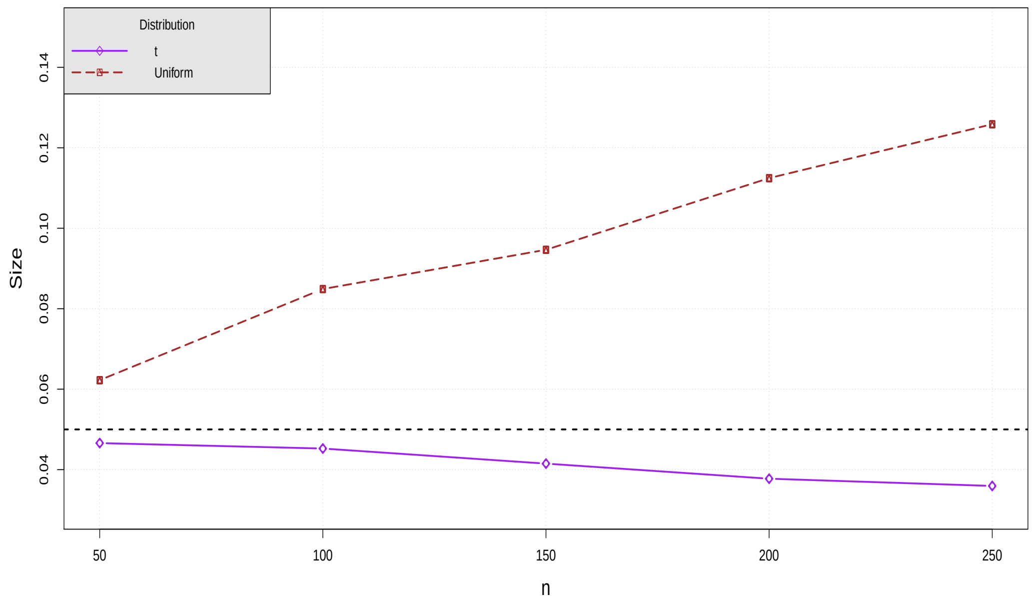


Figure 5.15: Size comparison when rejecting based on 0.05 level from simulated normals for R_n; $p = 2$

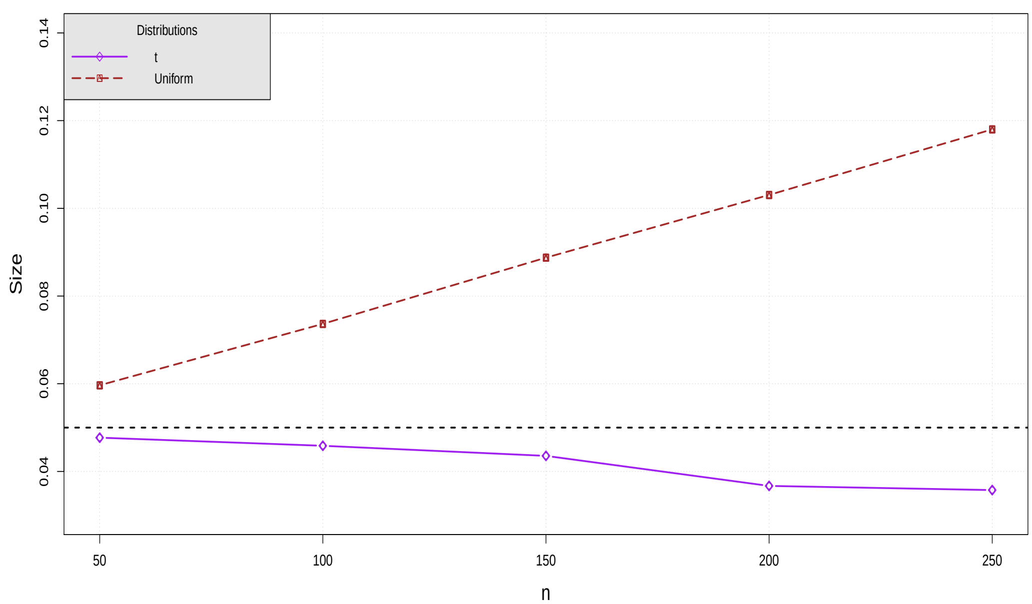


Figure 5.16: Size comparison when rejecting based on 0.05 level from simulated normals for R_n; $p = 6$

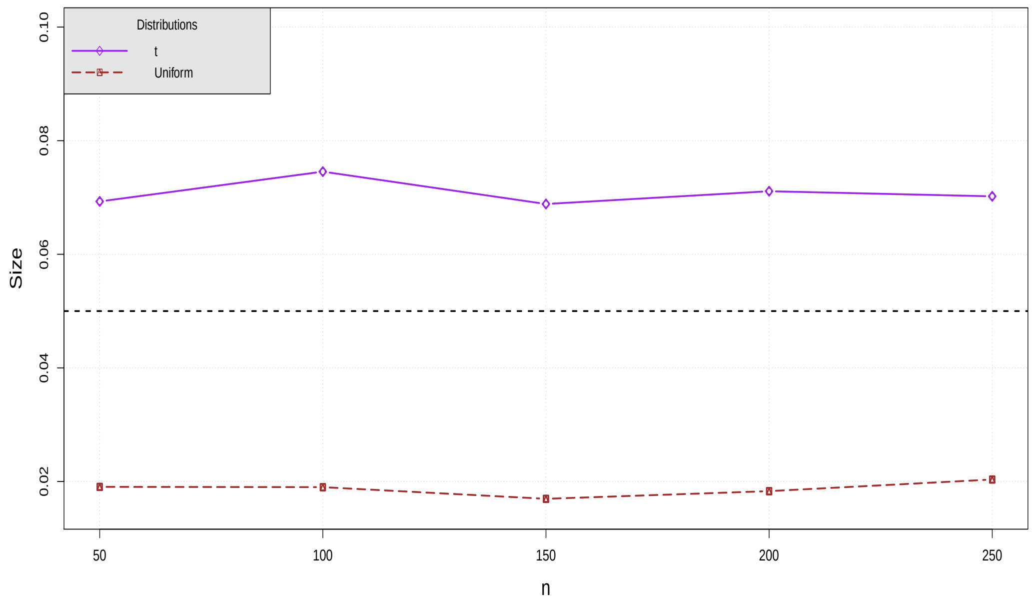


Figure 5.17: Size comparison when rejecting based on 0.05 level from simulated normals for H_n; $p = 2$

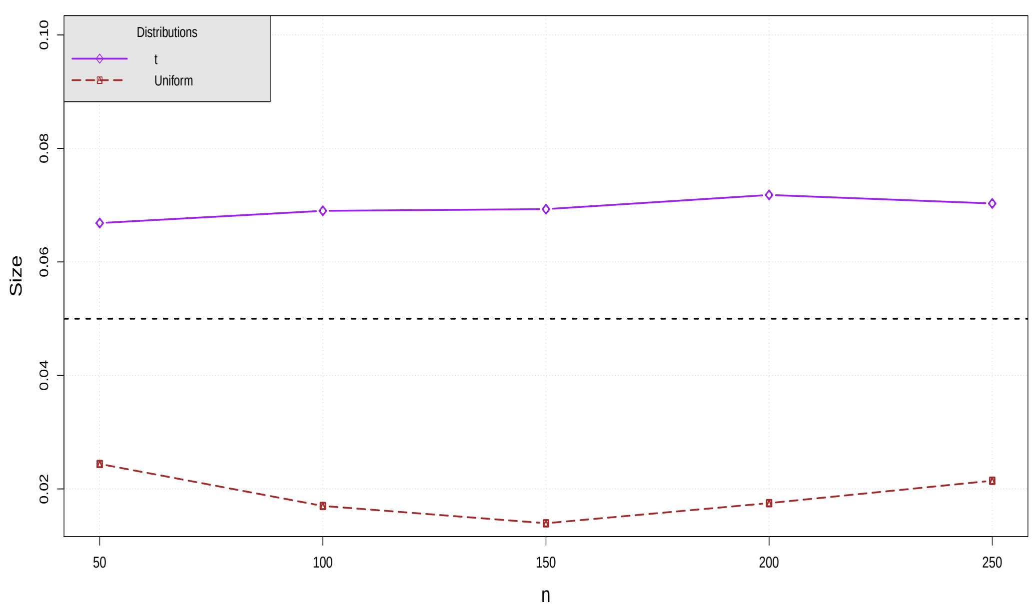


Figure 5.18: Size comparison when rejecting based on 0.05 level from simulated normals for H_n; $p = 6$

To explain the irrelevancy of $\boldsymbol{\beta}$ and σ in the simulations, it is clear that the test statistics depend solely on the residuals of the fitted model. The residuals are $(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{Y_n} = (\boldsymbol{I}-\boldsymbol{M})\boldsymbol{e_n}$, where $\boldsymbol{M}$ is the perpendicular projection operator onto the column space of $\boldsymbol{X_n}$. It is obvious that the residuals entirely depend on the design matrix $\boldsymbol{X_n}$ and the error vector $\boldsymbol{e_n}$. Thus, we effectively only need to simulate $\boldsymbol{e}^*$ so without loss of generality we can take $\boldsymbol{\beta} = \boldsymbol{0}$. Also, σ is irrelevant as long as the components of $\boldsymbol{e_n}$ follow a distribution with CDF $F\left(\frac{e}{\phi}\right)$ where ϕ is a scale parameter and the statistic $\hat{\sigma}_n^2$ is proportional to a quadratic form in $\boldsymbol{e_n}$, $\boldsymbol{e_n^T B e_n}$, where $\boldsymbol{B}$ is a non-zero non-negative definite matrix. For the MSE of model (1.1), $\boldsymbol{B} = (\boldsymbol{I}-\boldsymbol{M})$. For the estimator in (4.1), $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}_n,\boldsymbol{\Gamma}_k}$. ϕ and σ coincide if the distribution is normal. In general, σ is proportional to ϕ if the second moment exists. Then the components of $\boldsymbol{r_n} = \frac{\boldsymbol{e_n}}{\phi}$ follow the distribution $F(\cdot)$ with scale parameter equal to 1. The ith residual is $\hat{e}_i = \boldsymbol{a}_i^T \boldsymbol{e_n}$, where $\boldsymbol{a}_i^T$ is the ith row vector of the matrix $(\boldsymbol{I}-\boldsymbol{M})$. Then the partial sum

$$\begin{aligned}
\sum_{i=1}^{m} \frac{\hat{e}_i}{\hat{\sigma}_n} &= \frac{1}{\hat{\sigma}_n} \sum_{i=1}^{m} \boldsymbol{a}_i^T \boldsymbol{e_n} \\
&= \frac{c\phi}{\sqrt{\boldsymbol{e_n^T B e_n}}} \sum_{i=1}^{m} \boldsymbol{a}_i^T \boldsymbol{r_n} \\
&= \frac{c}{\sqrt{\boldsymbol{r_n^T B r_n}}} \sum_{i=1}^{m} \boldsymbol{a}_i^T \boldsymbol{r_n}
\end{aligned}$$

where c is a known proportionality constant. Clearly the partial sum depends only on $\boldsymbol{r}_n$ that has a parameter free distribution hence the choice of σ does not matter. In all simulations, the convenient value $\sigma = 1$ was used.

Finally, the partial sums are built on $\tilde{n}$ residuals rather than n so that the asymptotic results work. Recall that $\tilde{n} = \lceil n/(\log\log n)^{1+\delta} \rceil$ where $\delta > 1$ for Q_n and H_n and greater than 0 for the rest of the statistics. As δ increases, $\tilde{n}$ decreases. The choice of δ affects the convergence rate of the test statistics. The higher δ gets, the slower is the convergence rate and finite sample test statistics less resemble their limiting distributions. As δ gets lowered, $\tilde{n}$ gets larger. Recall that T_n, W_n, V_n and R_n are divided by an increasing function of $\tilde{n}$, $\sqrt{\tilde{n}}$. Whereas Q_n and H_n are subtracted by an increasing function of $\tilde{n}$, $b_{\tilde{n}}$. This leads to reduction in the test statistics if $\tilde{n}$ is excessively large hence affecting the power. Large values of δ might exclude observations that may be needed to reveal lack of fit. We depend on simulations results to decide on an appropriate value of δ. After extensive simulations among $\delta \in \{0.5, 1, 1.5, 2, 2.5, 3, 4, 5\}$ we found that $\delta = 2$ seems a reasonable choice in terms of power. This coincides with C-L's suggestion for their test statistics T_n and Q_n.

Chapter 6

Power Comparisons

In this chapter we consider power comparisons between the test statistics using simulations. In each example, the departure from the linearity assumption ranges from none to severe. Type I error, or the size of the tests, is set equal to 0.05.

First, a set of covariates is generated once before forming a design matrix, $\boldsymbol{X_n}$. Then 95% quantiles of each test statistic are computed using $\boldsymbol{X_n}$ and $B = 19,999$ and assuming normality unless stated otherwise. 10,000 response vectors, $\boldsymbol{y}_1 \ldots \boldsymbol{y}_{10,000}$, are simulated according to some relationship with the covariates. Each $\boldsymbol{y}_i$ is linearly regressed on $\boldsymbol{X_n}$ before computing the test statistics. A lack of fit is declared and the model is rejected according to a particular test statistic if its value exceeds its previously computed empirical 95% quantile. The empirical power is defined as the rejection rate. If the model is correctly specified, the empirical power for each test must be close to 0.05 or to reject about 500 times out of 10,000. The power at the null model should differ from 0.05 only by the sampling error in the two simulations. We typically expect that the empirical power increases as the departure from the model increases.

Simple linear regression, SLR, is discussed first followed by multiple linear regression, MLR.

6.1 SLR

For SLR, the fitted relationship between the response variable y and x is

$$y = \beta_0 + \beta_1 x + \epsilon$$

In examples below, a lack of fit takes the form of a nonlinear function of x, controlled by a constant θ, added to this relationship. The constant θ characterizes the amount of miss-specification in the model.

EXAMPLE 6.1.1: The covariate x is sampled from $N(0,1)$ with $n = 70$, and the response y is drawn independently from

$$y = 1 + 2x + \theta x^2 + \epsilon, \quad \epsilon \sim N(0, 2.5^2)$$

θ ranges between 0 and 1. It is assumed that $E(y) = \beta_0 + \beta_1 x$. The model is correctly specified at $\theta = 0$. Figure 6.1 displays the power performance for each test.

First note that the power at $\theta = 0$ is about 0.05 corresponding to Type I error or tests size. It is evident that the tests based on partial sums of residuals outperform the tests based on absolute residuals (R_n and H_n). The power of H_n did not exceed 0.5 until $\theta = 1$. R_n was only capable of 0.48. Whereas W_n and V_n exceeded power 0.5 at $\theta = 0.6$ and achieved a maximum power of 0.94. The inferiority of R_n and H_n is due to the fact that first $\tilde{n}$ residuals are largely dominated by positive residuals. Taking the absolute value of the residuals is not crucial here when forming the partial sums. W_n is the most powerful test followed closely and almost indistinguishably by V_n. We will continue seeing this behavior for most of the examples. At $\theta = 0.5$, W_n is 17%, 27%, 245% and 165% more powerful than T_n, Q_n, R_n and H_n respectively.

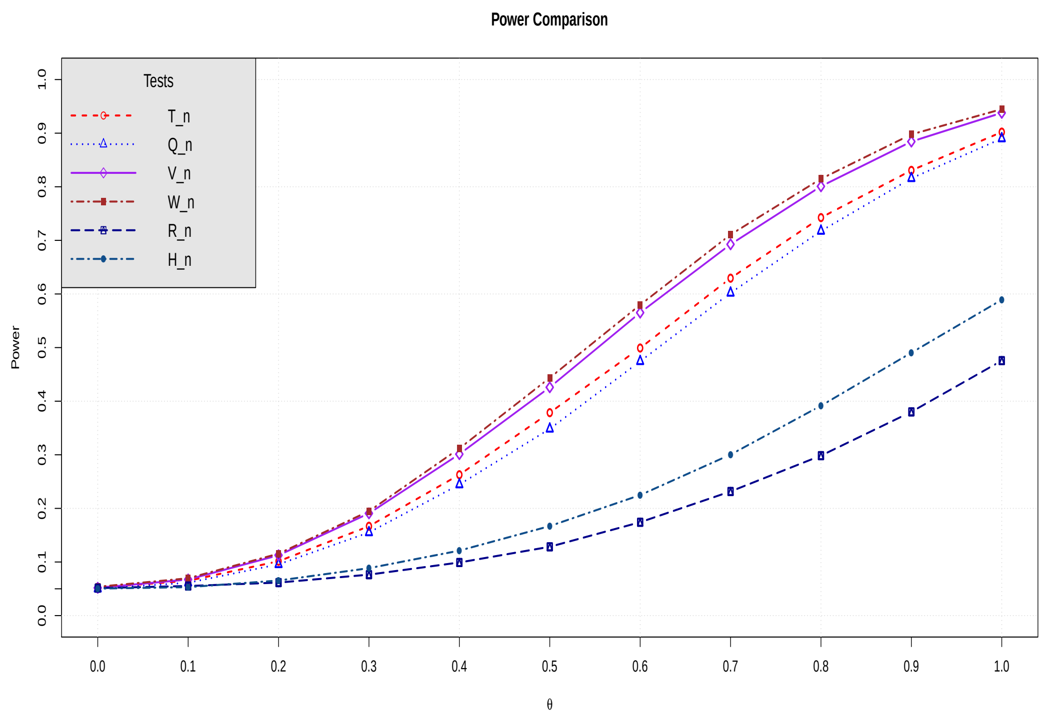


Figure 6.1: Power Comparison - Example 6.1.1; $y = 1 + 2x + \theta x^2 + \epsilon$

In general, W_n and V_n seem to be more sensitive than Q_n and T_n when the lack fit is distributed randomly over the first ordered residuals. Q_n is superior when the lack of fit occurs in low ordered residuals whereas T_n works better when the lack of fit takes place over relatively higher ordered residuals. The difference between R_n and H_n can be understood in terms of the difference between T_n and Q_n but for absolute residuals.

Example 6.1.2: x is sampled from $U(1, 4)$ with $n = 70$ and the response y follows the relationship

$$y = 3x^{-\theta} + \epsilon, \quad \epsilon \sim N(0, 0.5^2)$$

θ ranges between 0 and 1.5. The results are shown in Figure 6.2.

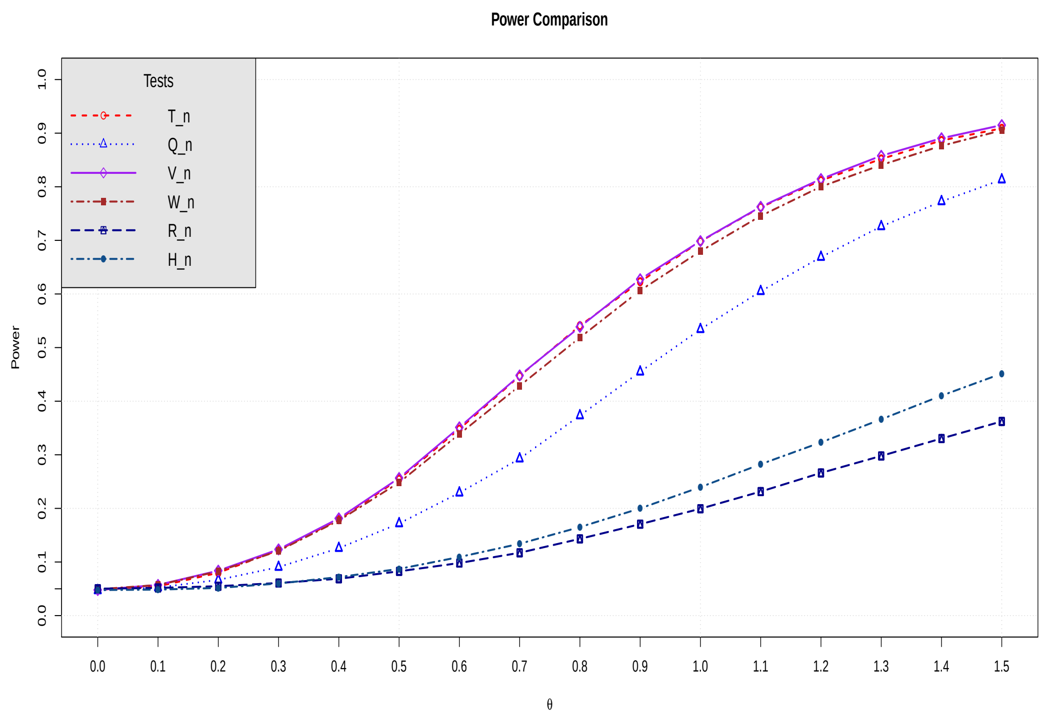


Figure 6.2: Power Comparison - Example 6.1.2; $y = 3x^{-\theta} + \epsilon$

There are some similar patterns to the previous example. T_n has improved and Q_n has declined. The tests based on partial sums of residuals surpass absolute residuals based tests. The maximum power of the latter tests is less than 0.5 as a result of dominating positive residuals in the low ordered residuals. The performances of W_n, V_n and T_n are almost identical. W_n is 33%, 200% and 250% more powerful than Q_n, H_n and R_n respectively when $\theta = 0.9$.
Here the lack of fit is not concentrated in the first few ordered residuals explaining the superiority of W_n over Q_n.

Example 6.1.3: x is sampled from $U(-1, 1)$ with $n = 70$ and the response y follows the relationship

$$y = 1 + 2x + \theta x^3 + \epsilon, \quad \epsilon \sim N(0, 0.1^2)$$

θ ranges between 0 and 3. The results are presented in Figure 6.3.

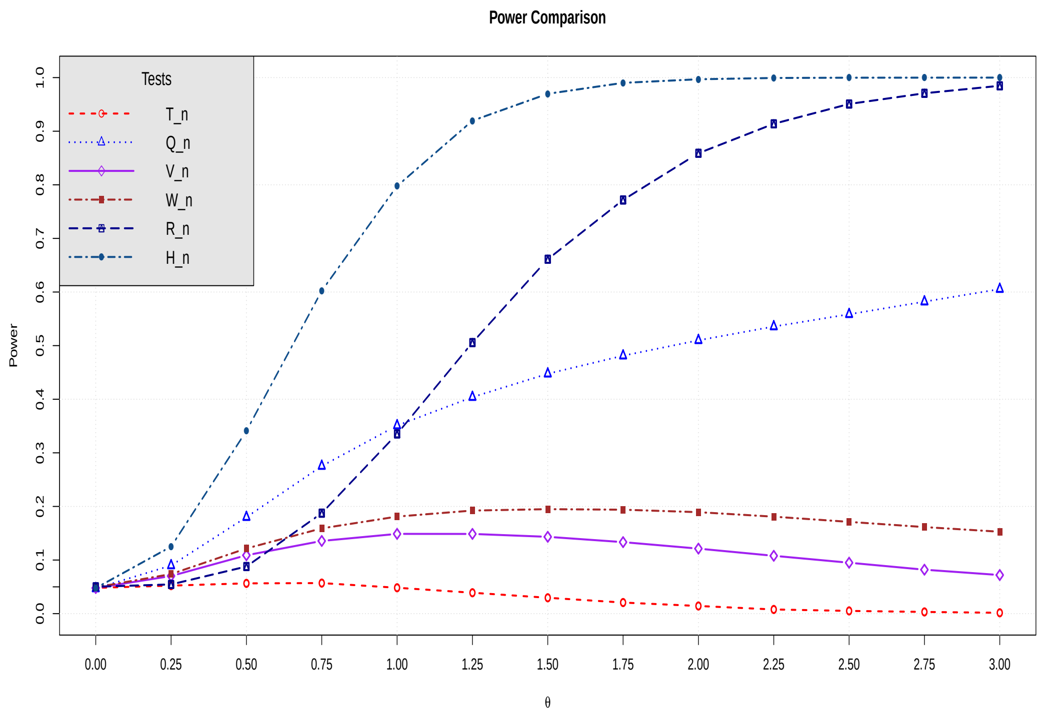


Figure 6.3: Power Comparison - Example 6.1.3; $y = 1 + 2x + \theta x^3 + \epsilon$

Notice the high quality power performance of H_n. It reaches power of 0.8 at $\theta = 1$. In contrast, the partial sums of residuals based tests suffer from low power except perhaps Q_n. They did not exceed 0.2. In fact, their power slightly decreases as θ increases.

The superiority of H_n over R_n is due to the fact that the lack of fit exists in the first few ordered residuals. For the same reason, Q_n has relatively good performance over T_n, W_n and V_n. The residuals have similar magnitudes but different signs. Thus, they cancel each other when taking the partial sum leading to small statistics for tests based on signed residuals. It is noted that Q_n has higher power than R_n for $\theta \leq 1$. H_n is 400%, 116% and 47% more powerful than W_n, Q_n and R_n respectively at $\theta = 1.5$.

The example is repeated assuming now $\epsilon \sim U(-0.1\sqrt{3}, 0.1\sqrt{3})$. The mean and the variance of this distribution match those for $N(0, 0.1^2)$. The $B = 19,999$ simulations draws are based on $U(-1,1)$. The tests based on absolute residuals are adjusted to accommodate for the mean and the variance of the absolute value of $U(-1,1)$. The results are shown in Figure 6.4.

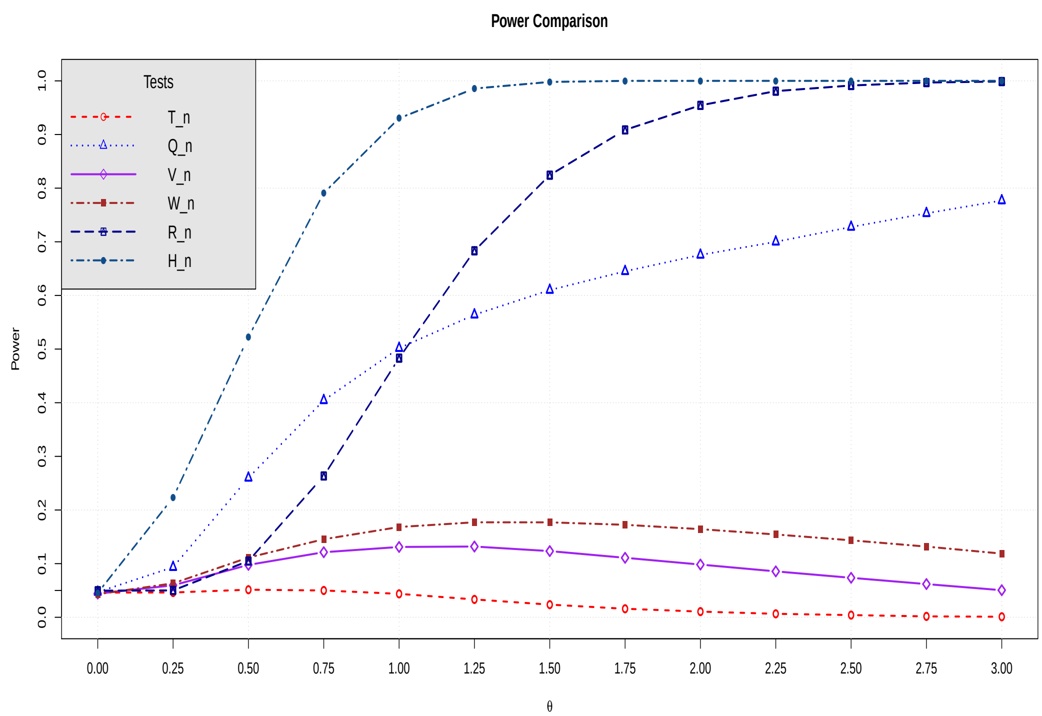


Figure 6.4: Power Comparison - Example 6.1.3; $\epsilon \sim U(-0.1\sqrt{3}, 0.1\sqrt{3})$

Clearly, the pattern has not changed. H_n is still in the lead and R_n dominates

Q_n for $\theta > 1$. As expected, the power of T_n, W_n and V_n has nearly stayed the same. These statistics have shown in Chapter (5) robustness against distributional assumption. Notice also the increase in power for the rest of statistics. For H_n the power has risen from 0.60 to 0.79 at $\theta = 0.75$. Similarly, the power of R_n and Q_n have increased about 0.17 at $\theta = 1.5$.

Again, the example is repeated but now we assume that ϵ follows a t distribution with 6 degrees of freedom and scale parameter equals $0.1\sqrt{\frac{4}{6}}$. The scale parameter is chosen so that the errors have the same mean and variance for $N(0, 0.1^2)$. The $B = 19,999$ simulations draws are based on $t(6)$. The tests based on absolute residuals are adjusted according to the mean and the variance of the absolute value of $t(6)$. The results are shown in Figure 6.5.

The pattern is not different but the distribution effect has become more apparent. The power of H_n has decreased by 0.13 and 0.32 at $\theta = 0.75$ when compared to normal and uniform respectively. Similarly, the power of R_n has dropped by 0.15 and 0.31 at $\theta = 1.5$. The maximum power of Q_n is 0.38. It achieved maximums of 0.60 and 0.77 when assuming normal and uniform, respectively. This illustrates the importance of the distributional assumption to these statistics. However, H_n still has showed powerful performance regardless the distribution of data. For the rest of the statistics, the power practically has not changed.

6.2 MLR

For MLR, the fitted relationship between the response variable y and the covariates x_1 and x_2 is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

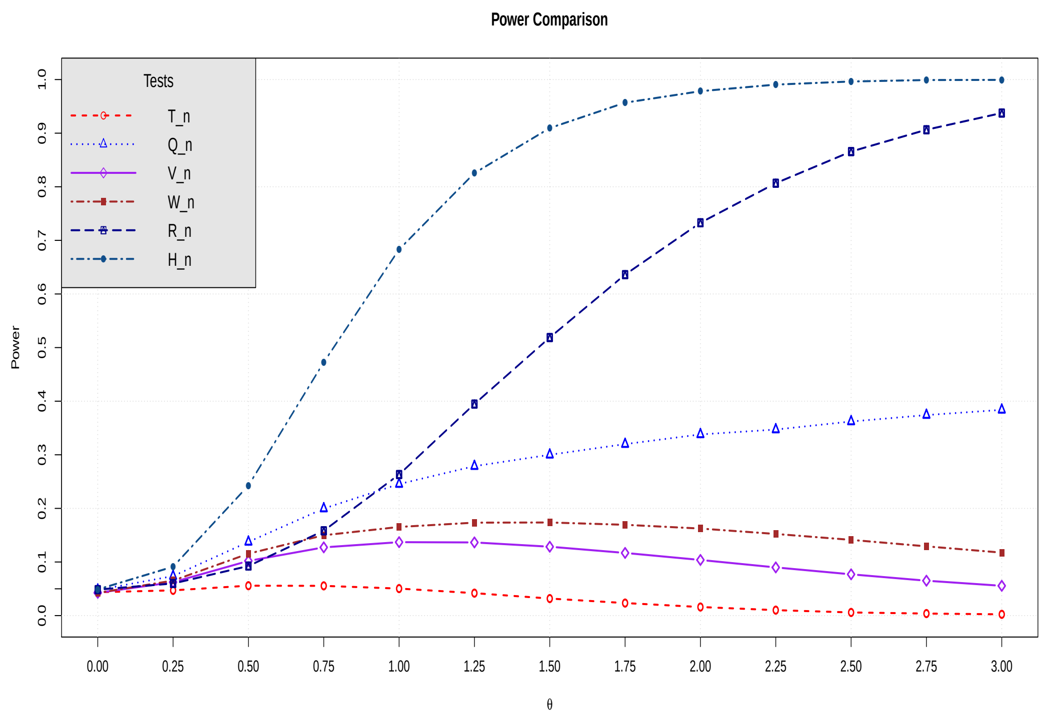


Figure 6.5: Power Comparison - Example 6.1.3; $\epsilon \sim t(6, 0.1\sqrt{\frac{4}{6}})$

Similar to SLR, a function of x_1 and x_2 controlled by a constant θ is added to this relationship. The covariates x_1 and x_2 are simulated once. Then 95% quantiles of each test statistic are computed using $B = 19,999$. At each value of θ, y is simulated 10,000 times. The empirical power for each test statistic is calculated at each θ.

Example 6.2.1: The effect of excluding the interaction term between two covariates is first assessed. x_1 and x_2 with $n = 70$ are sampled independently from $N(0,1)$ and the response y is drawn from

$$y = 2 + x_1 + x_2 + \theta x_1 x_2 + \epsilon, \quad \epsilon \sim N(0, 1.5^2)$$

θ ranges between 0 and 2. The fitted model assumes $E(y) = \beta_0 + \beta_1 x + \beta_2 x$. It does not include the interaction term. It assumes that the mean function is additive. The model is correctly specified when $\theta = 0$. The ordering is imposed using x_1 and x_2. The results are provided in Figure 6.6.

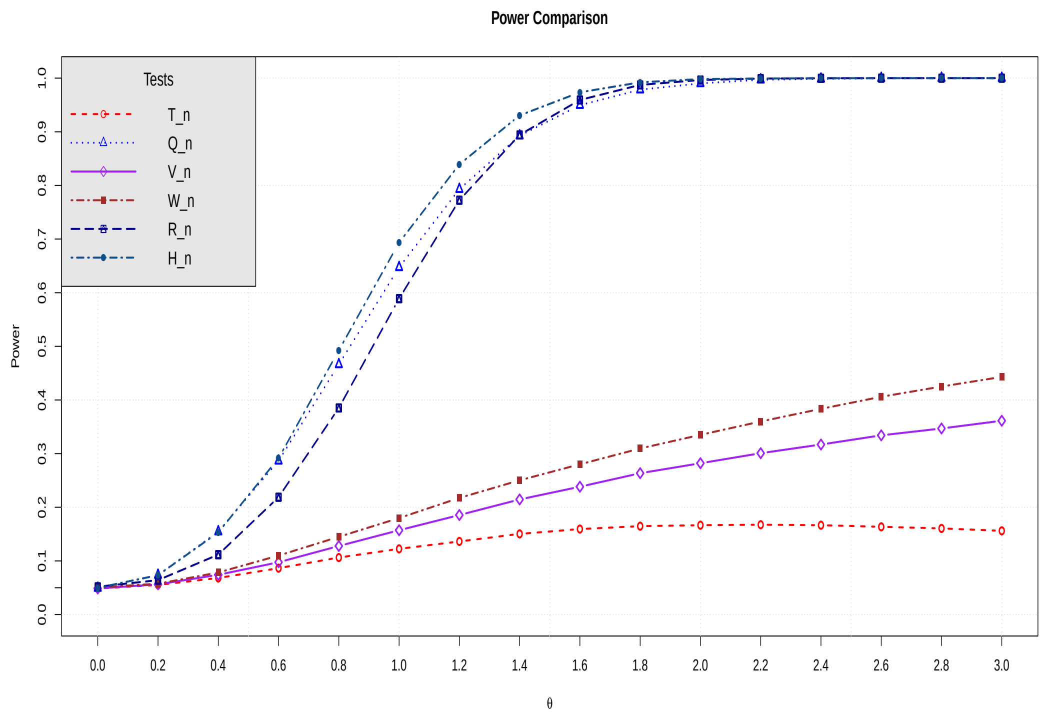


Figure 6.6: Power Comparison - Example 6.2.1; $y = 2 + x_1 + x_2 + \theta x_1 x_2 + \epsilon$

The patterns are similar to Example 6.1.3 with a difference in performance of Q_n. Tests based on partial sums of residuals, except for Q_n, produced low power. Among this group, W_n is the best across θ with a maximum of 0.44. Whereas, tests based on absolute residuals produced satisfactory performance. H_n has the highest power across θ followed closely by Q_n and almost indistinguishably for $\theta \leq 0.8$.

At $\theta = 2$, H_n, R_n and Q_n have reached power 1 while the power for the rest of the tests is below 0.4. At $\theta = 1.4$, H_n is 270% , 333% and 520% more powerful than W_n, V_n and T_n respectively. For the same value of theta, W_n is 16% and 66% more powerful than V_n and T_n respectively.

EXAMPLE 6.2.2: x_1 and x_2 are sampled from $U(-2,2)$ and χ_1^2 respectively with $n = 70$. The response y follows the relationship

$$y = 2 + 3x_1 - x_2 + \theta x_1^2 + \epsilon, \quad \epsilon \sim N(0, 2^2)$$

θ ranges between 0 and 1. The fitted linear model does not include the quadratic term x_1^2. The data are ordered according to x_1. The results are shown in Figure 6.7.

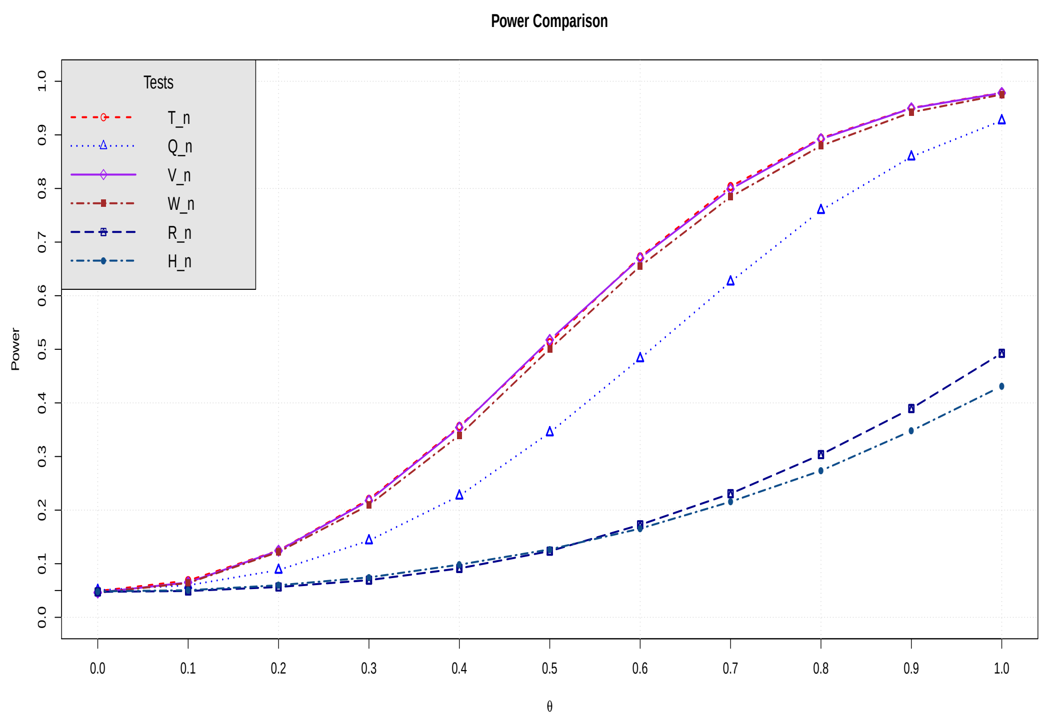


Figure 6.7: Power Comparison - Example 6.2.2; $y = 2 + 3x_1 - x_2 + \theta x_1^2 + \epsilon$

The results are similar to those of Example 6.1.2. The T_n, W_n and V_n produced almost identical powers. The difference between these tests and Q_n is more obvious.

None of the tests based on absolute residuals, R_n and H_n exceeded 0.5 with a slightly better performance for the former for $\theta \geq 0.6$. T_n is 49% and 315% more powerful than Q_n and R_n respectively at $\theta = 0.6$.

To assess the effect of the distribution of error vector, the example is repeated twice. First, we assume $\epsilon \sim U(-2\sqrt{3}, 2\sqrt{3})$ and the simulations are drawn from $U(-1,1)$. Then we alter the distribution to a t distribution with 6 d.f and scale parameter $2\sqrt{\frac{4}{6}}$ and the simulations are based on t with 6 d.f. Note that both distributions have mean 0 and standard deviation 2 as $N(0, 2^2)$. Similar to Example 6.1.3, R_n and H_n are adjusted according to the assumed distribution. The results are shown in Figures 6.8 and 6.9.

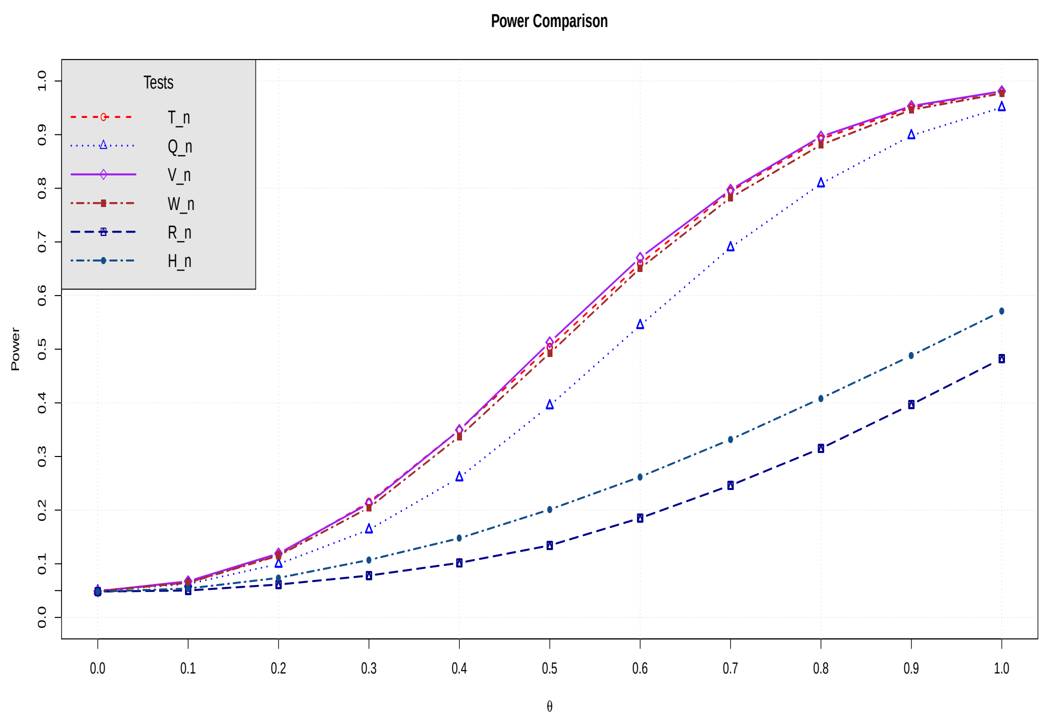


Figure 6.8: Power Comparison - Example 6.2.2; $\epsilon \sim U(-2\sqrt{3}, 2\sqrt{3})$

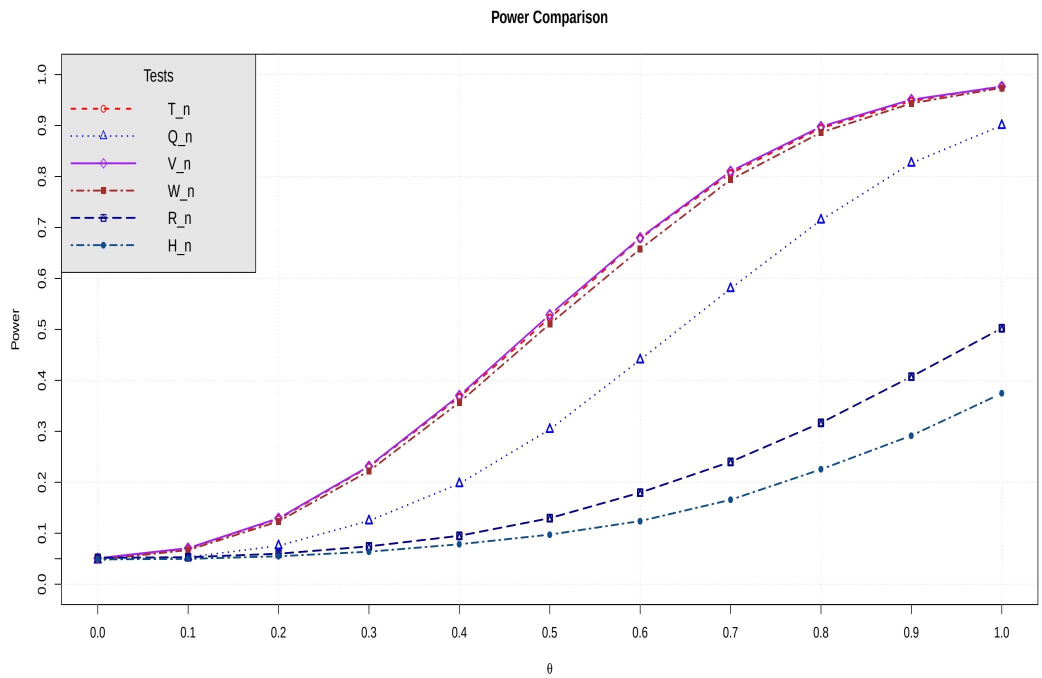


Figure 6.9: Power Comparison - Example 6.2.2; $\epsilon \sim t(6, 2\sqrt{\frac{4}{6}})$

T_n, W_n and V_n are the least affected by the changing the distribution. They produced almost identical power across distributions. The power of Q_n has increased a little and decreased by almost the same amount when assuming uniform and t correspondingly. The power of H_n has increased and exceeded R_n for every value of θ when assuming uniform distribution. Whereas R_n has a slight advantage when the distribution is t as in the normal case.

Generally, It can be observed that assuming uniform leads to a higher power for Q_n, R_n and H_n and lower power when assuming t with respect to normality. The difference might be large as seen in Example (6.1.3) or small as in Example (6.2.2). T_n, W_n and V_n are little affected by changing the distribution showing robustness against distributional assumptions.

EXAMPLE 6.2.3: x_1 and x_2 are sampled from $N(0,1)$ and $F(4,10)$ respectively with $n = 70$. The response y follows the relationship

$$y = 2 + 2x_1 + 3x_2^{\theta} + \epsilon, \quad \epsilon \sim N(0, 2^2)$$

θ ranges between 1 and 2. The fitted model corresponds to $\theta = 1$. The data are ordered according to x_2. The results are shown in Figure 6.10.

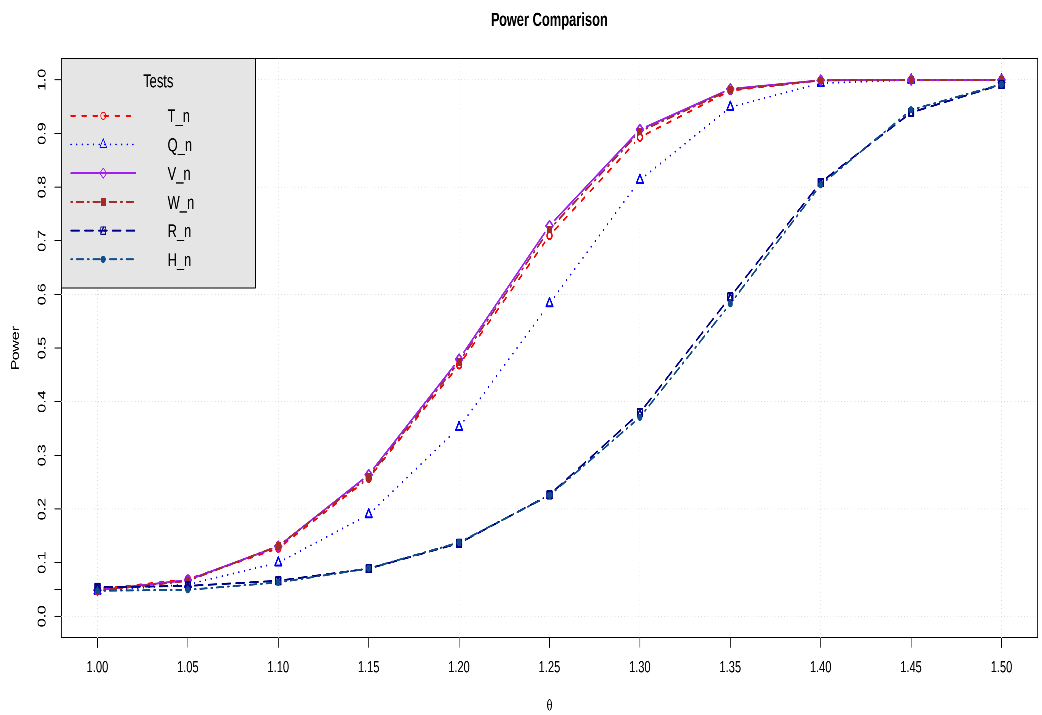


Figure 6.10: Power Comparison - Example 6.2.3; $y = 2 + 2x_1 + 3x_2^{\theta} + \epsilon$

As θ increases the power of all tests increase. For all tests, a power of 1.00 has been obtained when $\theta = 1.5$. T_n, W_n and V_n reached a power of 1 faster than the rest of the tests followed by Q_n. T_n, W_n and V_n have achieved power of 0.9 at $\theta = 1.3$. For the same value of θ, R_n and H_n have attained a power around 0.4.

Chapter 7

Summary

This dissertation introduced methods for testing the lack of fit in linear models. More precisely, we proposed statistical procedures to investigate the validity of the mean function specification, i.e. linearity assumption. C-L suggested two lack-of-fit tests based on partial sums of residuals. Following C-L, we presented additional test statistics based on partial sums of residuals. We gave assumptions required to achieve convergence in distribution before deriving their asymptotic distributions. Ordering methods and a consistent estimator of σ were introduced. We studied the small sample behavior for each test statistic. It was clear that the test statistics suffer from slow convergence leading to poor asymptotic approximation of the P values for small, moderate, and even somewhat large samples. Thus, we presented an effective approximation to P values through Monte Carlo simulations. We proposed new tests based on partial sums of absolute residuals. The use of absolute residuals allowed detection of lack of fit that was previously not possible with partial sums of residuals. Finally, the C-L tests and the new tests were compared through several examples and simulation studies in terms of their abilities in detecting an existing lack-of-fit.

Appendix A

Asymptotic distributions

All of the asymptotic distributions are extensions of results by Darling and Erdös (1956) and Erdös or Kac (1945) modified to deal with the necessity of estimating $\boldsymbol{\beta}$ and σ in the linear model. In fact, for known $\boldsymbol{\beta}$ and σ the results follow directly from Darling and Erdös (1956) or Erdös and Kac (1945). Assumptions (a) and (b) of section 2 are needed to handle the estimation. The following three lemmas are proved in Christensen and Lin (2012). Lemma (4) follows from Lemma (1) and Lemma (2).

Lemma 1. $\sqrt{n} \parallel \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \parallel / a_n$ is bounded a.s, where $a_n = \sqrt{2 \log\log n}$ and $\hat{\boldsymbol{\beta}}_n$ is the least squares estimate of $\boldsymbol{\beta}$.

Lemma 2. If assumption (a) is satisfied, $\left| \sum_{i=1}^{m} \frac{\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma\sqrt{\tilde{n}}} \right|$ and $\sum_{i=1}^{m} \left| \frac{\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma\sqrt{\tilde{n}}} \right|$ converge in probability to 0 as $\tilde{n} \to \infty$, for any integer $m \in \{1, 2..., \tilde{n}\}$, where $\tilde{n} = \lceil n/(\log\log n)^{1+\delta} \rceil$ for $\delta > 0$.

Lemma 3. If assumption (a) is satisfied, $a_{\tilde{n}} \max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} \frac{\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma\sqrt{m}} \right|$ and $a_{\tilde{n}} \sum_{i=1}^{m} \left| \frac{\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma\sqrt{m}} \right|$ converge in probability to 0 as $\tilde{n} \to \infty$, for any integer $m \in$

$\{1, 2..., \tilde{n}\}$, where $\tilde{n} = \lceil n/(\log\log n)^{1+\delta} \rceil$ for $\delta > 1$ and $a_{\tilde{n}} = \sqrt{2\log\log\tilde{n}}$.

Lemma 4. It follows from Lemma (1) and Lemma (2), $\frac{1}{\sqrt{\tilde{n}}}\max_{1\le m\le\tilde{n}}\left|\sum_{i=1}^{m}\frac{\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})}{\sigma}\right|$ and $\frac{1}{\tilde{n}}\max_{1\le m\le\tilde{n}}\sum_{i=1}^{m}\left(\frac{\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})}{\sigma}\right)^2$ converge in probability to 0 as $n\to\infty$.

A.1 Proof of Theorem 1

Proof of Theorem 1. If $\boldsymbol{\beta}$ and σ are known, then $e_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$ for $i \in \{1, ..., n\}$ are independently distributed with $E(e_i) = 0$ and $Var(e_i) = \sigma^2$. By Erdös and Kac (1945) part(4), as $\tilde{n}\to\infty$,

$$\frac{1}{\sqrt{\tilde{n}^3}}\sum_{m=1}^{\tilde{n}}\left|\sum_{i=1}^{m}\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right| \xrightarrow{\mathcal{L}} W,$$

with the distribution of W indicated in Theorem 1. First consider,

$$\begin{aligned}
&\frac{1}{\sqrt{\tilde{n}^3}}\sum_{m=1}^{\tilde{n}}\left|\sum_{i=1}^{m}\frac{y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n}{\sigma}\right| \\
&= \frac{1}{\sqrt{\tilde{n}^3}}\sum_{m=1}^{\tilde{n}}\left|\sum_{i=1}^{m}\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}+\boldsymbol{x}_i^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_n)}{\sigma}\right| \\
&= \frac{1}{\sqrt{\tilde{n}^3}}\sum_{m=1}^{\tilde{n}}\left|\sum_{i=1}^{m}\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}+\sum_{i=1}^{m}\frac{\boldsymbol{x}_i^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_n)}{\sigma}\right| \\
&\le \frac{1}{\sqrt{\tilde{n}^3}}\sum_{m=1}^{\tilde{n}}\left|\sum_{i=1}^{m}\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right| + \frac{1}{\sqrt{\tilde{n}^3}}\sum_{m=1}^{\tilde{n}}\left|\sum_{i=1}^{m}\frac{\boldsymbol{x}_i^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_n)}{\sigma}\right|
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\frac{1}{\sqrt{\tilde{n}^3}} \sum_{m=1}^{\tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} \right| \\
&= \frac{1}{\sqrt{\tilde{n}^3}} \sum_{m=1}^{\tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n + \boldsymbol{x}_i^T (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right| \\
&= \frac{1}{\sqrt{\tilde{n}^3}} \sum_{m=1}^{\tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\sigma} + \sum_{i=1}^{m} \frac{\boldsymbol{x}_i^T (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right| \\
&\leq \frac{1}{\sqrt{\tilde{n}^3}} \sum_{m=1}^{\tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - x_i^T \hat{\boldsymbol{\beta}}_n}{\sigma} \right| + \frac{1}{\sqrt{\tilde{n}^3}} \sum_{m=1}^{\tilde{n}} \left| \sum_{i=1}^{m} \frac{\boldsymbol{x}_i^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)}{\sigma} \right|
\end{aligned}$$

It suffices to show that $\frac{1}{\sqrt{\tilde{n}^3}} \sum_{m=1}^{\tilde{n}} \left| \sum_{i=1}^{m} \frac{\boldsymbol{x}_i^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)}{\sigma} \right|$ converges in probability to 0 to prove that $\frac{1}{\sqrt{\tilde{n}^3}} \sum_{m=1}^{\tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\sigma} \right|$ and $\frac{1}{\sqrt{\tilde{n}^3}} \sum_{m=1}^{\tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} \right|$ have the same limiting distribution.

By Lemma (4),

$$\begin{aligned}
\frac{1}{\sqrt{\tilde{n}^3}} \sum_{m=1}^{\tilde{n}} \left| \sum_{i=1}^{m} \frac{\boldsymbol{x}_i^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)}{\sigma} \right| &= \frac{1}{\tilde{n}} \sum_{m=1}^{\tilde{n}} \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{m} \frac{\boldsymbol{x}_i^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)}{\sigma} \right| \\
&\leq \frac{1}{\tilde{n}} \sum_{m=1}^{\tilde{n}} \max_{1 \leq m \leq \tilde{n}} \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{m} \frac{\boldsymbol{x}_i^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)}{\sigma} \right| \\
&= \max_{1 \leq m \leq \tilde{n}} \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{m} \frac{\boldsymbol{x}_i^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)}{\sigma} \right| \xrightarrow{p} 0.
\end{aligned}$$

Moreover,

$$W_n = \frac{1}{\sqrt{\tilde{n}^3}} \sum_{m=1}^{\tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\hat{\sigma}_n} \right| = \frac{\sigma}{\hat{\sigma}_n} \frac{1}{\sqrt{\tilde{n}^3}} \sum_{m=1}^{\tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\sigma} \right|.$$

By condition (b), $\sigma/\hat{\sigma}_n \xrightarrow{p} 1$, hence $W_n \xrightarrow{\mathcal{L}} W$. □

A.2 Proof of Theorem 2

Proof of Theorem 2. If $\boldsymbol{\beta}$ and σ are known, then $e_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$ for $i \in \{1, ..., n\}$ are independently distributed with $E(e_i) = 0$ and $Var(e_i) = \sigma^2$. It follows immediately by Erdös and Kac (1945) part(3)

$$\frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}\left(\sum_{i=1}^{m}\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2 \xrightarrow{\mathcal{L}} V,$$

with the distribution of V indicated in Theorem (1).

Let

$$a_m = \sum_{i=1}^{m}\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}$$

and

$$b_m = \sum_{i=1}^{m}\frac{\boldsymbol{x}_i^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)}{\sigma}$$

Then,

$$\begin{aligned}
&\frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}\left(\sum_{i=1}^{m}\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n}{\sigma}\right)^2 \\
= &\frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}\left(\sum_{i=1}^{m}\frac{(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n) + \boldsymbol{x}_i^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)}{\sigma}\right)^2 \\
= &\frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}\left(\sum_{i=1}^{m}\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2 + \frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}\left(\sum_{i=1}^{m}\frac{\boldsymbol{x}_i^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)}{\sigma}\right)^2 \\
+ &\frac{2}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}\left(\sum_{i=1}^{m}\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)\left(\sum_{i=1}^{m}\frac{\boldsymbol{x}_i^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)}{\sigma}\right) \\
= &\frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}a_m^2 + \frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}b_m^2 + \frac{2}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}a_m b_m
\end{aligned}$$

We have already established that the first term $\frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}a_m^2$ converges in distribution to V. It suffices to show that the second term $\frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}b_m^2$ and the third term

$\frac{2}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}} a_m b_m$ converge in probability to 0.

By Lemma (4),

$$0 \leq \frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}} b_m^2 \leq \frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}} \max_{1\leq m\leq \tilde{n}} b_m^2 = \max_{1\leq m\leq \tilde{n}} \frac{1}{\tilde{n}} b_m^2 \xrightarrow{p} 0$$

Therefore, $\frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}} b_m^2 \xrightarrow{p} 0$ as $\tilde{n} \to \infty$.

For the third term,

$$\begin{aligned}
\left|\frac{2}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}} a_m b_m\right| &\leq \frac{2}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}} |a_m||b_m| \\
&\leq \frac{2}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}} |a_m| \max_{1\leq m\leq \tilde{n}} |b_m| \\
&\leq \max_{1\leq m\leq \tilde{n}} |b_m| \frac{2}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}} |a_m| \\
&= \frac{1}{\sqrt{\tilde{n}}} \max_{1\leq m\leq \tilde{n}} |b_m| \frac{2}{\tilde{n}^{\frac{3}{2}}}\sum_{m=1}^{\tilde{n}} |a_m|
\end{aligned}$$

By Lemma (4), $\frac{1}{\sqrt{\tilde{n}}}\max_{1\leq m\leq \tilde{n}} |b_m| \xrightarrow{p} 0$ and by Erdös and Kac (1945) part(4), $\frac{1}{\tilde{n}^{\frac{3}{2}}}\sum_{m=1}^{\tilde{n}} |a_m|$ converges in distribution.

So, by Slutsky's theorem, $\frac{1}{\sqrt{\tilde{n}}}\max_{1\leq m\leq \tilde{n}} |b_m| \frac{1}{\tilde{n}^{\frac{3}{2}}}\sum_{m=1}^{\tilde{n}} |a_m| \xrightarrow{p} 0$.

Thus,

$$\left|\frac{2}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}} a_m b_m\right| \xrightarrow{p} 0$$

and

$$\frac{2}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}} a_m b_m \xrightarrow{p} 0$$

as $\tilde{n} \to \infty$. This establishes that $\frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}\left(\sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n}{\sigma}\right)^2$ has limiting distribution V.

Finally,

$$V_n = \frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}\left(\sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n}{\hat{\sigma}_n}\right)^2 = \frac{\sigma^2}{\hat{\sigma}_n^2}\frac{1}{\tilde{n}^2}\sum_{m=1}^{\tilde{n}}\left(\sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n}{\sigma}\right)^2.$$

By assumption (b), $\sigma/\hat{\sigma}_n \xrightarrow{p} 1$, hence $V_n \xrightarrow{\mathcal{L}} V$. □

A.3 Proof of Theorem 3

Proof of Theorem 3. Suppose $\boldsymbol{\beta}$ is known, then $e_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$ for $i \in \{1, ..., n\}$ are independently distributed with $E(e_i) = 0$ and $Var(e_i) = \sigma^2$. With $e_i/\sigma \sim N(0,1)$, hence $|e_i/\sigma|$ are independent identically distributed ($i.i.d.$) random variables with expected value $\sqrt{\frac{2}{\pi}}$ and variance $1 - \frac{2}{\pi}$. Hence $w_i = \frac{|e_i|}{\sigma\sqrt{1-\frac{2}{\pi}}} - \sqrt{\frac{2}{\pi-2}}$ are $i.i.d$ random variables with expected value 0 and variance 1. By Erdös and Kac (1945) part(1) , as $\tilde{n} \to \infty$,

$$\begin{aligned}
\frac{1}{\sqrt{\tilde{n}}} \max_{1\le m\le \tilde{n}} \sum_{i=1}^{m} w_i &= \frac{1}{\sqrt{\tilde{n}}} \max_{1\le m\le \tilde{n}} \left[\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}} \sum_{i=1}^{m} |e_i| - m\sqrt{\frac{2}{\pi-2}} \right] \\
&= \frac{1}{\sqrt{\tilde{n}}} \max_{1\le m\le \tilde{n}} \left[\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}} \sum_{i=1}^{m} \left| y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} \right| - m\sqrt{\frac{2}{\pi-2}} \right] \xrightarrow{\mathcal{L}} Z.
\end{aligned}$$

Let $\hat{\boldsymbol{\beta}}_n$ be the least square estimator of $\boldsymbol{\beta}$ and let k and q be numbers that satisfy

$$\begin{aligned}
&\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}} \sum_{i=1}^{k} \left| y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n \right| - k\sqrt{\frac{2}{\pi-2}} \\
&= \max_{1\le m\le \tilde{n}} \left[\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}} \sum_{i=1}^{m} \left| y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n \right| - m\sqrt{\frac{2}{\pi-2}} \right]
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}} \sum_{i=1}^{q} \left| y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} \right| - q\sqrt{\frac{2}{\pi-2}} \\
&= \max_{1\le m\le \tilde{n}} \left[\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}} \sum_{i=1}^{m} \left| y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} \right| - m\sqrt{\frac{2}{\pi-2}} \right]
\end{aligned}$$

$$\text{Now, } \frac{1}{\sqrt{\tilde{n}}}\left[\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}}\sum_{i=1}^{k}\left|y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n\right|-k\sqrt{\frac{2}{\pi-2}}\right]$$

$$\begin{aligned}
&= \frac{1}{\sqrt{\tilde{n}}}\left[\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}}\sum_{i=1}^{k}\left|y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}+\boldsymbol{x}_i^T\boldsymbol{\beta}-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n\right|-k\sqrt{\frac{2}{\pi-2}}\right]\\
&\leq \frac{1}{\sqrt{\tilde{n}}}\left[\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}}\sum_{i=1}^{k}\left|y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right|-k\sqrt{\frac{2}{\pi-2}}\right]\\
&+ \frac{1}{\sigma\sqrt{\tilde{n}(1-\frac{2}{\pi})}}\sum_{i=1}^{k}\left|\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\right|\\
&\leq \frac{1}{\sqrt{\tilde{n}}}\max_{1\leq m\leq\tilde{n}}\left[\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}}\sum_{i=1}^{m}\left|y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right|-m\sqrt{\frac{2}{\pi-2}}\right]\\
&+ \frac{1}{\sigma\sqrt{\tilde{n}(1-\frac{2}{\pi})}}\sum_{i=1}^{k}\left|\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\right|
\end{aligned}$$

Also,

$$\frac{1}{\sqrt{\tilde{n}}}\left[\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}}\sum_{i=1}^{q}\left|y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right|-q\sqrt{\frac{2}{\pi-2}}\right]$$

$$\begin{aligned}
&= \frac{1}{\sqrt{\tilde{n}}}\left[\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}}\sum_{i=1}^{q}\left|y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n+\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n-\boldsymbol{x}_i^T\boldsymbol{\beta}\right|-q\sqrt{\frac{2}{\pi-2}}\right]\\
&\leq \frac{1}{\sqrt{\tilde{n}}}\left[\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}}\sum_{i=1}^{q}\left|y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n\right|-q\sqrt{\frac{2}{\pi-2}}\right]\\
&+ \frac{1}{\sigma\sqrt{\tilde{n}(1-\frac{2}{\pi})}}\sum_{i=1}^{q}\left|\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\right|\\
&\leq \frac{1}{\sqrt{\tilde{n}}}\max_{1\leq m\leq\tilde{n}}\left[\frac{1}{\sigma\sqrt{1-\frac{2}{\pi}}}\sum_{i=1}^{m}\left|y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n\right|-m\sqrt{\frac{2}{\pi-2}}\right]\\
&+ \frac{1}{\sigma\sqrt{\tilde{n}(1-\frac{2}{\pi})}}\sum_{i=1}^{q}\left|\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\right|.
\end{aligned}$$

Thus $A - B \leq C \leq A + B$. By Lemma (2), both $\sum_{i=1}^{q} \left| \boldsymbol{x}_i^T (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \right| / \sigma\sqrt{\tilde{n}}$ and $\sum_{i=1}^{k} \left| \boldsymbol{x}_i^T (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \right| / \sigma\sqrt{\tilde{n}}$ converge in probability to 0 as $n \to \infty$, therefore

$$\tilde{R}_n = \frac{1}{\sqrt{\tilde{n}}} \max_{1 \leq m \leq \tilde{n}} \left[\frac{1}{\sigma\sqrt{1 - \frac{2}{\pi}}} \sum_{i=1}^{m} \left| y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n \right| - m\sqrt{\frac{2}{\pi - 2}} \right]$$

has the same limiting distribution as

$$\frac{1}{\sqrt{\tilde{n}}} \max_{1 \leq m \leq \tilde{n}} \left[\frac{1}{\sigma\sqrt{1 - \frac{2}{\pi}}} \sum_{i=1}^{m} \left| y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} \right| - m\sqrt{\frac{2}{\pi - 2}} \right].$$

Hence $\tilde{R}_n \xrightarrow{\mathcal{L}} Z$. By assumption (b), $\sigma/\hat{\sigma}_n \xrightarrow{p} 1$. Thus $\dfrac{\sigma}{\hat{\sigma}_n}\tilde{R}_n \xrightarrow{\mathcal{L}} Z$. R_n is identical to $\tilde{R}_n$ except that $\hat{\sigma}_n$ replaces σ. R_n can be written as

$$\frac{1}{\sqrt{\tilde{n}}} \max_{1 \leq m \leq \tilde{n}} \left[\frac{1}{\hat{\sigma}_n\sqrt{1 - \frac{2}{\pi}}} \sum_{i=1}^{m} \left| y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n \right| - \frac{\sigma}{\hat{\sigma}_n} m\sqrt{\frac{2}{\pi - 2}} + \frac{\sigma}{\hat{\sigma}_n} m\sqrt{\frac{2}{\pi - 2}} - m\sqrt{\frac{2}{\pi - 2}} \right]$$

Applying the following inequality that can be applied to any two sequences of real numbers a_m and b_m, $\max_m a_m - \max_m |b_m| \leq \max_m (a_m + b_m) \leq \max_m a_m + \max_m b_m$, we obtain

$$\frac{\sigma}{\hat{\sigma}_n} R_n - \frac{1}{\sqrt{\tilde{n}}} \max_{1 \leq m \leq \tilde{n}} \left| m\sqrt{\frac{2}{\pi - 2}} \left(\frac{\sigma}{\hat{\sigma}_n} - 1 \right) \right|$$
$$\leq R_n \leq \frac{\sigma}{\hat{\sigma}_n} \tilde{R}_n + \frac{1}{\sqrt{\tilde{n}}} \max_{1 \leq m \leq \tilde{n}} \left[m\sqrt{\frac{2}{\pi - 2}} \left(\frac{\sigma}{\hat{\sigma}_n} - 1 \right) \right]$$

It suffices to show that $\dfrac{1}{\sqrt{\tilde{n}}} \max_{1 \leq m \leq \tilde{n}} \left[m\sqrt{\dfrac{2}{\pi - 2}} \left(\dfrac{\sigma}{\hat{\sigma}_n} - 1 \right) \right]$ converges to 0 to establish that R_n and $\frac{\sigma}{\hat{\sigma}_n}\tilde{R}_n$ have the same limiting distribution which is Z.

$$\begin{aligned}
\frac{1}{\sqrt{\tilde{n}}}\max_{1\le m\le\tilde{n}}\left[m\sqrt{\frac{2}{\pi-2}}\left(\frac{\sigma}{\hat{\sigma}_n}-1\right)\right] &= \frac{1}{\sqrt{\tilde{n}}}\sqrt{\frac{2}{\pi-2}}\left(\frac{\sigma}{\hat{\sigma}_n}-1\right)\max_{1\le m\le\tilde{n}} m \\
&= \sqrt{\frac{2}{\pi-2}}\sqrt{\tilde{n}}\left(\frac{\sigma}{\hat{\sigma}_n}-1\right) \\
&= \sqrt{\frac{2}{\pi-2}}\frac{\sqrt{\tilde{n}}}{\sqrt{n}}\sqrt{n}\left(\frac{\sigma}{\hat{\sigma}_n}-1\right)
\end{aligned}$$

Using assumption (b) and the slow convergence of $\tilde{n}$ to infinity, we get the convergence to 0. By (b), $\sqrt{n}\left(\frac{\sigma}{\hat{\sigma}_n}-1\right)$ is bounded in probability and $\frac{\sqrt{\tilde{n}}}{\sqrt{n}}$ converges to 0. Then by slutsky's theorem, $\frac{1}{\sqrt{\tilde{n}}}\max_{1\le m\le\tilde{n}}\left[m\sqrt{\frac{2}{\pi-2}}\left(\frac{\sigma}{\hat{\sigma}_n}-1\right)\right]$ converges to 0. The convergence of H_n follows similarly but Darling and Erdös (1955) Theorem 1 is used instead of Erdös and Kac (1945) part(1) and Lemma 3 instead of Lemma 2. □

A.4 Additional test Statistics

A.4.1 Z_n

Following C-L, if $\boldsymbol{\beta}$ and σ are known, then $e_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$ for $i \in \{1, ..., n\}$ are independently distributed with $E(e_i) = 0$ and $Var(e_i) = \sigma^2$. By Erdös and Kac (1945) part(1) , as $\tilde{n} \to \infty$,

$$\frac{1}{\sqrt{\tilde{n}}}\max_{1\le m\le\tilde{n}}\sum_{i=1}^{m}\frac{y_i - x_i^T\beta}{\sigma} \xrightarrow{\mathcal{L}} U,$$

where U has a half-normal distribution.

Let $\hat{\boldsymbol{\beta}}_n$ be the least square estimator of $\boldsymbol{\beta}$ and let k and q be numbers that satisfy

$$\sum_{i=1}^{k}\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n}{\sigma} = \max_{1\le m\le\tilde{n}}\sum_{i=1}^{m}\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n}{\sigma},$$

and

$$\sum_{i=1}^{q} \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} = \max_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma},$$

so that,

$$\begin{aligned}
\frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\sigma} &= \frac{1}{\sqrt{\tilde{n}}} \sum_{i=1}^{k} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\sigma} \\
&= \frac{1}{\sqrt{\tilde{n}}} \sum_{i=1}^{k} \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} + \frac{1}{\sqrt{\tilde{n}}} \sum_{i=1}^{k} \frac{\boldsymbol{x}_i^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)}{\sigma} \\
&\le \frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} + \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{k} \frac{\boldsymbol{x}_i^T (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right|.
\end{aligned}$$

Also,

$$\begin{aligned}
\frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\sigma} &\ge \frac{1}{\sqrt{\tilde{n}}} \sum_{i=1}^{q} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\sigma} \\
&= \frac{1}{\sqrt{\tilde{n}}} \sum_{i=1}^{q} \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} - \frac{1}{\sqrt{\tilde{n}}} \sum_{i=1}^{q} \frac{\boldsymbol{x}_i^T (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \\
&\ge \frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} - \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{q} \frac{\boldsymbol{x}_i^T (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right|.
\end{aligned}$$

By Lemma (2), both $\left| \sum_{i=1}^{q} \boldsymbol{x}_i^T (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \right| / \sigma\sqrt{\tilde{n}}$ and $\left| \sum_{i=1}^{k} \boldsymbol{x}_i^T (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \right| / \sigma\sqrt{\tilde{n}}$ converge in probability to 0 as $n \to \infty$, therefore

$$\frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\sigma}$$

has the same limiting distribution as

$$\frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma}$$

Now,

$$Z_n = \frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\hat{\sigma}_n} = \frac{\sigma}{\hat{\sigma}_n \sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_n}{\sigma}.$$

By assumption (b), $\sigma/\hat{\sigma}_n \xrightarrow{p} 1$, hence $Z_n \xrightarrow{\mathcal{L}} U$. □

A.4.2 M_n

If $\boldsymbol{\beta}$ and σ are known, then $-e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} - y_i$ for $i \in \{1, ..., n\}$ are independently distributed with $E(-e_i) = 0$ and $Var(-e_i) = \sigma^2$. By Erdös and Kac (1945) part(1) , as $\tilde{n} \to \infty$,

$$\frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{x_i^T\beta - y_i}{\sigma} \xrightarrow{\mathcal{L}} U,$$

where U has a half-normal distribution.

Then, for $u > 0$,

$$Pr\left[\frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{x_i^T\beta - y_i}{\sigma} < u\right] = Pr\left[\frac{1}{\sqrt{\tilde{n}}} \min_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{y_i - x_i^T\beta}{\sigma} > -u\right].$$

Hence, $Pr[\frac{1}{\sqrt{\tilde{n}}} \min_{1 \le m \le \tilde{n}} \sum_{i=1}^{m} \frac{y_i - x_i^T\beta}{\sigma} < -u]$ converges to $\left(\frac{2}{\pi}\right)^{\frac{1}{2}} \int_u^\infty \exp(-\frac{r^2}{2})\, dr$ for $r > 0$. The rest of the proof is similar to the proof of Z_n. □

A.4.3 G_n

It follows from Darling and Erdös (1955) Part(1) and Lemma (3). The rest of the proof is similar to the proofs of Z_n and M_n. □

References

[1] Cameron, R.H., and Martin, W. T., (1944), " The Wiener Measure of Hilbert Neighborhoods in the Space of Real Continuous functions," *Journal of Mathematics and Physics,* 23, 195-209.

[2] Christensen, R., and Sun, S.K, (2010), "Alternative Goodness-of-Fit Tests for Linear Models," *Journal of the American Statistical Association,* 105, 291-301.

[3] Christensen, R., and Lin, Yong (2014), " Lack-of-fit tests based on partial sums of residuals, " *Communications in Statistics: Theory and Methods,* to appear.

[4] Darling, D. A., and Erdös, P. (1956), "A Limit Theorem for the Maximum of Normalized Sums of Independent Random Variables," *Duke Mathematical Journal,* 23, 143-155.

[5] Eicker, F. (1979), "The Asymptotic Distribution of the Suprema of the Standardized Empirical Processes," *The Annals of Statistics,* 7, 116-138.

[6] Erdös, P., and Kac M. (1946), "On Certain Limit Theorems of the Theory of probability," *Bulletin of American Mathematical Society*, 52, 292-302.

[7] Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*, New York: Springer.

[8] Koval′, V.A. (2002), "The Law of the Iterated Logarithm for Matrix-Normed Sums of Independent Random Variables and Its Applications," *Mathematical Notes*, 72, 331-336.

[9] MacKinnon G.J., "Bootstrap Inference in Econometrics," *The Canadian Journal of Economics* , 35, 615-645.

[10] Takács,L. (1993), "On the Distribution of the Integral of the Absolute Value of the Brownian Motion," *The Annals of Applied Probability,* 3, 186-197.