2009

Generalized Inference in Linear Regression Models

Quazi Ibrahim
University of Windsor

Generalized Inference in Linear Regression Models

by

Quazi Imad Uddin Ibrahim

A Thesis
Submitted to the Faculty of Graduate Studies
through the Department of Mathematics and Statistics
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada

2009

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

In this thesis, we consider inference problems in linear regression under both homoscedasticity and heteroscedasticity of the error noise. Namely, we construct generalized confidence regions and generalized confidence intervals for regression coefficients of linear regression models. Regressor variables are considered non-stochastic. Independent normal errors with zero mean and constant or varying dispersion are considered. The regression data from two different regimes are considered. In testing the equality of the regression coefficients in the two regimes under heteroscedasticity, we develop the generalized pivotal quantities of their differences and the generalized p-values. Generalized methods of inference are especially useful in multiparameter cases where nontrivial tests are difficult to obtain. We propose generalized test variables and generalized p-values to test the equality of the sets of regression coefficients of the two regimes. The test can be applied efficiently for all sample sizes and for homoscedastic as well as heteroscedastic cases. The simulation study shows that the proposed method preserves the nominal significance level and maintain satisfactory power under heteroscedasticity, and for small and moderate sample sizes. We also construct the generalized confidence region for the difference of the two sets of regression coefficients. When the regression coefficients remained the same for the two regimes under heteroscedasticity, we propose generalized confidence regions and generalized confidence intervals for the regression parameters.

We applied the proposed method on the community health study data of Sarnia in 2005 and the US gasoline consumption data before and after the 1973 oil crisis. The analysis results show that, for both data sets, the regime change is statistically significant at 5% level.

Dedication

This thesis is dedicated to my family. I am grateful to my family for their continuous support and encouragement. I remember their love, affection and inspiration throughout my life.

Acknowledgements

All praises are due to almighty Allah. I express my sacred gratitude to the omnipotent and omniscient Allah, Who gave us the wisdom of creation. Without His blessings this task could not have ended.

I would like to thank my supervisors Dr. Karen Fung and Dr. Sévérien Nkurunziza. I am very grateful to them for their continuing support and encouragement during the preparation of the thesis. Their guidance and sincere co-operation helped me to complete the task.

I would like to thank Dr. Ejaz Ahmed, Department of Mathematics and Statistics, and Dr. Robin Gras, School of Computer Science, for serving on my thesis committee.

I express my cordial gratitude to all my respected teachers of the Department of Mathematics and Statistics of the University of Windsor.

I also acknowledge the help, co-operation of all the members of staff of this department.

I would like to thank Kazi Azad, Tanvir Quadir, S.M. Enayetur Raheem, Fuqi Chen and all of my class mates and friends for their help, co-operation and encouragement to accomplish this research work successfully.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Literature review

In estimation, we often infer the true value of the parameter or a function of the parameter is contained in an interval with certain confidence. These intervals are called confidence intervals. To define confidence intervals consider a random sample $Y = (Y_1, Y_2, \ldots, Y_n)$ from a probability density function (pdf) or a probability mass function (pmf) $f(y|\theta)$ where θ is an unknown parameter. Suppose $T_1(Y)$ and $T_2(Y)$ are two statistics such that

$$Pr[T_1(Y) \leq \theta \leq T_2(Y)] = \gamma, \quad \gamma \in (0, 1).$$

If the realized values of $T_1(Y)$ and $T_2(Y)$ are a and b respectively, [a, b] is called a 100γ percent confidence interval for θ. Here γ is called the confidence coefficient. Typical values of γ are 0.9, 0.95 and 0.99. One method of constructing confidence interval is to use a pivotal quantity.

Definition (Pivotal quantity): Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample from a probability density function (pdf) or a probability mass function (pmf) $f(y|\theta)$ where θ is an unknown parameter and $Q = g(Y, \theta)$ is a function of Y and θ. If Q has a probability distribution independent of any unknown parameters, it is called a pivotal quantity.

Thus for a fixed γ, there exist real numbers q_1 and q_2 ($q_1 < q_2$) such that

$$Pr[q_1 \leq g(Y,\theta) \leq q_2] = \gamma, \quad \gamma \in (0,1).$$

If $q_1 \leq g(Y,\theta) \leq q_2 \;\; \Leftrightarrow \;\; T_1(Y) \leq \theta \leq T_2(Y)$ where $T_1(Y)$ and $T_2(Y)$ are functions of sample only, the random interval $[T_1(Y),\, T_2(Y)]$ is called a 100γ percent confidence interval for θ. For an observed sample point $y = (y_1, y_2, \ldots, y_n)$, $[T_1(y),\, T_2(y)]$ is also called a 100γ percent confidence interval for θ.

In complex situations involving nuisance parameters, often the uniformly most accurate confidence intervals are unavailable. For instance, the uniformly most accurate unbiased confidence intervals for the difference in means of two independent normal populations do not exist unless the population variances are assumed equal. When the variances are heterogeneous, this problem is known as the Behrens-Fisher problem (Welch, 1938). To overcome this problem, Weerahandi (1993) introduced the concept of generalized pivotal quantity and generalized confidence interval.

Definition (Generalized pivotal quantity): Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random sample from a distribution involving parameters θ and δ. We are interested in constructing a confidence interval for θ. Let $y = (y_1, y_2, \ldots, y_n)$ be the observed sample. The generalized pivotal quantity, denoted by $R(Y, y, \theta, \delta)$, has the following three properties:

1. R is a function of Y, y, θ and δ,
2. the distribution of R is independent of θ and δ and
3. $R(y, y, \theta, \delta)$ does not depend on δ.

In this thesis, we consider a more specific generalized pivotal function that satisfies $R(y, y, \theta, \delta) = \theta$. Accordingly, a $100(1-\alpha)$ percent generalized confidence interval for θ is $[R_{\alpha/2}, R_{1-\alpha/2}]$ where $R_{\alpha/2}$ and $R_{1-\alpha/2}$ are the $100(\alpha/2)^{th}$ and $100(1-\alpha/2)^{th}$ percentiles of $R(Y, y, \theta, \delta)$. It is noticed that generalized confidence intervals can be constructed for small as well as for large samples.

In passing, recall the duality between the uniformly most accurate confidence interval and the uniformly most powerful (UMP) test. Thus, for the above mentioned problem where classical inference does not provide an optimal (small sample) confidence interval, the UMP unbiased test does not exist too. Tsui and Weerahandi (1989) introduced generalized test variables and generalized p-values to deal with this problem.

Definition (Generalized test variable): Let $Y = (Y_1, Y_2, \dots, Y_n)$ be a random sample from a distribution involving parameters θ and δ. We are interested in testing the hypothesis

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta \neq \theta_0.$$

Let $y = (y_1, y_2, \dots, y_n)$ be the observed sample. The generalized test variable, denoted by $T(Y, y, \theta, \delta)$, is a function of (Y, y, θ, δ) that satisfies the following requirements:

1. For given y and (θ_0, δ) the distribution of T is independent of the nuisance parameter δ.

2. $t = T(y, y, \theta, \delta)$ does not depend on any unknown parameters.

3. For given y and δ, $P(T \geq t)$ is stochastically monotone in θ, i.e. stochastically increasing or decreasing in θ.

In general, for a given y and δ we can take

$$T(Y, y, \theta, \delta) = R(Y, y, \theta, \delta) - \theta$$

and one can verify that, the distribution of T for given y and δ is stochastically monotone in θ. In this case the generalized p-value for testing the hypothesis is

$$\begin{aligned} P &= 2\min\left\{\sup_{\theta=\theta_0} P(T \geq t),\ \sup_{\theta=\theta_0} P(T \leq t)\right\} \\ &= 2\min\left\{\sup_{\theta=\theta_0} P(R \geq \theta),\ \sup_{\theta=\theta_0} P(R \leq \theta)\right\} \\ &= 2\min\left\{P(R \geq \theta_0),\ P(R \leq \theta_0)\right\}. \end{aligned}$$

In the same setting, if $T(Y, y, \theta, \delta)$ satisfies the following conditions, it can be considered as a generalized test variable too (Gamage et al; 2004):

1. The distribution of $T(Y, y, \theta_0, \delta)$ is free of the nuisance parameter δ.

2. $t = T(y, y, \theta_0, \delta)$ is free of δ.

3. $P(T \geq t)$ is nondecreasing in θ for fixed y and δ.

McNally, Iyer and Mathew (2003) used the generalized test variables and the generalized p-values to test population and individual bioequivalence. They showed that these tests perform better than confidence interval methods and have superior power for assessing population bioequivalence.

Lin and Lee (2004) constructed a generalized pivotal quantity to estimate the common mean of several normal populations when the variances are unknown and unequal. The proposed generalized pivotal quantity was based on the best linear unbiased estimator of the common mean.

Gamage, Mathew and Weerahandi (2004) developed a procedure based on generalized p-values to test the equality of the mean vectors of two multivariate normal populations with unequal covariance matrix. They showed the type I error probability of their generalized p-value test did not exceed the nominal level. They constructed a generalized confidence region for the difference between the mean vectors. A solution of the heteroscedastic MANOVA problem using generalized p-value was also given.

Factors that influence the gun accuracy of an M1 series tank are of considerable interest in US army. One of the factors is gun tubes. Mathew and Webb (2005) developed generalized confidence intervals and generalized test variable to compare variability among two types of gun tubes (new tubes and control tubes). They considered mixed models for their generalized inference.

Hannig, Iyer and Patterson (2006) proposed fiducial generalized pivotal quantities as a subclass of generalized pivotal quantities. They showed that generalized confidence intervals constructed based on fiducial generalized pivotal quantities have asymptotically correct frequentist coverage. They found that the subfamily of fiducial generalized pivots has a close connection with fiducial inference proposed by R.

A. Fisher.

In Metrology, a measurand is based on a sequence of measurements each with type-A and type-B errors. The measurements may come from a single experiment or several separate experiments. Wang and Iyer (2006) proposed a generalized confidence interval for a measurand based on the two measurement models with different sets of assumptions on type-B errors.

Krishnamoorthy, Mathew and Ramachandran (2006) constructed generalized p-values and generalized confidence intervals to test and compute confidence interval for the mean of a lognormal distribution. They assessed occupational exposure using the lognormal mean. They showed that their proposed methods are easy to implement and applicable to small sample sizes. They extended their procedures to compare two lognormal means and to infer a lognormal variance.

Krishnamoorthy, Mathew and Ramachandran (2007) developed generalized pivotal quantities (GPQs) for the overall mean and the variance components for one-way random effects model. The GPQs were then used to construct tolerance limits in the one-way random effects model and to construct upper confidence limits for the exceedance probabilities of occupational exposure limit.

Bebu and Mathew (2007) proposed a generalized confidence interval for the ratio of the means of a bivariate log-normal distribution. They also suggested the same approach to obtain a confidence interval for the ratio of the variances. Simulated coverage probabilities of the proposed generalized confidence intervals were found satisfactory irrespective of the sample size. The power of the tests based on the GPQs were also found satisfactory.

Li, Xu and Li (2007) proposed a method of constructing generalized p-value via the fiducial inference. They discussed the properties of the power of the generalized test. They illustrated their methods for the two-parameter exponential distribution and unbalanced two-fold nested design.

1.2 Objective

Our objective is to find generalized confidence intervals for regression coefficients, dispersion parameters and the expected response for simple and multiple linear regression models with non-stochastic explanatory variables but under different assumptions of the error distribution:

(i) Error distribution is normal with zero mean and constant variance.

(ii) Error distribution is normal with zero mean and varying dispersion.

(iii) Heteroscedasticity in two different regimes.

With the same assumptions, we construct generalized confidence regions for multiple linear regression parameters. Also, for testing the equality of corresponding regression coefficients in two different regimes with heteroscedasticity, we develop generalized test variables, confidence regions, confidence intervals and p-values.

Chapter 2

Generalized Confidence Intervals for Simple Linear Regression Parameters

In this chapter we construct generalized confidence intervals for simple linear regression parameters under different scenarios. Consider a simple linear regression model of the form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \; i = 1, 2, ..., n, \tag{2.1}$$

where Y is the response variable, X is the explanatory variable, ε is the random error term, and β_0 and β_1 are the regression coefficients. The regressor X is considered fixed throughout this chapter. In section 2.1, we assume independent normal errors with zero mean and constant variance. Based on this assumption we develop generalized pivotal quantities for the regression coefficients, dispersion parameter and expected response for a given value of X and then obtain their generalized confidence intervals. In section 2.2, errors are considered independently normally distributed with zero mean and varying dispersion, i.e. $\varepsilon_i \sim N(0, \sigma_i^2)$. In particular, we take $\sigma_i^2 = \sigma^2 X_i^2$, where σ is constant (Dougherty; 1992). We construct generalized pivotal quantities for β_0, β_1 and σ^2. The notion of regimes is introduced in section 2.3. Regimes can be different time periods, different regions etc. We consider data from two regimes.

We assume the regression coefficients remain the same for the two regimes. The dispersion in error terms are assumed the same within the regime but different between regimes. We propose generalized pivotal quantities for the regression parameters in such case. In section 2.4, we test the equality of the corresponding regression coefficients of the two regimes. When the regimes' error variances are different, it becomes a Behrens-Fisher problem problem in regression setting. We propose generalized pivotal quantities for the difference of the corresponding regression coefficients. The generalized p-values for testing the equality of corresponding slopes and intercepts of two regimes are then given.

2.1 Error terms are normal random variables with zero mean and constant variance

Suppose the error terms are iid normal with zero mean and constant variance σ^2. The maximum likelihood estimators for β_0, β_1 and σ^2 are $b_0 = \bar{Y} - b_1\bar{X}$, $b_1 = \frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2}$ and $S_Y^2 = \frac{\sum(Y_i-\hat{Y}_i)^2}{n}$, respectively, where $\bar{X} = \frac{\sum X_i}{n}$, $\bar{Y} = \frac{\sum Y_i}{n}$ and $\hat{Y}_i = b_0 + b_1X_i$. Interval estimation will be based on these maximum likelihood estimators. The estimator $b = (b_0, b_1)'$ follows a bivariate normal distribution as

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i-\bar{X})^2} & \frac{-\bar{X}}{\sum(X_i-\bar{X})^2} \\ \frac{-\bar{X}}{\sum(X_i-\bar{X})^2} & \frac{1}{\sum(X_i-\bar{X})^2} \end{pmatrix}\right),$$

and S_Y^2 follows a chi-square distribution as

$$\frac{nS_Y^2}{\sigma^2} \sim \chi^2_{n-2} \text{ (see Appendix A.1).}$$

Also, b_0 and b_1 are independent of S_Y^2 (see Appendix A.2).

2.1.1 Generalized confidence interval (GCI) for β_1

Denote generalized pivotal quantity (GPQ) for β_1 by R_{β_1}. We define

$$\begin{aligned} R_{\beta_1} &= b_1 - \frac{b_1-\beta_1}{\frac{\sigma}{\sqrt{\sum (X_i-\bar{X})^2}}} \times \frac{\sigma}{\sqrt{\sum (X_i-\bar{X})^2}} \times \frac{s_y}{S_Y} \\ &= b_1 - \frac{b_1-\beta_1}{\frac{\sigma}{\sqrt{\sum (X_i-\bar{X})^2}}} \times \frac{1}{\sqrt{\frac{S_Y^2}{\sigma^2}}} \times \frac{s_y}{\sqrt{\sum (X_i-\bar{X})^2}}, \end{aligned}$$

where s_y is the observed value of S_Y. Here, $\frac{b_1-\beta_1}{\frac{\sigma}{\sqrt{\sum (X_i-\bar{X})^2}}} \sim N(0,1)$ and $\frac{nS_Y^2}{\sigma^2} \sim \chi^2_{n-2}$. Further, they are independent as b_1 and S_Y^2 are independent.

Therefore,

$$R_{\beta_1} = b_1 - T_{n-2} \times \frac{(\sqrt{n/(n-2)})s_y}{\sqrt{\sum (X_i-\bar{X})^2}}, \tag{2.2}$$

where T_{n-2} has a t-distribution with $n-2$ degrees of freedom. From equation (2.2), the distribution of R_{β_1} is independent of any unknown parameters. Also, the observed value of R_{β_1} is β_1.

To construct a $100(1-\gamma)$ percent GCI for β_1, we set

$$Pr(R_{\beta_1} < c) = 1 - \gamma/2.$$

This implies that

$$Pr\left(b_1 - T_{n-2}.\frac{(\sqrt{n/(n-2)})s_y}{\sqrt{\sum (X_i-\bar{X})^2}} < c\right) = 1-\gamma/2.$$

Therefore,

$$Pr\left(T_{n-2} > \frac{\sqrt{\sum (X_i-\bar{X})^2}}{(\sqrt{n/(n-2)})s_y}.(b_1-c)\right) = 1-\gamma/2.$$

Suppose $t_{n-2,\gamma/2}$ is the $100(1-\gamma/2)^{th}$ percentile of T_{n-2}, then we get

$$\frac{\sqrt{\sum (X_i-\bar{X})^2}}{(\sqrt{n/(n-2)})s_y}.(b_1-c) = -t_{n-2,\gamma/2}.$$

Hence,

$$c = b_1 + t_{n-2,\gamma/2}.\frac{\left(\sqrt{n/(n-2)}\right) s_y}{\sqrt{\sum (X_i - \bar{X})^2}}.$$

Again, let

$$Pr(R_{\beta_1} < d) = \gamma/2.$$

This implies that

$$Pr\left(b_1 - T_{n-2}.\frac{(\sqrt{n/(n-2)})s_y}{\sqrt{\sum (X_i - \bar{X})^2}} < d\right) = \gamma/2,$$

or,

$$Pr\left(T_{n-2} > \frac{\sqrt{\sum (X_i - \bar{X})^2}}{(\sqrt{n/(n-2)})s_y}.(b_1 - d)\right) = \gamma/2.$$

From the t-table we have $Pr(T_{n-2} > t_{n-2,\gamma/2}) = \gamma/2$. Then,

$$\frac{\sqrt{\sum (X_i - \bar{X})^2}}{(\sqrt{n/(n-2)})s_y}.(b_1 - d)) = t_{n-2,\gamma/2}.$$

Therefore,

$$d = b_1 - t_{n-2,\gamma/2}.\frac{(\sqrt{n/(n-2)})s_y}{\sqrt{\sum (X_i - \bar{X})^2}}.$$

Thus, the $100(1-\gamma)$ percent GCI for β_1 is

$$[d,\ c] = \left[b_1 - t_{n-2,\gamma/2}.\frac{(\sqrt{n/(n-2)})s_y}{\sqrt{\sum (X_i - \bar{X})^2}},\ b_1 + t_{n-2,\gamma/2}.\frac{(\sqrt{n/(n-2)})s_y}{\sqrt{\sum (X_i - \bar{X})^2}}\right].$$

2.1.2 GCI for β_0

The GPQ for β_0 is

$$R_{\beta_0} = b_0 - \frac{b_0 - \beta_0}{\sigma\sqrt{1/n + \bar{X}^2/\sum (X_i - \bar{X})^2}} \times \sigma\sqrt{1/n + \bar{X}^2/\sum (X_i - \bar{X})^2} \times \frac{s_y}{S_Y}, \quad (2.3)$$

then,

$$R_{\beta_0} = b_0 - \frac{Z}{\sqrt{(S_Y^2/\sigma^2)}} \times s_y \sqrt{(1/n + \bar{X}^2/\sum(X_i - \bar{X})^2)} \qquad \text{with } Z \sim N(0,1).$$

Hence,

$$R_{\beta_0} = b_0 - T_{n-2}\sqrt{(n/(n-2))s_y^2(1/n + \bar{X}^2/\sum(X_i - \bar{X})^2)},$$

whose distribution is independent of any unknown parameters. Again, from (2.3) $R_{\beta_0} = \beta_0$ for observed value (X, s_y^2). As in the case of generalized estimation of β_1, a similar set of operations and inverse operations gives $100(1-\gamma)$ percent GCI for β_0 :

$$\left[b_0 - t_{n-2,\gamma/2} \times S(b_0), \quad b_0 + t_{n-2,\gamma/2} \times S(b_0)\right],$$

where,

$$S(b_0) = \sqrt{(n/(n-2))\, s_y^2 \left(1/n + \bar{X}^2/\sum(X_i - \bar{X})^2\right)}\ .$$

2.1.3 GCI for dispersion parameter

The GPQ for σ^2 is

$$R_{\sigma^2} = \frac{\sigma^2}{S_Y^2} \times s_y^2 = \frac{s_y^2}{S_Y^2/\sigma^2} = \frac{ns_y^2}{nS_Y^2/\sigma^2} = \frac{ns_y^2}{\chi_{n-2}^2}. \tag{2.4}$$

To construct the $100(1-\gamma)$ percent GCI for σ^2 let

$$Pr(R_{\sigma^2} < c) = 1 - \gamma/2.$$

Then,

$$Pr\left(\frac{ns_y^2}{\chi_{n-2}^2} < c\right) = 1 - \gamma/2,$$

or,

$$Pr\left(\chi_{n-2}^2 > \frac{ns_y^2}{c}\right) = 1 - \gamma/2.$$

Suppose $\chi_{n-2,\gamma/2}^2$ is the $100\gamma/2^{th}$ percentile of χ_{n-2}^2. It implies that

$$\frac{ns_y^2}{c} = \chi_{n-2,\gamma/2}^2,$$

or,

$$c = \frac{ns_y^2}{\chi^2_{n-2,\gamma/2}}.$$

Again $Pr(R_{\sigma^2} < d) = \gamma/2$ gives

$$\frac{ns_y^2}{d} = \chi^2_{n-2,1-\gamma/2}.$$

Therefore,

$$d = \frac{ns_y^2}{\chi^2_{n-2,1-\gamma/2}}.$$

Thus, the $100(1-\gamma)$ percent GCI for σ^2 is

$$[d,\ c] = \left[\frac{ns_y^2}{\chi^2_{n-2,1-\gamma/2}},\ \frac{ns_y^2}{\chi^2_{n-2,\gamma/2}}\right].$$

2.1.4 GCI for expected response for given X

Let μ_X be the expected response for a given X, i.e.,

$$\mu_X = E[Y \mid X] = \beta_0 + \beta_1 X.$$

Therefore, an estimate of μ_X is

$$\hat{\mu}_X = b_0 + b_1 X.$$

Since $\hat{\mu}_X$ is a linear combination of two jointly normal random variables b_0 and b_1, $\hat{\mu}_X$ is also normally distributed. Now,

$$E[\hat{\mu}_X] = E[b_0 + b_1 X] = \beta_0 + \beta_1 X = \mu_X.$$

$$V[\hat{\mu}_X] = V[b_0 + b_1 X] = V[\bar{Y} - b_1\bar{X} + b_1 X] = V[\bar{Y} + b_1(X - \bar{X})].$$

One can verify that

$$Cov[\bar{Y}, b_1] = E\left(\bar{\varepsilon} \times \frac{\sum(X_i - \bar{X})\varepsilon_i}{\sum(X_i - \bar{X})^2}\right) = \frac{\sum(X_i - \bar{X})\sigma^2}{n\sum(X_i - \bar{X})^2} = 0.$$

Then,

$$V[\hat{\mu}_X] = \frac{\sigma^2}{n} + (X - \bar{X})^2 \frac{\sigma^2}{\sum (X_i - \bar{X})^2} = \sigma^2 \left(\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right).$$

Therefore ,

$$\hat{\mu}_X \sim N \left(\mu_X, \sigma^2 \left(\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right) \right).$$

The GPQ for μ_X is

$$\begin{aligned} R_{\mu_X} &= \hat{\mu}_X - \frac{\hat{\mu}_X - \mu_X}{\sigma \sqrt{\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum (X_i - \bar{X})^2}}} \times \sigma \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum (X_i - \bar{X})^2}} \times \frac{s_y}{S_Y}, \qquad (2.5) \\ &= \hat{\mu}_X - \frac{Z}{\sqrt{S_Y^2/\sigma^2}} \times s_y \sqrt{1/n + (X - \bar{X})^2 / \sum (X_i - \bar{X})^2}, \\ &= \hat{\mu}_X - T_{n-2} \sqrt{(n/(n-2)) s_y^2 (1/n + (X - \bar{X})^2 / \sum (X_i - \bar{X})^2)}, \end{aligned}$$

which gives the $100(1-\gamma)$ percent GCI for μ_X as

$$\left[\hat{\mu}_X - t_{n-2,\gamma/2} \sqrt{\hat{V}(\hat{\mu}_X)},\ \hat{\mu}_X + t_{n-2,\gamma/2} \sqrt{\hat{V}(\hat{\mu}_X)}\ \right],$$

where $t_{n-2,\gamma/2}$ is the $100(1-\gamma/2)^{th}$ percentile of T_{n-2} and

$$\hat{V}(\hat{\mu}_X) = (n/(n-2))\, s_y^2 \left(1/n + (X - \bar{X})^2 / \sum (X_i - \bar{X})^2 \right).$$

2.2 Error distribution is normal with zero mean and varying dispersion

Consider the model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{where} \quad i = 1, 2, \dots, n.$$

We assume independent $\varepsilon_i \sim N(0, \sigma_i^2)$, where σ_i^2 are not necessarily equal. Heterogeneous error variance is often observed in practice (Gujarati; 1995). In matrix notation, we can express the model as

$$Y = X\beta + \varepsilon,$$

where

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

We assume that $\varepsilon \sim N_n(0, V)$, where

$$V = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}.$$

Often it is observed that variability increases as X increases. If we assume $\sigma_i^2 = \sigma^2 X_i^2$, where σ^2 is a constant, we can make the following transformation

$$\frac{Y_i}{X_i} = \frac{\beta_0}{X_i} + \beta_1 + \frac{\varepsilon_i}{X_i}$$

or,

$$\acute{Y_i} = \beta_0 \acute{X_i} + \beta_1 + \acute{\varepsilon_i}.$$

We may rewrite the above expression as

$$\acute{Y_i} = \beta_1 + \beta_0 \acute{X_i} + \acute{\varepsilon_i}. \tag{2.6}$$

Now, $\acute{\varepsilon_i} \sim N(0, \sigma^2)$. Thus homoscedasticity is maintained in the model (2.6) and it becomes the usual estimation problem in simple linear regression with constant error variance and non-stochastic X. For notational simplicity consider $\acute{Y_i}$ as Y_i, $\acute{X_i}$ as X_i and $\acute{\varepsilon_i}$ as ε_i. The maximum likelihood estimators for β_0, β_1 and σ^2 are

$$b_0 = \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2},\ b_1 = \bar{Y} - b_0 \bar{X} \text{ and } S_Y^2 = \frac{\sum (Y_i - \hat{Y_i})^2}{n},$$

where $\bar{X} = \sum X_i / n$, $\bar{Y} = \sum Y_i / n$ and $\hat{Y_i} = b_1 + b_0 X_i$. These estimators are distributed as

$$\begin{pmatrix} b_1 \\ b_0 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \beta_1 \\ \beta_0 \end{pmatrix}, \ \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} & \frac{-\bar{X}}{\sum (X_i - \bar{X})^2} \\ \frac{-\bar{X}}{\sum (X_i - \bar{X})^2} & \frac{1}{\sum (X_i - \bar{X})^2} \end{pmatrix} \right)$$

and

$$\frac{nS_Y^2}{\sigma^2} \sim \chi^2_{n-2}.$$

The GPQs for β_0, β_1 and σ^2 are obtained as before:

$$R_{\beta_0} = b_0 - \frac{b_0 - \beta_0}{\frac{\sigma}{\sqrt{\sum (X_i - \bar{X})^2}}} \times \frac{\sigma}{\sqrt{\sum (X_i - \bar{X})^2}} \times \frac{s_y}{S_Y}, \tag{2.7}$$

$$R_{\beta_1} = b_1 - \frac{b_1 - \beta_1}{\sigma\sqrt{1/n + \bar{X}^2/\sum (X_i - \bar{X})^2}} \times \sigma\sqrt{1/n + \bar{X}^2/\sum (X_i - \bar{X})^2} \times \frac{s_y}{S_Y}, \tag{2.8}$$

$$R_{\sigma^2} = \frac{\sigma^2}{S_Y^2} \times s_y^2, \tag{2.9}$$

where s_y^2 is observed value of S_Y^2.

2.3 Heteroscedasticity in two different regimes

Often data are collected in two different regimes, for example, the pre-depression period and the depression period. The dispersion in error terms remains the same within the regime but varies between regimes. Consider the model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i; \quad i = 1, 2, \ldots, n,$$

where ε_i is normally distributed with zero mean and

$$Var(\varepsilon_i) = \begin{cases} \sigma_1^2 \quad for \quad i = 1, 2, \ldots, n_1, \\ \sigma_2^2 \quad for \quad i = n_1 + 1, n_1 + 2, \ldots, n. \end{cases} \tag{2.10}$$

In matrix notation

$$Y = X\beta + \varepsilon,$$

here $\varepsilon \sim N_n(\underset{\sim}{0}, V)$, where

$$V = \begin{pmatrix} \sigma_1^2 I_{n_1} & \underset{\sim}{0} \\ \underset{\sim}{0} & \sigma_2^2 I_{n-n_1} \end{pmatrix}.$$

2.3.1 GCIs for β_0 and β_1 when σ_1^2 and σ_2^2 are known

When σ_1^2 and σ_2^2 are known, the estimation problem is straightforward. The generalized least square estimates of regression parameters are obtained by

$$b = (b_0(X,Y), b_1(X,Y))' = (X'V^{-1}X)^{-1}X'V^{-1}Y,$$

which follows a bivariate normal distribution as

$$b \sim N_2\left(\beta, (X'V^{-1}X)^{-1}\right).$$

Suppose

$$(X'V^{-1}X)^{-1} = \begin{pmatrix} g_1(X,\sigma_1^2,\sigma_2^2) & g_{12}(X,\sigma_1^2,\sigma_2^2) \\ g_{12}(X,\sigma_1^2,\sigma_2^2) & g_2(X,\sigma_1^2,\sigma_2^2) \end{pmatrix},$$

where $g_1(X,\sigma_1^2,\sigma_2^2) = Var(b_0)$, $g_2(X,\sigma_1^2,\sigma_2^2) = Var(b_1)$ and $g_{12}(X,\sigma_1^2,\sigma_2^2) = Cov(b_0,b_1)$.

The GPQ for β_0 and β_1 are respectively:

$$\begin{aligned} R_{\beta_0} &= b_0(X,y) - \frac{b_0(X,Y)-\beta_0}{(g_1(X,\sigma_1^2,\sigma_2^2))^{1/2}} \times \left(g_1(X,\sigma_1^2,\sigma_2^2)\right)^{1/2} \\ &= b_0(X,y) - Z \times \left(g_1(X,\sigma_1^2,\sigma_2^2)\right)^{1/2} \end{aligned} \tag{2.11}$$

and

$$\begin{aligned} R_{\beta_1} &= b_1(X,y) - \frac{b_1(X,Y)-\beta_1}{(g_2(X,\sigma_1^2,\sigma_2^2))^{1/2}} \times \left(g_2(X,\sigma_1^2,\sigma_2^2)\right)^{1/2} \\ &= b_1(X,y) - Z \times \left(g_2(X,\sigma_1^2,\sigma_2^2)\right)^{1/2}, \end{aligned} \tag{2.12}$$

where Z is $N(0,1)$. Let $n_2 = n - n_1$ and introduce j such that $i = n_1+1, n_1+2, ..., n$ is the same as $j = 1, 2, ..., n_2$. Then g_1 and g_2 can be expressed as

$$g_1(X,\sigma_1^2,\sigma_2^2) = \frac{\frac{\sum {X_i}^2}{\sigma_1^2} + \frac{\sum {X_j}^2}{\sigma_2^2}}{(\frac{n_1}{\sigma_1^2} + \frac{n_2}{\sigma_2^2})(\frac{\sum {X_i}^2}{\sigma_1^2} + \frac{\sum {X_j}^2}{\sigma_2^2}) - (\frac{n_1\bar{X}_1}{\sigma_1^2} + \frac{n_2\bar{X}_2}{\sigma_2^2})^2},$$

$$g_2(X,\sigma_1^2,\sigma_2^2) = \frac{\frac{n_1}{\sigma_1^2} + \frac{n_2}{\sigma_2^2}}{(\frac{n_1}{\sigma_1^2} + \frac{n_2}{\sigma_2^2})(\frac{\sum {X_i}^2}{\sigma_1^2} + \frac{\sum {X_j}^2}{\sigma_2^2}) - (\frac{n_1\bar{X}_1}{\sigma_1^2} + \frac{n_2\bar{X}_2}{\sigma_2^2})^2},$$

where $\bar{X}_1 = \sum X_i/n_1$ and $\bar{X}_2 = \sum X_j/n_2$. The generalized $(1-\alpha)\%$ confidence intervals for β_0 and β_1 are, respectively

$$\left[b_0(X,y) - Z_{\alpha/2} \times \left(g_1(X,\sigma_1^2,\sigma_2^2)\right)^{1/2},\ b_0(X,y) + Z_{\alpha/2} \times \left(g_1(X,\sigma_1^2,\sigma_2^2)\right)^{1/2}\right]$$

and

$$\left[b_1(X,y) - Z_{\alpha/2} \times \left(g_2(X,\sigma_1^2,\sigma_2^2)\right)^{1/2},\ b_1(X,y) + Z_{\alpha/2} \times \left(g_2(X,\sigma_1^2,\sigma_2^2)\right)^{1/2}\right].$$

2.3.2 GCIs for β_0 and β_1 when σ_1^2 and σ_2^2 are unknown

When σ_1^2 and σ_2^2 are unknown, we propose weighted estimators for the regression parameters. First, independent estimates of regression parameters for the two regimes are obtained using the least square method and then weighted average of the estimates give the proposed estimates for the combined sample.

Suppose $b_0^{(i)}$, $b_1^{(i)}$ are least square estimates of β_0 and β_1 respectively and $\hat{\sigma}_i^2 = S_i^2$ is the error mean square for regime i, where $i = 1, 2$. The proposed weighted estimator of β_0 is

$$b_0 = w_1 b_0^{(1)} + (1-w_1) b_0^{(2)},$$

where w_1 is the weight for regime 1 and it is determined such that $Var(b_0)$ is minimum. Now,

$$Var(b_0) = w_1^2 Var\left(b_0^{(1)}\right) + (1-w_1)^2 Var\left(b_0^{(2)}\right) = w_1^2\sigma_1^2 f(X_1) + (1-w_1)^2\sigma_2^2 f(X_2)$$

where

$$f(X) = \frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2},$$

$X_1 = (X_1, X_2, \ldots, X_{n_1})$ and $X_2 = (X_{n_1+1}, X_{n_1+2}, \ldots, X_n)$. Differentiating $Var(b_0)$ with respect to w_1 and then equating it to 0 we get

$$2w_1\sigma_1^2 f(X_1) - 2(1-w_1)\sigma_2^2 f(X_2) = 0,$$

which gives

$$w_1 = \frac{\sigma_2^2 f(X_2)}{\sigma_1^2 f(X_1) + \sigma_2^2 f(X_2)}.$$

The estimate is given by

$$\hat{w}_1 = \frac{s_2^2 f(X_2)}{s_1^2 f(X_1) + s_2^2 f(X_2)},$$

where s_1^2 and s_2^2 are the observed value of the error mean squares $S_1^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n_1 - 2}$ and $S_2^2 = \frac{\sum (Y_j - \hat{Y}_j)^2}{n_2 - 2}$, respectively.

Therefore,

$$b_0 = \hat{w}_1 b_0^{(1)} + (1 - \hat{w}_1) b_0^{(2)}.$$

Using (2.3) the GPQ for β_0 for regime i is

$$R_{\beta_0^{(i)}} = b_0^{(i)} - \frac{b_0^{(i)} - \beta_0}{\sigma_i \sqrt{f(X_i)}} \times \sigma_i \sqrt{f(X_i)} \times \frac{s_i}{S_i} = b_0^{(i)} - T_{n_i - 2} \times s_i \sqrt{f(X_i)}. \qquad (2.13)$$

Therefore, the GPQ for β_0 for combined sample is

$$R_{\beta_0} = \hat{w}_1 R_{\beta_0^{(1)}} + (1 - \hat{w}_1) R_{\beta_0^{(2)}}. \qquad (2.14)$$

For observed y of Y, R_{β_0} is equal to β_0 and its distribution is independent of any unknown parameters. The $100\alpha/2^{th}$ and $100(1-\alpha/2)^{th}$ percentiles of R_{β_0} form a $100(1-\alpha)$ percent GCI for it. The percentiles can be obtained through simulation.

Similarly, the GPQ for β_1 is

$$R_{\beta_1} = \hat{\delta}_1 R_{\beta_1^{(1)}} + (1 - \hat{\delta}_1) R_{\beta_1^{(2)}}, \qquad (2.15)$$

where $R_{\beta_1^{(i)}}$ is the GPQ for β_1 for regime i, $i = 1, 2$. Now, $R_{\beta_1^{(i)}}$ is defined as

$$R_{\beta_1^{(i)}} = b_1^{(i)} - \frac{b_1^{(i)} - \beta_1}{\sigma_i \sqrt{f^*(X_i)}} \times \sigma_i \sqrt{f^*(X_i)} \times \frac{s_i}{S_i} = b_1^{(i)} - T_{n_i - 2} \times s_i \sqrt{f^*(X_i)}, \qquad (2.16)$$

where

$$f^*(X) = \frac{1}{\sum (X_i - \bar{X})^2}$$

and $\hat{\delta}_1$ is the estimated weight which is defined as

$$\hat{\delta}_1 = \frac{s_2^2 f^*(X_2)}{s_1^2 f^*(X_1) + s_2^2 f^*(X_2)}.$$

The GCI for β_1 can then be obtained by getting the percentiles of R_{β_1} through simulation.

The GPQ for σ_i^2 is

$$R_{\sigma_i^2} = \frac{\sigma_i^2}{S_i^2} \times s_i^2 = \frac{(n_i - 2)s_i^2}{\chi^2_{n_i-2}}. \tag{2.17}$$

2.4 Testing equality of corresponding regression coefficients under heteroscedasticity

Consider the regression model

$$\begin{cases} Y_i = \beta_{10} + \beta_{11} X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n_1, \\ Y_j = \beta_{20} + \beta_{21} X_j + \varepsilon_j, \quad j = 1, 2, \ldots, n_2, \end{cases} \tag{2.18}$$

where $n_2 = n - n_1$ with ε_i $\underline{\underline{iid}}$ $N(0, \sigma_1^2)$ and ε_j $\underline{\underline{iid}}$ $N(0, \sigma_2^2)$. Also, ε_i and ε_j are independent.

We would like to test the hypotheses

$$H_{01} : \beta_{10} = \beta_{20} \text{ against } H_{11} : \beta_{10} \neq \beta_{20} \text{ and}$$

$$H_{02} : \beta_{11} = \beta_{21} \text{ against } H_{12} : \beta_{11} \neq \beta_{21}.$$

To test the hypotheses, we propose GPQs for the difference of the corresponding regression coefficients of two regimes. Our proposed GPQs can be used efficiently in testing the hypotheses irrespective of the sample sizes of the regimes. We will illustrate the construction of the GPQ for $\beta_{10} - \beta_{20}$. The GPQ for $\beta_{11} - \beta_{21}$ can be obtained in a similar fashion.

We know that

$$b_{10}(X_1, Y_1) \sim N\left(\beta_{10}, \sigma_1^2 f(X_1)\right), \quad b_{20}(X_2, Y_2) \sim N\left(\beta_{20}, \sigma_2^2 f(X_2)\right),$$

where,

$$f(X) = \frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2},$$

$Y_1 = (Y_1, Y_2, \ldots, Y_{n_1})$, $Y_2 = (Y_{n_1+1}, Y_{n_1+2}, \ldots, Y_n)$, $X_1 = (X_1, X_2, \ldots, X_{n_1})$ and $X_2 = (X_{n_1+1}, X_{n_1+2}, \ldots, X_n)$.

Again, (see Appendix A.1)

$$\frac{(n_1 - 2)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-2} \quad \text{and} \quad \frac{(n_2 - 2)S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-2},$$

where $S_1^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n_1-2}$ and $S_2^2 = \frac{\sum(Y_j - \hat{Y}_j)^2}{n_2-2}$ are the error mean squares for regimes 1 and 2, respectively. Further, these four random variables b_{10}, b_{20}, S_1^2 and S_2^2 are independent of each other, since b_{i0} and S_i^2, $i = 1, 2$, are independent of each other (Appendix A.2) and the samples from the two regimes are independent.

Therefore,

$$b_{10}(X_1, Y_1) - b_{20}(X_2, Y_2) \sim N\left(\beta_{10} - \beta_{20}, \sigma_1^2 f(X_1) + \sigma_2^2 f(X_2)\right).$$

The proposed GPQ for $\beta_{10} - \beta_{20}$ is

$$R_{\beta_{10}-\beta_{20}} = b_{10}(X_1, y_1) - b_{20}(X_2, y_2) - \left(\frac{b_{10}(X_1, Y_1) - b_{20}(X_2, Y_2) - (\beta_{10} - \beta_{20})}{(\sigma_1^2 f(X_1) + \sigma_2^2 f(X_2))^{1/2}}\right)$$

$$\times \left(\sigma_1^2 f(X_1)\frac{s_1^2}{S_1^2} + \sigma_2^2 f(X_2)\frac{s_2^2}{S_2^2}\right)^{1/2}. \tag{2.19}$$

Then,

$$R_{\beta_{10}-\beta_{20}} = b_{10}(X_1, y_1) - b_{20}(X_2, y_2) - Z\left(\frac{s_1^2 f(X_1)}{S_1^2/\sigma_1^2} + \frac{s_2^2 f(X_2)}{S_2^2/\sigma_2^2}\right)^{1/2}$$

with $Z \sim N(0,1)$.

Further, we have

$$\begin{aligned} R_{\beta_{10}-\beta_{20}} &= b_{10}(X_1,y_1) - b_{20}(X_2,y_2) - Z\left(\frac{(n_1-2)s_1^2 f(X_1)}{\chi^2_{n_1-2}} + \frac{(n_2-2)s_2^2 f(X_2)}{\chi^2_{n_2-2}}\right)^{1/2} \\ &= b_{10}(X_1,y_1) - b_{20}(X_2,y_2) - \frac{Z}{\left(\frac{\chi^2_{n_1-2}+\chi^2_{n_2-2}}{n_1+n_2-4}\right)^{1/2}} \times \left(\frac{\chi^2_{n_1-2}+\chi^2_{n_2-2}}{n_1+n_2-4}\right)^{1/2} \\ &\quad \times \left(\frac{(n_1-2)s_1^2 f(X_1)}{\chi^2_{n_1-2}} + \frac{(n_2-2)s_2^2 f(X_2)}{\chi^2_{n_2-2}}\right)^{1/2}, \end{aligned}$$

and finally,

$$\begin{aligned} R_{\beta_{10}-\beta_{20}} &= b_{10}(X_1,y_1) - b_{20}(X_2,y_2) - T_{n_1+n_2-4} \\ &\times \left(\frac{1}{n_1+n_2-4}\left(\frac{(n_1-2)s_1^2 f(X_1)}{B} + \frac{(n_2-2)s_2^2 f(X_2)}{1-B}\right)\right)^{1/2}, \end{aligned} \quad (2.20)$$

where

$$B = \frac{\chi^2_{n_1-2}}{\chi^2_{n_1-2}+\chi^2_{n_2-2}} \sim Beta(\frac{n_1-2}{2}, \frac{n_2-2}{2}).$$

We have seen from (2.19) for observed sample $R_{\beta_{10}-\beta_{20}}$ is equal to $\beta_{10}-\beta_{20}$. We also observe that its distribution is independent of any unknown parameters.

Similarly, the proposed GPQ for $\beta_{11}-\beta_{21}$ is

$$R_{\beta_{11}-\beta_{21}} = b_{11}(X_1,y_1) - b_{21}(X_2,y_2) - \left(\frac{b_{11}(X_1,Y_1) - b_{21}(X_2,Y_2) - (\beta_{11}-\beta_{21})}{(\sigma_1^2 f^*(X_1) + \sigma_2^2 f^*(X_2))^{1/2}}\right)$$

$$\times \left(\sigma_1^2 f^*(X_1)\frac{s_1^2}{S_1^2} + \sigma_2^2 f^*(X_2)\frac{s_2^2}{S_2^2}\right)^{1/2}, \quad (2.21)$$

where

$$f^*(X) = \frac{1}{\sum (X_i - \bar{X})^2}.$$

Then,

$$\begin{aligned} R_{\beta_{11}-\beta_{21}} &= b_{11}(X_1,y_1) - b_{21}(X_2,y_2) - T_{n_1+n_2-4} \\ &\times \left(\frac{1}{n_1+n_2-4}\left(\frac{(n_1-2)s_1^2 f^*(X_1)}{B} + \frac{(n_2-2)s_2^2 f^*(X_2)}{1-B}\right)\right)^{1/2} \end{aligned} \quad (2.22)$$

The percentiles of $R_{\beta_{10}-\beta_{20}}$ and $R_{\beta_{11}-\beta_{21}}$ would give the GCIs for $\beta_{10}-\beta_{20}$ and $\beta_{11}-\beta_{21}$ respectively. The percentiles can be obtained using simulation.

The generalized p-values for testing the equality of intercepts and slopes are respectively

$$P_{\beta_{10}-\beta_{20}} = 2\ \min\left\{P(R_{\beta_{10}-\beta_{20}} \geq 0),\ P(R_{\beta_{10}-\beta_{20}} \leq 0)\right\} \tag{2.23}$$

and

$$P_{\beta_{11}-\beta_{21}} = 2\ \min\left\{P(R_{\beta_{11}-\beta_{21}} \geq 0),\ P(R_{\beta_{11}-\beta_{21}} \leq 0)\right\}. \tag{2.24}$$

Chapter 3

Generalized Inference for Multiple Linear Regression Parameters

In this chapter we make generalized inference on parameters of a multiple linear regression model for a fixed set of values of the explanatory variables. Different assumptions are made about the error distribution. Consider the multiple linear regression model

$$Y_j = \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_p X_{pj} + \varepsilon_j,\ j = 1, 2, \ldots, n. \tag{3.1}$$

Here Y_j is the response variable for the j^{th} set of values of $(X_1, X_2, \ldots, X_p)$ and ε_j is the corresponding random error term. Also, $\beta_i,\ i = 1, 2, \ldots, p$ represent a total of p unknown parameters to be estimated. Intercept can be included in the model by defining $X_{1j} = 1$ for all j. In section 3.1, we assume iid normal errors with zero mean and constant variance. Then, we construct the generalized confidence region for the regression parameters' vector and generalized confidence interval for the regression coefficients. We also construct generalized confidence interval for the dispersion parameter. In section 3.2, heteroscedasticity in error terms is considered. We assume that variability in error terms is due to the measurement errors in one particular explanatory variable. Multivariate data from two different regimes with heteroscedasticity are considered in section 3.3. As in the simple linear regression case, we assume that the regression coefficients remain the same for the two regimes and that the error variance is stationary within the regime but different between regimes. In such case, we propose generalized confidence regions and confidence intervals for the regression parameters' vector and regression coefficients, respectively. In section 3.4, we test the equality of the sets regression coefficients of two regimes under heteroscedasticity. Chow (1960) proposed a test to do this task in the homoscedastic case. Later, Toyoda (1974) extended the Chow test for heteroscedastic regimes. He showed that the test is appropriate when the regimes' error variances are equal. Under heteroscedasticity the test works well if at least one regime has large sample size. If both regimes have small sample sizes, its level of significance is affected greatly even for moderate heteroscedasticity. For this multivariate Behrens-Fisher problem we propose a generalized test variable that can be used efficiently in testing the equality of the sets regression coefficients of the two regimes irrespective of their sample sizes. The generalized p-value for this test is given. We construct the generalized confidence region for the difference of the two sets of regression coefficients and then, the generalized confidence intervals for elements of that vector. Note that in the definition of GTV, given in section 1.1, the third property refers to the case where $\theta \in \mathbb{R}$. For the case where $\theta \in \mathbb{R}^n$, the concept of monotonocity needs some clarifications since the concept of order is not clearly defined in $\mathbb{R}^n$. Let, the norm of a vector $x \in \mathbb{R}^n$ is $\| x \|_A^2 = x'Ax$, where A is a positive definite matrix. We consider that the vector x is less than the vector $y \in \mathbb{R}^n$ $(x < y)$ if

$$\| x \|_A < \| y \|_A,$$

for any positive definite matrix A. Thus, a real valued function over $\mathbb{R}^n$, say, $f(x)$, $x \in \mathbb{R}^n$ is considered to be non-decreasing if for all $x_1 < x_2 \in \mathbb{R}^n$ we have

$$f(x_1) \leq f(x_2).$$

3.1 Error terms are normal random variables with zero mean and constant variance

Suppose error terms $\varepsilon_j,\ j = 1, 2, \ldots, n$ are iid normal with zero mean and constant variance σ^2. Defining

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix},\ \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},\ X = \begin{pmatrix} X_{11} & X_{21} & \ldots & X_{p1} \\ X_{12} & X_{22} & \ldots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & \ldots & X_{pn} \end{pmatrix} \text{ and } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

the model in (3.1) can be written in matrix notation as

$$Y = X\beta + \varepsilon, \tag{3.2}$$

where X has full column rank.

3.1.1 Generalized confidence region (GCR) for β

Given model (3.2), the least squares estimator for β is

$$\hat{\beta} = b = (X'X)^{-1}X'Y, \text{ where } b = (b_1, b_2, \ldots, b_p)',$$

and given the normality assumption on ε,

$$b \sim N_p\left(\beta, \sigma^2(X'X)^{-1}\right)$$

and by the result in Appendix A.1,

$$\frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p},$$

where

$$S^2 = (Y - \hat{Y})'(Y - \hat{Y})/(n-p) = \sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2/(n-p) = \text{Error MS}.$$

Let us define

$$T = \left(s^2(X'X)^{-1}\right)^{-1/2} b,$$

$$\Lambda = \left(s^2(X'X)^{-1}\right)^{-1/2} \sigma^2(X'X)^{-1} \left(s^2(X'X)^{-1}\right)^{-1/2},$$

$$\theta = \left(s^2(X'X)^{-1}\right)^{-1/2} \beta,$$

where s^2 is the observed value of S^2.

Therefore,

$$T \sim N_p(\theta, \Lambda).$$

It implies that

$$U = (T-\theta)'\Lambda^{-1}(T-\theta) \sim \chi^2_p.$$

One can verify that

$$U = \frac{1}{\sigma^2}(b-\beta)'(X'X)(b-\beta).$$

Further, let

$$V = \frac{S^2}{\sigma^2 s^2}.$$

One can verify that

$$V \sim \frac{1}{(n-p)s^2}\chi^2_{n-p} .$$

Again, U and V are independent of each other (see Appendix A2). Therefore,

$$F = \frac{U/p}{V} = \frac{\frac{1}{\sigma^2}(b-\beta)'(X'X)(b-\beta)/p}{\frac{S^2}{\sigma^2 s^2}} = s^2 F_{p,\ n-p}, \tag{3.3}$$

where $F_{p,\ n-p}$ has Fishers's distribution with $(p, n-p)$ degrees of freedom. Note that, for given Y and β the distribution of F is free of the nuisance parameter σ^2. Also, under $H_0 : \beta = \beta_0$, the observed value of F is

$$f_0 = \frac{1}{p}(b-\beta_0)'(X'X)(b-\beta_0) \tag{3.4}$$

that does not depend on σ^2. Further, the observed value of F is

$$f = \frac{1}{p}(b-\beta)'(X'X)(b-\beta),$$

a positive definite quadratic form in $(b-\beta)$. For given y, f is a non-decreasing function of β. It implies that $P(F \geq f)$ is stochastically non-increasing in β. Therefore, F is the generalized test variable for testing $H_0 : \beta = \beta_0$. The generalized p-value is

$$P(F \geq f_0).$$

Now, this generalized test variable F can be used to obtain a GCR for β. In fact, in multiparameter problems, generalized pivotal quantities are difficult or impossible to obtain. Instead, the distribution of a generalized test variable can be used to derive a generalized confidence region (Gamage et al; 2004). Let $F_{p,\ n-p,\ 1-\alpha}$ is the $100(1-\alpha)^{th}$ percentile of Fisher's distribution with $(p, n-p)$ degrees of freedom. Then

$$P(F \leq s^2 F_{p,\ n-p,\ 1-\alpha}) = 1-\alpha.$$

The $100(1-\alpha)$ percent generalized confidence region (GCR) for the elements of β is represented by the set of values of the vector b which satisfy the following inequality

$$(b-\beta)'(X'X)(b-\beta) \leq ps^2 F_{p,\ n-p,\ 1-\alpha}. \tag{3.5}$$

3.1.2 Generalized confidence interval (GCI) for β_i

Suppose,

$$\sigma^2(X'X)^{-1} = \sigma^2 \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1p} \\ d_{21} & d_{22} & \dots & d_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ d_{p1} & d_{p2} & \dots & d_{pp} \end{pmatrix}.$$

Denote the generalized pivotal quantity for β_i by R_{β_i}. Then

$$R_{\beta_i} = b_i - \frac{b_i - \beta_i}{\sigma\sqrt{d_{ii}}} \times \sigma\sqrt{d_{ii}} \times \frac{s_y}{S_Y}, \tag{3.6}$$

where s_y^2 is the observed value of

$$S_Y^2 = \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 / n.$$

Note that $\frac{nS_Y^2}{\sigma^2} \sim \chi^2_{n-p}$. One can verify that

$$R_{\beta_i} = b_i - T_{n-p} \sqrt{\frac{n}{n-p} s_y^2 \; d_{ii}} \; , \tag{3.7}$$

where T_{n-p} has a t-distribution with $n-p$ degrees of freedom. We would like to construct generalized confidence intervals for β_i' s. There are p of them. If we construct $100(1-\alpha)$ percent generalized confidence intervals for each of them, the probability that the p intervals will simultaneously be correct is at least $(1 - p\alpha)$. If p is large, the set of generalized confidence intervals becomes relatively uninformative. There are several approaches to maintain the overall confidence level to be at least $(1-\alpha)$. Among them Bonferroni approach(Alt; 1982), Scheffe method (Scheffe; 1959) and Working-Hotelling approach (Working and Hotelling; 1929) are widely used. For simplicity we consider the Bonferroni technique of splitting α. We construct generalized confidence intervals for β_i' s each with confidence coefficient $(1 - \alpha/p)$. In this way we maintain the overall confidence level to be at least $(1 - \alpha)$. Therefore, the $100(1 - \alpha)$ percent joint generalized confidence intervals for β_i' s are

$$\left(b_i - t_{n-p,\ \alpha/_{2p}} \sqrt{\frac{n}{n-p} s_y^2 \; d_{ii}},\ b_i + t_{n-p,\ \alpha/_{2p}} \sqrt{\frac{n}{n-p} s_y^2 \; d_{ii}} \right),$$

where $t_{n-p,\ \alpha/_{2p}}$ is the $100(1-\alpha/_{2p})^{th}$ percentile of T_{n-p}. The limitation of the Bonferroni approach is that when p is large, it will give conservative results. In such case, the other methods of maintaining overall confidence level can be applied in generalized inference.

3.1.3 GCI for dispersion parameter

The GPQ for σ^2 is

$$R_{\sigma^2} = \frac{\sigma^2}{S_Y^2} \times s_y^2.$$

One can verify that

$$R_{\sigma^2} = \frac{ns_y^2}{\chi^2_{n-p}} \tag{3.8}$$

Therefore, a $100(1-\gamma)$ percent GCI for σ^2 is

$$\left(\frac{ns_y^2}{\chi^2_{n-p,1-\gamma/2}}, \; \frac{ns_y^2}{\chi^2_{n-p,\gamma/2}} \right),$$

where $\chi^2_{n-p,1-\gamma/2}$ is the $100(1-\gamma/2)^{th}$ percentile of χ^2_{n-p}.

3.2 Error distribution is normal with zero mean and varying dispersion

Consider the model

$$Y_j = \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_p X_{pj} + \varepsilon_j, \; j = 1, 2, \ldots, n.$$

We assume independent $\varepsilon_j \sim N(0, \sigma_j^2)$, where σ_j^2 are not necessarily equal. In matrix notation

$$Y = X\beta + \varepsilon.$$

Here, we assume that $\varepsilon \sim N_n(0, \; V)$, where

$$V = \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix}.$$

Often this heteroscedasticity in error variance is due to systematic errors in measurements of one or more explanatory variables. For instance, suppose $\sigma_j^2 = \sigma^2 X_{ij}^2$, where

σ^2 is a constant. We can make the following transformation to stabilize the error variance:

$$\frac{Y_j}{X_{ij}} = \beta_1 \frac{X_{1j}}{X_{ij}} + \beta_2 \frac{X_{2j}}{X_{ij}} + \ldots + \beta_i \frac{X_{ij}}{X_{ij}} + \ldots + \beta_p \frac{X_{pj}}{X_{ij}} + \frac{\varepsilon_j}{X_{ij}}.$$

Or,

$$Y'_j = \beta_1 X'_{1j} + \beta_2 X'_{2j} + \ldots + \beta_i X'_{ij} + \ldots + \beta_p X'_{pj} + \varepsilon'_j, \tag{3.9}$$

with $X'_{ij} = 1$ for $\forall j$ and $\varepsilon'_j \sim N(0,\ \sigma^2)$. Therefore, the model (3.9) becomes a multiple linear regression model with normal and homoscedastic error term and non-stochastic X. GCR for β and GCIs for β_i and σ^2 can be obtained as explained in section 3.1.

3.3 Heteroscedasticity in two different regimes

Consider the model

$$Y_j = \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_p X_{pj} + \varepsilon_j,\ j = 1, 2, \ldots, n,$$

where ε_j is normally distributed with zero mean and

$$Var(\varepsilon_j) = \begin{cases} \sigma_1^2 & for \quad j = 1, 2, \ldots, n_1 \\ \sigma_2^2 & for \quad j = n_1 + 1, n_1 + 2, \ldots, n \end{cases} \tag{3.10}$$

i.e. σ_1^2 and σ_2^2 are error variances in the two different regimes. In matrix notation

$$Y = X\beta + \varepsilon,$$

where $\varepsilon \sim N_n(\underset{\sim}{0}, V)$ and

$$V = \begin{pmatrix} \sigma_1^2 I_{n_1} & \underset{\sim}{0} \\ \underset{\sim}{0} & \sigma_2^2 I_{n-n_1} \end{pmatrix}.$$

3.3.1 GCR and GCI for regression coefficients when σ_1^2 and σ_2^2 are known

For the known variance case, the least squares estimators of regression parameters are

$$\hat{\beta} = b = (b_1, b_2, \ldots, b_p)' = (X'V^{-1}X)^{-1}X'V^{-1}Y,$$

and b follows a p-variate normal distribution:

$$b \sim N_p\left(\beta, (X'V^{-1}X)^{-1}\right).$$

Therefore,

$$(b-\beta)'(X'V^{-1}X)(b-\beta) \sim \chi^2_p.$$

A $100(1-\gamma)$ percent GCR for β is the set of values of b which satisfy the following inequality

$$(b-\beta)'(X'V^{-1}X)(b-\beta) \leq \chi^2_{p,\ 1-\gamma/2}, \tag{3.11}$$

where $\chi^2_{p,\ 1-\gamma/2}$ is the $100(1-\gamma/2)^{th}$ percentile of χ^2_p distribution. Further, suppose

$$(X'V^{-1}X)^{-1} = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1p} \\ g_{21} & g_{22} & \cdots & g_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ g_{p1} & g_{p2} & \cdots & g_{pp} \end{pmatrix},$$

where g_{kk} is the variance of b_k and $g_{kk'}$ is the covariance of b_k and $b_{k'}$, $k \neq k' = 1, 2, \ldots, p$. Then GPQ for β_k is

$$R_{\beta_k} = b_k - \frac{b_k - \beta_k}{\sqrt{g_{kk}}} \times \sqrt{g_{kk}} = b_k - Z\sqrt{g_{kk}}, \tag{3.12}$$

where $Z \sim N(0,1)$. Therefore, a $100(1-\gamma)$ percent joint generalized confidence intervals for β_k s are

$$\left(b_k - Z_{\gamma/2p}\sqrt{g_{kk}},\ b_k + Z_{\gamma/2p}\sqrt{g_{kk}}\ \right),$$

where $Z_{\gamma/2p}$ is the $100(1-\gamma/2p)^{th}$ percentile of Z.

3.3.2 GCR and GCI for regression coefficients when σ_1^2 and σ_2^2 are unknown

Let us partition the model $Y = X\beta + \varepsilon$ for the two regimes as

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\beta + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix},$$

where Y_1 is of order $(n_1 \times 1)$ representing responses for the first regime,

Y_2 is of order $(n - n_1 \times 1)$ representing responses for the second regime,

X_1 is the design matrix of order $(n_1 \times p)$ for the first regime,

X_2 is the design matrix of order $(n - n_1 \times p)$ for the second regime,

β is the $(p \times 1)$ vector of regression coefficients,

ε_1 is the $(n_1 \times 1)$ error vector for regime 1 and

ε_2 is the $(n - n_1 \times 1)$ error vector for regime 2.

Then, we have in matrix notation the model for the i^{th} regime

$$Y_i = X_i\beta + \varepsilon_i, \ i = 1, 2,$$

where $\varepsilon_i \sim N_{n_i}(\underset{\sim}{0}, \sigma_i^2 I_{n_i})$. The generalized test variable for the i^{th} regime is

$$F_i = \frac{\frac{1}{\sigma_i^2}(b_i - \beta)'(X'_iX_i)(b_i - \beta)/p}{\frac{S_i^2}{\sigma_i^2 s_i^2}} = s_i^2 F_{p,\ n_i - p}, \tag{3.13}$$

where s_i^2 is the observed value of S_i^2 = Error MS for regime i, $F_{p,\ n_i-p}$ has a F distribution with $(p,\ n_i - p)$ degrees of freedom and

$$\hat{\beta}_i = b_i = (b_{i1}, b_{i2}, \ldots, b_{ip})' = (X'_iX_i)^{-1}X'_iY_i, \ i = 1, 2.$$

Let us define

$$F = \eta_1 F_1 + (1 - \eta_1)F_2 = \eta_1 s_1^2 F_{p,n_1-p} + (1 - \eta_1)s_2^2 F_{p,n_2-p}.$$

η_1 is determined in such a way that $Var(F)$ is minimum. One can verify that

$$\hat{\eta_1} = \frac{s_2^4 V_2}{s_1^4 V_1 + s_2^4 V_2},$$

where

$$V_i = Var(F_{p,\ n_i-p}) = \frac{2(n_i - p)^2(n_i - 2)}{p(n_i - p - 2)^2(n_i - p - 4)} \text{ for } n_i - p > 4, \ i = 1, 2.$$

Let F_γ satisfy $Pr(F \leq F_\gamma) = 1 - \gamma$. We can obtain F_γ through simulation. Then $100(1 - \gamma)$ percent GCR for β is obtained by the set of values of the vectors b_1 and b_2 which satisfy the following inequality

$$\hat{\eta_1}(b_1 - \beta)'(X'_1X_1)(b_1 - \beta) + (1 - \hat{\eta_1})(b_2 - \beta)'(X'_2X_2)(b_2 - \beta) \leq pF_\gamma. \tag{3.14}$$

Further, suppose

$$Cov(b_i) = \sigma_i^2 (X'_i X_i)^{-1} = \sigma_i^2 \begin{pmatrix} d_{i11} & d_{i12} & \dots & d_{i1p} \\ d_{i21} & d_{i22} & \dots & d_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ d_{ip1} & d_{ip2} & \dots & d_{ipp} \end{pmatrix},$$

where $\sigma_i^2 d_{ikk}$ is the variance of the least square estimator of β_k and $\sigma_i^2 d_{ikk'}$ is the covariance between the least square estimators of β_k and $\beta_{k'}$, $k \neq k' = 1, 2, \dots, p$ for i^{th} regime. The generalized pivotal quantity for β_k is

$$R_{\beta_k} = \hat{w}_k R_{\beta_k^{(1)}} + (1 - \hat{w}_k) R_{\beta_k^{(2)}}, \tag{3.15}$$

where $R_{\beta_k^{(i)}}$ is the GPQ for β_k from the i^{th} regime and

$$R_{\beta_k^{(i)}} = b_{ik} - T_{n_i - p} \times s_i \sqrt{d_{ikk}}. \tag{3.16}$$

Here s_i^2 is the observed value of S_i^2, the error mean square for regime i and

$$\hat{w}_k = \frac{s_2^2 d_{2kk}}{s_1^2 d_{1kk} + s_2^2 d_{2kk}}$$

is the estimated weight factor.

3.4 Testing equality of corresponding regression coefficients under heteroscedasticity

In section 3.3 we consider the case where the regression coefficients remained the same for the two regimes, only the error variances were different between regimes. In this section, we like to test the equality of the regression coefficients of the two regimes under heteroscedasticity. Consider the regression models

$$\begin{cases} Y_{1j} = \beta_{11} X_{11j} + \beta_{12} X_{12j} + \dots + \beta_{1p} X_{1pj} + \varepsilon_{1j}, \ j = 1, 2, \dots, n_1, \\ Y_{2j} = \beta_{21} X_{21j} + \beta_{22} X_{22j} + \dots + \beta_{2p} X_{2pj} + \varepsilon_{2j}, \ j = 1, 2, \dots, n_2, \end{cases} \tag{3.17}$$

where $\varepsilon_{1j} \sim iid\ N(0, \sigma_1^2)$ and $\varepsilon_{2j} \sim iid\ N(0, \sigma_2^2)$. In matrix notation

$$\begin{cases} Y_1 = X_1\beta_1 + \varepsilon_1, \\ Y_2 = X_2\beta_2 + \varepsilon_2, \end{cases}$$

where $\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{ip})'$,

$\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{in_i})' \sim N_{n_i}(0, \sigma_i^2 I_{n_i})$,

ε_1 and ε_2 are independent,

$Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ and

X_i is the design matrix of order $(n_i \times p)$ for the i^{th} regime.

The least squares estimators of the regression parameters for the i^{th} regime is

$$\hat{\beta}_i = b_i = (b_{i1}, b_{i2}, \ldots, b_{ip})' = (X'_iX_i)^{-1}X'_iY_i.$$

An unbiased estimator for the error variance σ_i^2 is S_i^2, error mean square for regime i. Further,

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \sim N_{2p}\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2(X'_1X_1)^{-1} & 0 \\ 0 & \sigma_2^2(X'_2X_2)^{-1} \end{pmatrix}\right), \tag{3.18}$$

and

$$S_i^2(X'_iX_i)^{-1}$$

is the random matrix of order $(p \times p)$ for the i^{th} regime, where (see Appendix A.1)

$$\frac{(n_i - p)S_i^2}{\sigma_i^2} \sim \chi^2_{n_i - p}. \tag{3.19}$$

We would like to test

$$H_0 : \beta_1 = \beta_2 \text{ against } H_1 : \beta_1 \neq \beta_2.$$

Suppose the observed value of S_i^2 is s_i^2. Let us define

$$Z \;=\; \left[\sigma_1^2(X'_1X_1)^{-1} + \sigma_2^2(X'_2X_2)^{-1}\right]^{-1/2}\left[(b_1 - b_2) - (\beta_1 - \beta_2)\right],$$

and

$$U \;=\; \frac{s_1^2}{S_1^2} \times \sigma_1^2 ({X'}_1 X_1)^{-1} + \frac{s_2^2}{S_2^2} \times \sigma_2^2 ({X'}_2 X_2)^{-1}.$$

From (3.18),

$$Z \;\sim\; N_p(0,\; I).$$

The matrix U can be rewritten as

$$U = \frac{s_1^2}{S_1^2/\sigma_1^2} \times ({X'}_1 X_1)^{-1} + \frac{s_2^2}{S_2^2/\sigma_2^2} \times ({X'}_2 X_2)^{-1},$$

and then, from (3.19),

$$U = \frac{(n_1 - p)s_1^2}{\chi^2_{n_1-p}} \times ({X'}_1 X_1)^{-1} + \frac{(n_2 - p)s_2^2}{\chi^2_{n_2-p}} \times ({X'}_2 X_2)^{-1}.$$

Now, define

$$T_I = Z'UZ = Z' \left(\frac{(n_1 - p)s_1^2}{\chi^2_{n_1-p}} \times ({X'}_1 X_1)^{-1} + \frac{(n_2 - p)s_2^2}{\chi^2_{n_2-p}} \times ({X'}_2 X_2)^{-1} \right) Z. \quad (3.20)$$

Clearly the distribution of T_I is independent of any unknown parameters. Further, for given y, the observed value of T_I is

$$\begin{aligned} t_I &= [(b_1 - b_2) - (\beta_1 - \beta_2)]'\, u^{-1/2} u u^{-1/2} \, [(b_1 - b_2) - (\beta_1 - \beta_2)] \\ &= [(b_1 - b_2) - (\beta_1 - \beta_2)]' \; [(b_1 - b_2) - (\beta_1 - \beta_2)]\,, \end{aligned}$$

where

$$u = \left[\sigma_1^2 ({X'}_1 X_1)^{-1} + \sigma_2^2 ({X'}_2 X_2)^{-1}\right]$$

is the observed value of U. Under H_0 the observed value of T_I does not depend on any unknown parameters. Further, for given y, t_I is a positive definite quadratic form in $[(b_1 - b_2) - (\beta_1 - \beta_2)]$. Therefore, t_I is a non-decreasing function of $\beta_1 - \beta_2$. It implies that $P(T_I \geq t_I)$ is stochastically non-increasing in $\beta_1 - \beta_2$. Thus, T_I is a generalized test variable. The generalized p-value is

$$P(T_I \geq t_I \mid H_0). \quad (3.21)$$

Now, we will construct generalized confidence region for $\beta_1 - \beta_2$ based on the distribution of the generalized test variable T_I. Let $t_{I_{1-\gamma}}$ satisfy that

$$P(T_I \leq t_{I_{1-\gamma}}) = 1 - \gamma.$$

Then, a $100(1-\gamma)$ percent generalized confidence region for the elements of $\beta_1 - \beta_2$ is represented by the set of values of the vector $b_1 - b_2$ which satisfy the following inequality

$$[(b_1 - b_2) - (\beta_1 - \beta_2)]' \, [(b_1 - b_2) - (\beta_1 - \beta_2)] \leq t_{I_{1-\gamma}}. \qquad (3.22)$$

In testing the hypothesis

$$H_0 : \beta_{1k} = \beta_{2k} \text{ against } H_1 : \beta_{1k} \neq \beta_{2k}, \; k = 1, 2, \ldots, p,$$

we use the GPQ for $\beta_{1k} - \beta_{2k}$ as we did for the simple linear regression model. The GPQ for $\beta_{1k} - \beta_{2k}$, denoted by $R_{\beta_{1k}-\beta_{2k}}$, is simply an extension of GPQs given in (2.20) and (2.22). Now,

$$\begin{aligned} R_{\beta_{1k}-\beta_{2k}} &= b_{1k}(X_1, y_1) - b_{2k}(X_2, y_2) - T_{n_1+n_2-2p} \\ &\times \left(\frac{1}{n_1 + n_2 - 2p} \left(\frac{(n_1 - p)s_1^2 d_{1kk}}{B} + \frac{(n_2 - p)s_2^2 d_{2kk}}{1 - B} \right) \right)^{1/2} \end{aligned} \qquad (3.23)$$

where s_i^2 is the observed value of S_i^2 = error mean square for regime i, $S_i^2 d_{ikk}$ is an unbiased estimator of the variance of b_{ik} and

$$B = \frac{\chi^2_{n_1-p}}{\chi^2_{n_1-p} + \chi^2_{n_2-p}} \sim Beta\left(\frac{n_1 - p}{2}, \frac{n_2 - p}{2} \right).$$

A joint $100(1-\alpha)$ percent generalized confidence interval for $\beta_{1k} - \beta_{2k}, \; k = 1, 2, \ldots, p$, is obtained by computing individual generalized confidence interval for $\beta_{1k} - \beta_{2k}$ each with confidence coefficient $(1 - \alpha/p)$ through simulation.

Chapter 4

Simulation Study

Simulation was carried out to study the performance of generalized confidence intervals (GCIs) of linear regression coefficients and dispersion parameters and generalized tests (GTs) for comparing regression coefficients for small and moderate sample sizes 3, 5, 10, 14, 15, 20, 30 and 60. Independent variables X were considered non-stochastic but different assumptions about the error distribution were made. For a fixed sample size, 10,000 samples were generated and the GCIs with typical 95 percent confidence level were computed for each sample. The percentage of intervals that included the true parameter was then obtained. For GTs, generalized p-value was computed for each sample and then the proportion of rejecting null hypothesis was computed under null and alternative hypothesis. In section 4.1, results are presented for simple linear regression models and in section 4.2 results are given for the multivariate case. The programs were written in R to run the simulation.

4.1 Simulation result for simple linear regression model

Consider a simple linear regression model of the form

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \ i = 1, 2, ..., n$$

with X fixed and independent normal ε_i's.

For the classical case when $\varepsilon_i \sim N(0, \sigma^2)$, we generated the observations y's for fixed X's and different set of values of β_0, β_1 and σ^2. The values of X's are given in Table 4.1. The values of X were randomly selected from numbers 1 to 1000 and then those were fixed for the simulation study. Parameter values are presented in Table 3.2. GCIs for β_0, β_1 and σ^2 were computed using (2.3), (2.2) and (2.4), respectively, for each sample obtained from the simulation scheme. Table 4.2 gives the percentages of GCIs that include the true parameter. For each parameter, the observed confidence level for GCI is close to the nominal 95 percent level even for small sample of size 3.

In case of heteroscedasticity $\varepsilon_i \sim N(0, \sigma_i^2)$, we assumed $\sigma_i^2 = \sigma^2 X_i^2$. We used the same set of values of β_0, β_1 and σ^2 for simulation as in the earlier case. The GCIs for β_0, β_1 and σ^2 were obtained using (2.7), (2.8) and (2.9) respectively. Observed confidence levels are presented in Table 4.3 for each parameter. As in the previous case, the observed confidence level for each parameter is close to the nominal 95 percent level.

In situations when the data were collected over two different regimes and error variance remained stationary within each regime but varied between regimes, we fixed the error variance for the first regime to be $\sigma_1^2 = 4$ and for the second regime to be $\sigma_2^2 = 9$. We also assumed that regime change did not affect the regression parameters and we kept their values at $\beta_0 = 5$ and $\beta_1 = 2$. Then GCIs of the regression parameters were computed when:

i. σ_1^2 and σ_2^2 were assumed known,

ii. σ_1^2 and σ_2^2 were assumed unknown and estimated by the error mean square of the respective regime.

When variances were unknown, the GCIs of the regression parameters were obtained using the proposed weighted generalized pivotal quantities (GPQs) given in (2.14) and (2.15). Then the observed confidence levels of GCIs were calculated.

Tables 4.4 and 4.5 give the simulation results for the two cases where n_1 and n_2

are the sample sizes of regime 1 and 2 respectively. In the known variance case, the observed confidence levels of GCIs of the regression coefficients are close to the nominal 95 percent level even when the regime sample sizes are small (Table 4.4). Sample size difference of regimes has no impact in observed confidence levels.

When the regimes' error variances are unknown, we observed that empirical confidence levels are just below the typical 95 percent for overall sample size $n = n_1 + n_2 \leq 20$ (Table 4.5). When n is 14, the observed confidence level of GCIs of the slope parameter is 94 percent for both equal and unequal sample sizes. Increments in sample sizes ($n > 20$) improve the confidence levels close to the nominal level.

In testing the hypothesis of equality of regression coefficients in the two regimes, we proposed the generalized pivotal quantities of their differences in (2.19) and (2.21). The null and alternative hypotheses are

$$H_{01} : \beta_{10} = \beta_{20} \text{ against } H_{11} : \beta_{10} \neq \beta_{20} \text{ and}$$

$$H_{02} : \beta_{11} = \beta_{21} \text{ against } H_{12} : \beta_{11} \neq \beta_{21}.$$

For this simulation study, we set $\beta_{10} = \beta_{20} = 5$ and $\beta_{11} = \beta_{21} = 2$. We estimated the proposed generalized confidence intervals (GCIs) as well as the classical confidence intervals (CI) for $\beta_{10} - \beta_{20}$ and $\beta_{11} - \beta_{21}$ for each sample obtained from the simulation scheme in homoscedastic and heteroscedastic settings. In the homoscedastic case, the classical confidence intervals were obtained using the exact t distributions each with $n_1 + n_2 - 4$ degrees of freedom. In the heteroscedastic case, the confidence intervals were obtained using approximate t distributions (Schechtman and Sherman; 2007). The approximate t statistic for comparing the slope parameters is

$$t = \frac{b_{11}(X_1, Y_1) - b_{21}(X_2, Y_2) - (\beta_{11} - \beta_{21})}{\left(S_1^2 f^*(X_1) + S_2^2 f^*(X_2)\right)^{1/2}},$$

where S_1^2 and S_2^2 are the error mean squares of the two regimes and

$$f^*(X) = \frac{1}{\sum (X_i - \bar{X})^2}.$$

The degrees freedom of this t statistic is obtained by applying the Satterthwaite's approximation (Satterthwaite; 1941) which is

$$r = \frac{\left(S_1^2 f^*(X_1) + S_2^2 f^*(X_2)\right)^2}{\frac{\left(S_1^2 f^*(X_1)\right)^2}{n_1-2} + \frac{\left(S_2^2 f^*(X_2)\right)^2}{n_2-2}}.$$

Similarly, the approximate t statistic can be obtained for comparing the intercepts of the two regimes. Then, we computed the percentages of times we rejected the null hypothesis for the generalized and classical tests and compare these values with the nominal value of 5 percent. In the case of homoscedasticity, we set $\sigma_1^2 = \sigma_2^2 = 4$. For heteroscedasticity, we considered two situations:

i. moderate heteroscedasticity ($\sigma_1^2 = 4$ and $\sigma_2^2 = 9$) and

ii. severe heteroscedasticity ($\sigma_1^2 = 4$ and $\sigma_2^2 = 25$).

Tables 4.6 and 4.7 give the simulation results. In the case of homoscedasticity, the test is found to be conservative for small samples ($n \leq 14$). As sample size increases, the test attains the nominal significance level of 0.05 for $n \geq 20$. In contrast, the classical test preserves the nominal level for all sample sizes considered. No impact on the significance levels was observed due to the sample size differences of regimes.

In the case of moderate heteroscedasticity, we observed that when the regimes' sample sizes are equal, the significance levels of the proposed generalized tests are close to the nominal 0.05 level for $n \geq 14$. For $n = 14$ ($n_1 = 7$ and $n_2 = 7$), the observed significance level of the test is 0.04 for comparing the slopes of the two regimes. When the regimes' sample sizes are unequal, the test is found to be conservative if overall sample size is small (≤ 14) and there exists large difference in regimes' sample sizes. When regimes' sample sizes are 4 and 10, the observed significance level of the test is 0.02 for comparing the slope. As sample size increases, the test preserves the significance levels close to the nominal level for $n \geq 20$. On the other hand, in the case of moderate heteroscedasticity, the classical test based on the equal variance assumption gives significance level close to the nominal level when the regimes' sample sizes are equal. When regimes' sample sizes are unequal, the classical test is either too conservative or too liberal. The classical test is found to be liberal when the large sample has smaller error variance than that of the small sample and is found to be conservative in the opposite case. Even when overall sample size is large (≥ 30), the classical test is conservative if the regimes' sample sizes differ significantly unlike the proposed test.

In the severe heteroscedastic case, the similar results are obtained for the generalized test as in the case of moderate heteroscedasticity. But the classical test is found to be conservative or liberal in most of the cases of the severe heteroscedasticity, except when the overall sample size is large ($n \geq 30$) and the regimes' sample sizes are approximately equal. In the heteroscedastic cases, when there is difference in regimes' sample sizes, the generalized test is comparatively better than the classical test in terms of empirical significance level.

Table 4.8 gives the observed significance levels of the proposed generalized test and the test based on approximate t distributions in testing equality of the regression coefficients in the two regimes in the moderate and severe heteroscedastic cases. We observed that the approximate t test preserves the nominal 5 percent significance level for all sample sizes considered in the heteroscedastic cases.

The powers of the generalized test and the classical test in testing equality of the slope parameters of the two regimes in the heteroscedastic cases are presented in Table 4.9. Figures 4.1 and 4.2 show the power curves of the two tests in testing equality of the slope parameters of the two regimes with severe heteroscedasticity ($\sigma_1^2 = 4$, $\sigma_2^2 = 25$) for different sample sizes, respectively. We observed that the generalized test gives higher power for overall sample size $n \geq 14$. In particular, Figure 4.3 illustrates the behaviour of the power of both tests in the severe heteroscedastic case for $n = 20$ and it is clear that the proposed test performs better than the classical test. In the case of moderate heteroscedasticity, both tests yield similar power for $n \geq 20$.

The size and power of the proposed test and the test based on approximate t in testing equality of the slope parameters in heteroscedastic cases (from moderate to severe) are presented in Table 4.10 for $n_1 = 15$ and $n_2 = 5$. Here, fixed X values considered in simulation are 10, 10.5, 11,..., 17 for regime 1 and 10, 10.5, 11,..., 12 for regime 2. We observed that the size and the power of the generalized test are comparable with that of the approximate t test. The observed level of significance (size) of the generalized test is close to the nominal 5 percent level for the moderate and the severe heteroscedastic cases. When the slope difference is 8, the power of the generalized test is 1 for the moderate heteroscedasticity ($\sigma_1^2 = 1, \sigma_2^2 = 2$) and it is 0.98 for the severe heteroscedasticity ($\sigma_1^2 = 1, \sigma_2^2 = 4$). In the extreme case when $\sigma_1^2 = 1, \sigma_2^2 = 8$, the power is 0.838 that is also quite satisfactory.

Table 4.1: Fixed X values for different sample sizes

Sample sizes							
3	5	10	14	20	30	60	
278	195	304	478	616	753	177	932
99	751	838	896	304	637	767	756
735	209	747	657	519	850	621	887
	701	535	71	756	509	984	904
	127	280	975	357	919	185	514
		711	915	224	855	664	605
		426	961	791	852	921	259
		140	257	294	395	168	758
		235	544	876	720	528	926
		270	189	703	315	800	776
			668	582	542	400	858
			512	811	768	288	429
			668	5	121	516	744
			866	511	871	815	330
				459	883	201	435
				496	557	66	298
				219	295	296	40
				64	649	839	321
				207	778	386	255
				854	924	802	41
					483	222	311
					298	537	126
					193	850	122
					568	266	645
					498	357	747
					871	250	147
					55	17	93
					378	62	511
					417	134	883
					860	300	495

Table 4.2: Empirical confidence levels for 95% generalized confidence intervals of the classical linear regression parameters

				Observed confidence level for		
β_0	β_1	σ^2	Sample size	β_0	β_1	σ^2
5	2	4	3	0.951	0.950	0.950
			5	0.951	0.953	0.953
			14	0.953	0.954	0.946
			30	0.950	0.951	0.950
5	0.5	4	3	0.952	0.953	0.951
			5	0.952	0.952	0.950
			14	0.949	0.950	0.948
			30	0.951	0.953	0.950
5	0.5	0.25	3	0.949	0.948	0.949
			5	0.949	0.953	0.948
			14	0.950	0.950	0.950
			30	0.952	0.950	0.949
10	-2	0.0025	3	0.949	0.947	0.949
			5	0.949	0.949	0.952
			14	0.947	0.947	0.947
			30	0.948	0.948	0.950

Table 4.3: Observed confidence levels for 95% generalized confidence intervals of the classical linear regression parameters in the case of heteroscedasticity

				Observed confidence level for		
β_0	β_1	σ^2	Sample size	β_0	β_1	σ^2
5	2	4	3	0.945	0.946	0.947
			5	0.948	0.950	0.947
			14	0.950	0.954	0.947
			30	0.949	0.951	0.952
5	0.5	4	3	0.952	0.952	0.948
			5	0.952	0.951	0.947
			14	0.947	0.948	0.948
			30	0.950	0.950	0.953
5	0.5	0.25	3	0.952	0.950	0.951
			5	0.951	0.953	0.953
			14	0.950	0.951	0.952
			30	0.947	0.952	0.944
10	-2	0.0025	3	0.947	0.949	0.944
			5	0.947	0.946	0.948
			14	0.953	0.950	0.951
			30	0.951	0.952	0.946

Table 4.4: Observed confidence levels for 95% generalized confidence intervals of the simple linear regression coefficients when error variances $\sigma_1^2 = 4$ and $\sigma_2^2 = 9$ of the two regimes are assumed known

		Observed confidence level for	
n_1	n_2	β_0	β_1
5	5	0.951	0.954
7	7	0.951	0.952
10	10	0.952	0.951
15	15	0.953	0.956
30	30	0.948	0.950
6	4	0.953	0.953
4	10	0.950	0.948
8	12	0.951	0.949
15	5	0.951	0.950
5	15	0.949	0.948
13	17	0.949	0.949
26	34	0.948	0.950
12	48	0.950	0.951

Table 4.5: Observed confidence levels for 95% generalized confidence intervals of the simple linear regression coefficients when error variances $\sigma_1^2 = 4$ and $\sigma_2^2 = 9$ of the two regimes are assumed unknown

		Observed confidence level for	
n_1	n_2	β_0	β_1
5	5	0.923	0.927
7	7	0.933	0.939
10	10	0.939	0.939
15	15	0.943	0.947
30	30	0.945	0.946
6	4	0.919	0.919
4	10	0.942	0.941
8	12	0.940	0.935
15	5	0.931	0.933
5	15	0.934	0.932
13	17	0.940	0.940
26	34	0.944	0.945
12	48	0.943	0.944

Table 4.6: Observed significance levels of the generalized and classical test in testing the equality of the regression coefficients in homoscedastic case ($\sigma_1^2 = \sigma_2^2 = 4$) at 0.05 level of significance

		Observed significance level for			
		Generalized test		Classical test	
n_1	n_2	$\beta_{10} = \beta_{20}$	$\beta_{11} = \beta_{21}$	$\beta_{10} = \beta_{20}$	$\beta_{11} = \beta_{21}$
5	5	0.027	0.022	0.048	0.046
7	7	0.031	0.031	0.052	0.053
10	10	0.040	0.035	0.054	0.054
15	15	0.040	0.046	0.049	0.049
30	30	0.054	0.050	0.049	0.047
6	4	0.018	0.030	0.052	0.050
4	10	0.022	0.025	0.052	0.052
8	12	0.037	0.038	0.052	0.053
15	5	0.034	0.034	0.052	0.053
5	15	0.041	0.040	0.048	0.047
13	17	0.041	0.041	0.048	0.050
26	34	0.043	0.045	0.054	0.052
12	48	0.049	0.045	0.050	0.052

Table 4.7: Observed significance levels of the generalized and classical tests in testing the equality of the regression coefficients in heteroscedastic cases at 0.05 level of significance

				Observed significance level for			
				Generalized test		Classical test	
σ_1^2	σ_2^2	n_1	n_2	$\beta_{10} = \beta_{20}$	$\beta_{11} = \beta_{21}$	$\beta_{10} = \beta_{20}$	$\beta_{11} = \beta_{21}$
4	9	5	5	0.017	0.019	0.047	0.056
		7	7	0.036	0.040	0.040	0.055
		10	10	0.037	0.038	0.038	0.043
		15	15	0.043	0.045	0.041	0.053
		30	30	0.046	0.045	0.050	0.050
		6	4	0.023	0.038	0.073	0.099
		4	10	0.019	0.019	0.030	0.025
		8	12	0.032	0.036	0.027	0.027
		15	5	0.038	0.041	0.096	0.110
		5	15	0.036	0.043	0.012	0.013
		13	17	0.041	0.044	0.038	0.037
		26	34	0.042	0.041	0.034	0.034
		12	48	0.045	0.042	0.012	0.013
4	25	5	5	0.026	0.033	0.056	0.076
		7	7	0.043	0.045	0.071	0.086
		10	10	0.039	0.045	0.032	0.039
		15	15	0.050	0.047	0.032	0.048
		30	30	0.051	0.046	0.053	0.052
		6	4	0.036	0.050	0.122	0.152
		4	10	0.027	0.023	0.021	0.017
		8	12	0.033	0.033	0.024	0.015
		15	5	0.042	0.039	0.178	0.199
		5	15	0.028	0.032	0.003	0.002
		13	17	0.044	0.043	0.029	0.033
		26	34	0.051	0.047	0.022	0.021
		12	48	0.055	0.048	0.002	0.003

Table 4.8: Observed significance levels of the generalized test and the test based on approximate t distribution in testing the equality of the regression coefficients in heteroscedastic cases at 0.05 level of significance

				Observed significance level for			
				Generalized test		Test based on approx. t	
σ_1^2	σ_2^2	n_1	n_2	$\beta_{10} = \beta_{20}$	$\beta_{11} = \beta_{21}$	$\beta_{10} = \beta_{20}$	$\beta_{11} = \beta_{21}$
4	9	5	5	0.017	0.019	0.043	0.043
		7	7	0.036	0.040	0.044	0.048
		10	10	0.037	0.038	0.052	0.051
		15	15	0.043	0.045	0.054	0.053
		30	30	0.046	0.045	0.047	0.050
		6	4	0.023	0.038	0.055	0.066
		4	10	0.019	0.019	0.049	0.052
		8	12	0.032	0.036	0.046	0.049
		15	5	0.038	0.041	0.059	0.061
		5	15	0.036	0.043	0.060	0.056
		13	17	0.041	0.044	0.051	0.053
		26	34	0.042	0.041	0.049	0.053
		12	48	0.045	0.042	0.051	0.049
4	25	5	5	0.026	0.033	0.048	0.054
		7	7	0.043	0.045	0.051	0.053
		10	10	0.039	0.045	0.049	0.050
		15	15	0.050	0.047	0.049	0.047
		30	30	0.051	0.046	0.048	0.053
		6	4	0.036	0.050	0.063	0.055
		4	10	0.027	0.023	0.049	0.050
		8	12	0.033	0.033	0.051	0.048
		15	5	0.042	0.039	0.064	0.064
		5	15	0.028	0.032	0.051	0.052
		13	17	0.044	0.043	0.049	0.044
		26	34	0.051	0.047	0.048	0.052
		12	48	0.055	0.048	0.041	0.049

Table 4.9: Power of the tests in testing the equality of the slope parameters ($\beta_{11} = \beta_{21} = 2$) in the two regimes with heteroscedasticity at 0.05 level of significance

					Slope difference ($\beta_{11} - \beta_{21}$)						
Test	σ_1^2	σ_2^2	n_1	n_2	-0.05	-0.03	-0.025	0	0.025	0.03	0.05
GCI	4	9	6	4	0.448	0.214	0.156	0.040	0.168	0.208	0.454
			4	10	1	0.950	0.892	0.028	0.905	0.962	1
			8	12	1	1	0.998	0.036	0.999	1	1
			13	17	1	1	1	0.046	1	1	1
			26	34	1	1	1	0.052	1	1	1
			12	48	1	1	1	0.048	1	1	1
	4	25	6	4	0.253	0.111	0.095	0.040	0.088	0.115	0.200
			4	10	1	0.890	0.763	0.020	0.764	0.893	0.999
			8	12	1	0.997	0.950	0.030	0.962	0.990	1
			13	17	1	0.999	0.990	0.045	0.994	1	1
			26	34	1	1	1	0.056	1	1	1
			12	48	1	1	1	0.051	1	1	1
CI	4	9	6	4	0.872	0.531	0.453	0.095	0.379	0.503	0.891
			4	10	1	0.999	0.986	0.019	0.985	1	1
			8	12	1	1	0.997	0.028	0.996	1	1
			13	17	1	1	1	0.037	1	1	1
			26	34	1	1	1	0.040	1	1	1
			12	48	1	1	1	0.016	1	1	1
	4	25	6	4	0.643	0.351	0.325	0.181	0.307	0.363	0.670
			4	10	1	0.888	0.727	0.110	0.722	0.873	1
			8	12	1	0.979	0.871	0.018	0.872	0.965	1
			13	17	1	1	0.989	0.030	0.989	1	1
			26	34	1	1	1	0.028	1	1	1
			12	48	1	1	1	0.003	1	1	1

Table 4.10: Size and power of the generalized test and the test based on approximate t in testing the equality of the slope parameters ($\beta_{11} - \beta_{21} = 0$) in the two regimes of sample sizes $n_1 = 15$ and $n_2 = 5$ in heteroscedastic cases at 0.05 level of significance

Slope difference	σ_1^2	σ_2^2	Generalized test	Test based on approx. t
0	1	2	0.054	0.054
	1	4	0.058	0.057
	1	8	0.046	0.053
2	1	2	0.343	0.372
	1	4	0.195	0.218
	1	8	0.139	0.129
4	1	2	0.821	0.838
	1	4	0.585	0.581
	1	8	0.315	0.313
6	1	2	0.989	0.986
	1	4	0.877	0.861
	1	8	0.606	0.600
8	1	2	1	1
	1	4	0.981	0.979
	1	8	0.838	0.831

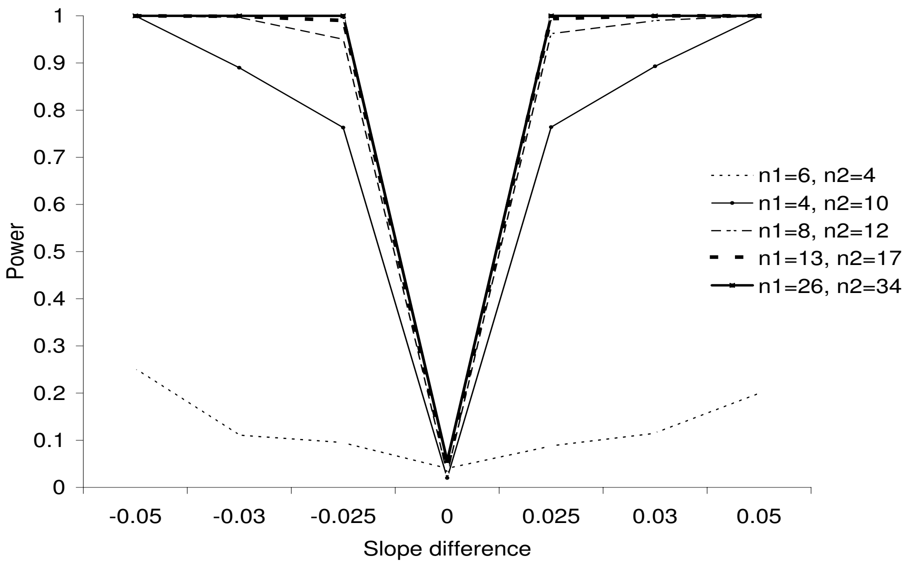

Figure 4.1: Power of the generalized test at 0.05 level of significance in testing the equality of the slope parameters ($\beta_{11} = \beta_{12}$) in the two regimes with severe heteroscedasticity $\sigma_1^2 = 4$ and $\sigma_2^2 = 25$ for different sample sizes

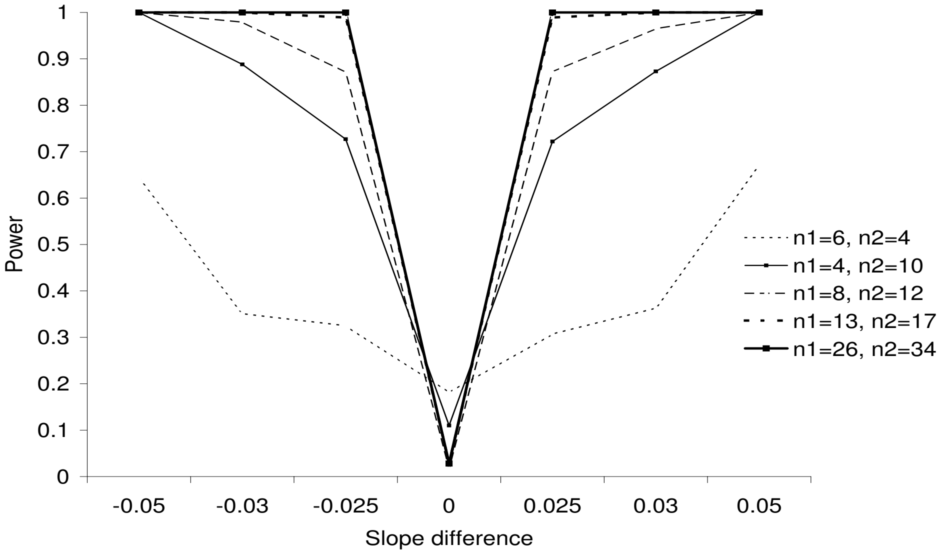

Figure 4.2: Power of the classical test based on the equal variance assumption in testing the equality of the slope parameters ($\beta_{11} = \beta_{12}$) in the two regimes with severe heteroscedasticity $\sigma_1^2 = 4$ and $\sigma_2^2 = 25$ for different sample sizes at 0.05 level of significance

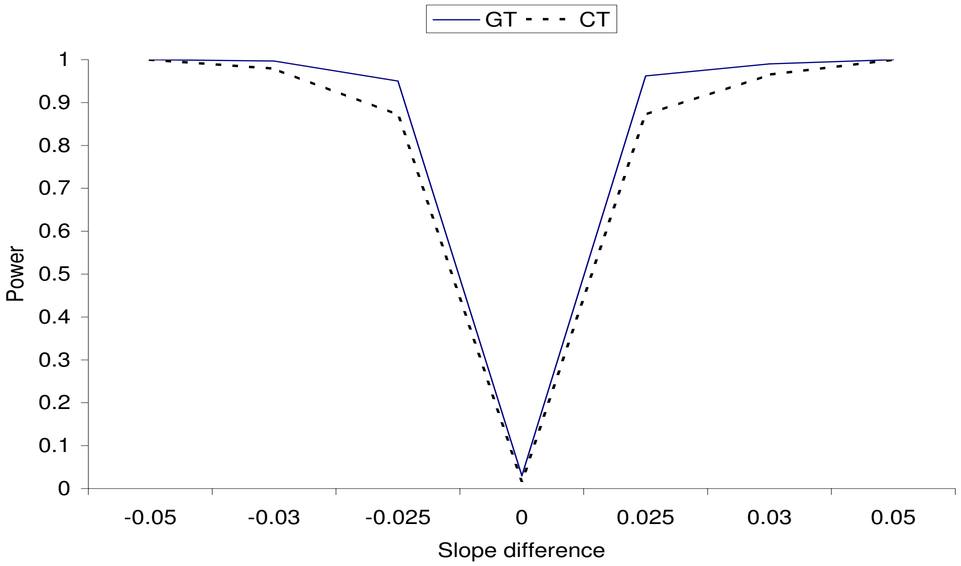

Figure 4.3: Power of the generalized and classical tests in testing the equality of the slope parameters ($\beta_{11} = \beta_{12}$) in the two regimes with severe heteroscedasticity $\sigma_1^2 = 4$ and $\sigma_2^2 = 25$ for overall sample size 20 ($n_1 = 8$ and $n_2 = 12$) at 0.05 level of significance

4.2 Simulation result for multiple linear regression model

Consider the regression models of two regimes

$$\begin{cases} Y_{1j} = \beta_{10} + \beta_{11}X_{11j} + \beta_{12}X_{12j} + \varepsilon_{1j}, \ j = 1, 2, \ldots, n_1, \\ Y_{2j} = \beta_{20} + \beta_{21}X_{21j} + \beta_{22}X_{22j} + \varepsilon_{2j}, \ j = 1, 2, \ldots, n_2, \end{cases} \tag{4.1}$$

where $\varepsilon_{1j} \sim iid\ N(0, \sigma_1^2)$ and $\varepsilon_{2j} \sim iid\ N(0, \sigma_2^2)$. In matrix notation

$$\begin{cases} Y_1 = X_1\beta_1 + \varepsilon_1, \\ Y_2 = X_2\beta_2 + \varepsilon_2, \end{cases}$$

where $\beta_i = (\beta_{i0}, \beta_{i1}, \beta_{i2})'$,

$\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{in_i})' \sim N_{n_i}(0, \sigma_i^2 I_{n_i})$,

ε_1 and ε_2 are independent,

$Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{i_{n_i}})'$ and

X_i is the design matrix of order $(n_i \times 3)$ for the i^{th} regime.

To test the hypothesis

$$H_0 : \beta_1 = \beta_2 \text{ against } H_1 : \beta_1 \neq \beta_2,$$

we proposed a generalized test variable in (3.20). Simulation was done to obtain the empirical significance level and the power of the test at 0.05 level of significance for regimes' sample sizes 5, 10 and 15. The explanatory variables were considered fixed. The values of X_{i1} were obtained from a sequence starting from 1 and then increased by 5 at every step. The values of X_{i2} were randomly sampled from 1000 to 10000 and then they were fixed for the simulation. In table 4.11, these X values are given. For heteroscedasticity, three situations were considered:

i. moderate heteroscedasticity ($\sigma_1^2 = 1$ and $\sigma_2^2 = 2$),

ii. severe heteroscedasticity ($\sigma_1^2 = 1$ and $\sigma_2^2 = 4$) and

iii. extreme heteroscedasticity ($\sigma_1^2 = 1$ and $\sigma_2^2 = 8$).

Observations y's were generated using (4.1) for $\beta_{i0} = 5$, $\beta_{i1} = 0.5$ and $\beta_{i2} = 1$. For a fixed sample size, 10,000 samples were generated and for each sample the generalized p-value for the test was computed using (3.21). Then proportion of rejecting the null hypothesis was calculated. This gives the empirical significance level for the generalized test. The power of the test was obtained by assigning the difference between corresponding regression coefficients $\beta_{1k} - \beta_{2k} \neq 0$, k=1,2,3 following the same procedure.

Empirical significance levels of the proposed generalized test at 0.05 level are given in Table 4.12 for the different sample sizes and the heteroscedastic cases. When the regimes' sample sizes are 15 and 5, the observed significance level of the test is 0.045 in the case of moderate heteroscedasticity. In the cases of the severe and extreme heteroscedasticity, the observed levels are 0.048 and 0.041, respectively. When regimes' sample sizes are 15 and 10, the empirical significance levels of the test are 0.052, 0.049 and 0.047 for moderate to extreme heteroscedastic cases, respectively. For equal sample case ($n_1 = n_2 = 15$), the similar results are observed. Thus for small samples and in the heteroscedastic cases the test preserves the significance level very close to the nominal 0.05 level.

In Table 4.13, the power of the generalized test is given for regimes' sample sizes 15 and 5 and in the heteroscedastic cases at 0.05 level of significance. We observed that when the difference between the sets of regression coefficients is $\beta_1 - \beta_2 = (20, 10, 15)'$, the power of the test is 0.91 in the moderate heteroscedastic case. When the difference is increased to $(30, 15, 20)'$, the power of the test is 0.996 in the moderate heteroscedastic case and it is 0.915 in the severe heteroscedastic case. In the case of extreme heteroscedasticity, a large difference in the sets of regression coefficients is expected. We observed that in the extreme heteroscedastic case when the difference between the coefficients is large $(30, 20, 25)'$, the power of the test is 0.800, which is also quite satisfactory. Therefore, the proposed generalized test maintains good power in testing equality of the two sets of regression coefficients of the two regimes in the heteroscedastic cases.

Table 4.11: Fixed X values for different sample sizes

Sample size	X_{i1}	X_{i2}
5	1	6333
	6	4066
	11	5388
	16	1551
	21	2596
10	1	3956
	6	1831
	11	3339
	16	8105
	21	8199
	26	4202
	31	5671
	36	8820
	41	9758
	46	5558
15	1	2950
	6	4639
	11	2875
	16	7977
	21	9563
	26	3239
	31	3635
	36	2923
	41	2672
	46	7584
	51	3521
	56	7141
	61	9702
	66	1451
	71	7655

Table 4.12: Observed significance levels of the generalized test (GT) in testing the equality of the two sets of regression coefficients in the two regimes with heteroscedasticity at 0.05 level of significance

n_1	n_2	σ_1^2	σ_2^2	Observed significance level of the GT
15	5	1	2	0.045
		1	4	0.048
		1	8	0.041
15	10	1	2	0.052
		1	4	0.049
		1	8	0.047
15	15	1	2	0.050
		1	4	0.044
		1	8	0.051

Table 4.13: Power of the generalized test (GT) in testing the equality of the two sets of regression coefficients in the two regimes of sample sizes $n_1 = 15$ and $n_2 = 5$ with heteroscedasticity at 0.05 level of significance

$\beta_{10} - \beta_{20}$	$\beta_{11} - \beta_{21}$	$\beta_{12} - \beta_{22}$	σ_1^2	σ_2^2	Power
0	0	0	1	2	0.045
			1	4	0.048
			1	8	0.041
10	5	10	1	2	0.541
			1	4	0.355
			1	8	0.211
20	10	15	1	2	0.908
			1	4	0.712
			1	8	0.444
30	15	20	1	2	0.996
			1	4	0.915
			1	8	0.732
30	20	25	1	2	0.998
			1	4	0.964
			1	8	0.800

4.3 Conclusion

In situations when the data were collected over two different regimes and the regression coefficients remained the same between regimes only the error variance varied by regimes, we proposed generalized confidence intervals (GCIs) for the regression coefficients. The simulation study shows that the GCIs preserve the confidence levels close to the nominal level. In testing equality of the regression coefficients in the two regimes, the generalized pivotal quantities of their differences and the generalized p-values were developed. When the regimes' error variances are different, the testing problem becomes a Behrens-Fisher problem in regression setting. In such case, the generalized test is comparatively better than the classical test based on equal variance assumption when regimes' sample sizes are unequal. The generalized test is found comparable with the test based on approximate t distribution in the heteroscedastic cases. Generalized methods are especially useful in multiparameter cases where nontrivial tests are difficult to obtain. To test the equality of the sets regression coefficients of two regimes under heteroscedasticity, we proposed a generalized test. The test preserves the nominal significance level and maintain satisfactory power.

Chapter 5

Application of Proposed Generalized Methods on Real Data

In this section, we applied the proposed methods on two data sets: the community health study data of Sarnia in 2005 and the US gasoline consumption data before and after the 1973 oil crisis. The programs were written in R and the statistical package SPSS was used in the data analyses.

5.1 Is air pollution in Sarnia causing respiratory problems among adults?

Sarnia is a major city in Southwestern Ontario, Canada. The city is also known as 'Chemical Valley' because over 40 percent of Canada's chemical industries are there. Toxic releases of those industries pollute the environment of the city and put the residents' life at risk of developing different diseases. The pollutants include respiratory toxicant, lead, mercury, benzene and nickel (Fung et al; 2007). Our objective is to examine whether the pollution has any impact on chronic respiratory problems. A community health study was conducted in the city in 2005. Residents of age 18 years or older were interviewed. A predesigned questionnaire was used in the survey. Most of the questions were about the mental and physical health of the respondents. Besides, they recorded responses such as length of stay in the region, presence in the region in the summer period, smoke, odours of the chemical plants etc. that can be used as a proxy for the exposure to the environment. Socio-demographic information of the residents was also documented.

5.1.1 Description of the data

The data were collected from the 5 regions of the city with postal codes N7S, N7T, N7V, N7W and N7X. The number of respondents interviewed from these regions are 268, 392, 120, 16, and 8, respectively. For each respondent the number of respiratory problems was counted. The respiratory problems included hay fever or other allergies, an attack of shortness of breath at any time in the last 12 months, an asthmatic attack in the last 12 months, current asthma and other respiratory problems. A respondent's length of stay in the community in completed years was recorded. It can be a proxy for the length of exposure to pollutions in Sarnia. Age of the respondent was categorized as 18-40 years and 40+, as older people are more vulnerable of developing health problems.

5.1.2 Analysis results

The scatter plot of the years of living in Sarnia and the number of respiratory problems (Figure 5.1) does not show a clear picture of relationship between them. Respondents were then grouped according to their length of stay and the mean of the number of respiratory problems was computed for each group. Figure 5.2 gives the scatter plot of the years of living in Sarnia and the mean number of respiratory problems for the five regions. A pattern of relationship emerges. If we discard the groups that have no respiratory problems from the analysis, a linear trend may exist (Figure 5.3). In this case the correlation coefficient is significant ($r = 0.264$, p - value $= 0.001$). We fitted a simple linear regression model with mean number of respiratory problems as response and years of living as regressor. Note that the data had been aggregated by regions and years of living and the model was fitted using the aggregated data. Table 5.1

gives analysis of variance for the model. Estimates with their 95 percent generalized confidence intervals (GCIs) and generalized p - values (GP-values) are given in Table 5.2. We observed that years of living in Sarnia is a significant regressor of mean number of respiratory problems. This gives an indication that long term exposure to the environmental pollution in Sarnia may create respiratory health hazard.

Since the data were collected from 5 different regions of Sarnia, we further examined whether the relationship holds in each region. Table 5.3 gives the correlation between years of living and mean number of respiratory problems by region. The significant relationship is observed only in region N7S. It is to be noted that in region N7X none had any respiratory problems. The estimates of the parameters for simple linear regression model with their corresponding GCIs (95 percent) and GP-values by region are given in Table 5.4. We observed that years of living is a significant regressor for mean number of respiratory problems in region N7S.

To illustrate our proposed generalized method for comparing two corresponding regression coefficients, we considered region N7S as regime 1 and N7T as regime 2. Table 5.5 presents the 95 percent GCIs for the differences of the regression coefficients with GP-values. It is shown that years of living in Sarnia has a significant different effect on the mean number of respiratory problems in region N7S from that of region N7T.

One might argue that the observed positive relationship between years of living in Sarnia and mean number of respiratory problems is due to the respondent's age. We controlled the effect of age by fitting a multiple linear regression model of mean number or respiratory problems on years of living and age of respondents. Here, the second stage of aggregation was done on the data. The data had been aggregated by region, years of living and age of the respondent. The analysis results are given in Table 5.6 and 5.7. We observed that the positive relationship between years of living in Sarnia and mean number of respiratory problems still holds at $\alpha = 0.05$.

5.1.3 Conclusion

Continuous disposal of industrial toxicant made the environment of Sarnia polluted and put the residents' lives at risk. Surveys are needed to measure the extent of pollution and it's impact on human lives and environment. Measures should also be taken to recycle the industrial waste.

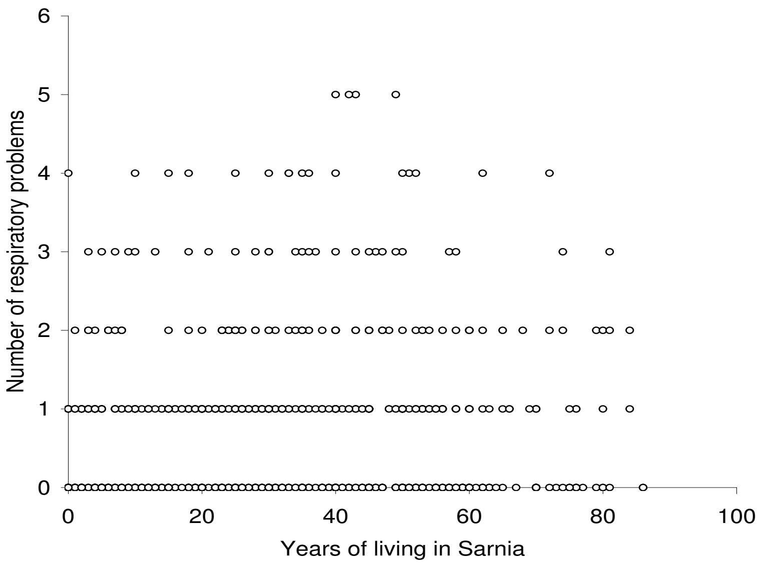

Figure 5.1: The scatter plot of the years of living in Sarnia and the number of respiratory problems

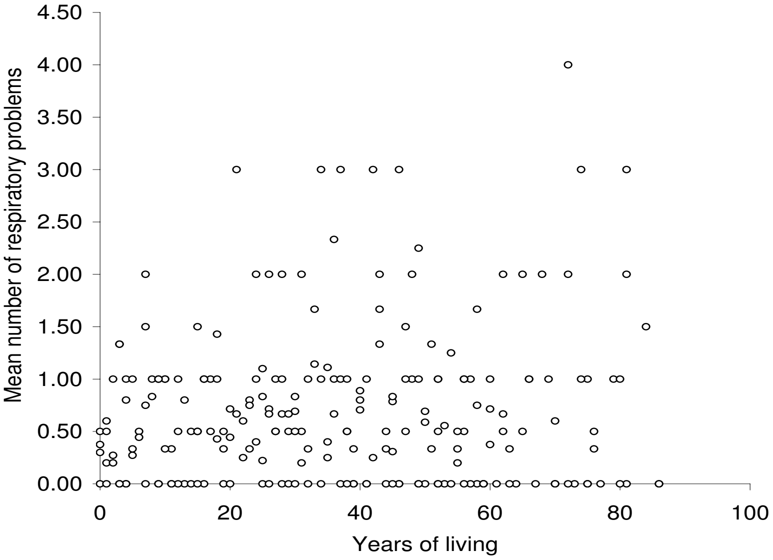

Figure 5.2: The scatter plot of the years of living and the mean number of respiratory problems for the five regions in Sarnia

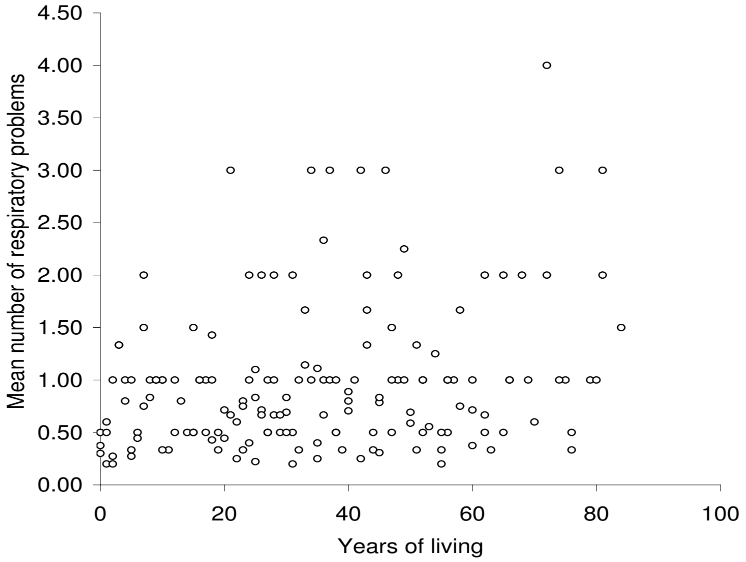

Figure 5.3: The scatter plot of the years of living and the mean number of respiratory problems for the five regions in Sarnia after excluding groups with no respiratory problems

Table 5.1: Analysis of variance for the linear regression model of mean number of respiratory problems on years of living in Sarnia

Source of variation	Sum of squares	df	Mean square	F	Sig
Years of living	5.466	1	5.466	11.893	0.001
Residual	73.069	159	0.460		
Total	78.535	160			

Table 5.2: Estimates of the parameters for the linear regression model of mean number of respiratory problems on years of living in Sarnia

Variables in the model	Estimate	95% GCI	GP-value
Intercept	0.684	(0.485, 0.882)	0.000
Years of living	0.008	(0.004, 0.013)	0.001

Table 5.3: the Correlation coefficient between years of living in Sarnia and the mean number of respiratory problems by region

Region	r	p-value	sample size
N7S	0.412	0.001	58
N7T	0.085	0.509	63
N7V	0.279	0.105	35
N7W	0.779	0.121	5
N7X	-	-	-

Table 5.4: Estimates of the parameters for the linear regression model of mean number of respiratory problems on years of living in Sarnia by region

Region	Variable	Parameters	Estimates	95% GCI	GP-value
N7S	Intercept	β_{10}	0.538	(0.224, 0.852)	0.001
	Years of living	β_{11}	0.013	(0.005, 0.021)	0.001
N7T	Intercept	β_{20}	0.801	(0.506, 1.095)	0.000
	Years of living	β_{21}	0.002	(-0.005, 0.009)	0.509
N7V	Intercept	β_{30}	0.659	(0.135, 1.182)	0.015
	Years of living	β_{31}	0.010	(-0.002, 0.023)	0.105
N7W	Intercept	β_{40}	0.977	(-0.493, 2.447)	0.125
	Years of living	β_{41}	0.026	(-0.012, 0.630)	0.121

Table 5.5: The 95 percent Generalized confidence intervals for the difference of the regression coefficients in regions N7S and N7T

Difference of the parametrs	GCI	GP-value
$\beta_{10} - \beta_{20}$	(-0.694, 0.163)	0.228
$\beta_{11} - \beta_{21}$	(0.0004, 0.021)	0.039

Table 5.6: Analysis of variance for the linear regression model of mean number of respiratory problems on years of living in Sarnia and age of respondents

Source of variation	Sum of squares	df	Mean square	F	Sig
Years of living	3.369	1	3.369	7.122	0.008
Age of respondents	0.269	1	0.269	0.569	0.452
Residual	90.396	191	0.473		
Total	94.034	193			

Table 5.7: Estimates of the parameters for the linear regression model of mean number of respiratory problems on years of living in Sarnia and age of respondents

Variables in the model	Estimate	95%GCI	GP-value
Intercept	0.836	(0.637, 1.035)	0.000
Years of living	0.007	(0.002, 0.012)	0.008
Age category	-0.049	(-0.280, 0.181)	0.673

5.2 Impact of gasoline price on total US gasoline consumption before and after 1973 oil crisis

The Oil crisis began when members of the Organization of the Petroleum Exporting Countries (OPEC) with Egypt and Syria placed an oil embargo on the United States on October 15, 1973. This came as a punishment for the US decision to resupply the Israeli military during the October War, the fourth Arab-Israeli war. The embargo caused a persistent effect in the US economy because the industrialized US economy was heavily dependent on the crude oil and OPEC was their predominant supplier.

Here we studied the US gasoline consumption data from 1960 to 1995. We fitted a linear regression model to the data. Then we divided the data into two regimes: before 1973 as regime 1 and on or after 1973 as regime 2 and examined using our proposed generalized test variable whether the set of regression coefficients of the regime 1 model is different from that of regime 2 model.

5.2.1 Description of the data

The data were taken from the Council of Economic Advisers, Report of the President 1996 (http://pages.stern.nyu.edu/∼ wgreene/Text/tables/TableF2-2.txt). The data consist of total US gasoline consumption (expenditure/price index), price index for gasoline, percapita disposable income, price index for new cars, price index for used cars, price index for public transport, aggregate price index for consumer durables, aggregate price index for consumer non-durables, aggregate price index for consumer services, and US total population.

5.2.2 Analysis results

A linear regression model was selected using stepwise selection criteria taking total US gasoline consumption as the response variable and all the remaining variables as the regressors. Collinearity among the independent variables (Myers; 1990) was also taken care in the model selection. The final model included price index for gasoline and price index for used cars as the regressors. The overall model fit was good ($R^2 = 0.83$ and $F_{2,32} = 37$ with p-value < 0.001). Table 5.8 presents the estimated regression coefficients with their GCIs and GP-values. Collinearity diagnostics for the model are presented in Table 5.9. We observed that the values of the variance inflation factor are less than 10, none of the condition index is greater than 30 and the smallest eigen value is not closed to zero. All these diagnostics indicate that collinearity is not a problem for the model. The normality assumption of the errors was also satisfied (Shapiro-Wilk W=0.967, p-value=0.361)

We observed that (Table 5.8) the price index of gasoline is not a significant regressor for total US gasoline consumption for the period of 1960 to 1995. This result is opposite to what we expected. The reason could be that we analyzed the data of two very different economic era, before the 1973 oil crisis and after the crisis, together. Figures 5.4 and 5.5 give the plots of the consumption of gasoline and the price index for gasoline by years, respectively. We observed that before the crisis, there was not that much variation in gasoline price and its consumption increased steadily. But after the crisis, we observed much more variation in gasoline price and its consumption. The scatter plot of the price and consumption of gasoline before and after the oil crisis is given in Figure 5.6. Before the crisis, a positive linear relationship between price and consumption of gasoline was observed. But it might not be the case after the crisis.

Then, we splitted the data in to two regimes: before the oil crisis and on or after the crisis and fitted the model separately for the two regimes. The estimated regression coefficients for the two regimes are given in Table 5.10. We observed that before the oil crisis, the price index of gasoline was positively related with the total consumption of gasoline. Although increase in price index resulted in increase in consumption, it is not surprising, because at the time, price of gasoline was inexpensive. However, between 1973 to 1995, we observed a significant reverse relationship between price index for gasoline and total gasoline consumption. After the oil embargo in 1973, the price of gasoline increased abruptly and became expensive. As a result, increase in price index resulted in decrease in gasoline consumption.

We observed that the estimated regression coefficients of the two regimes are quite different from each other. Severe heterogeneity in error variances of the two regimes was also observed. Error mean squares for regimes 1 and 2 are 17.416 and 99.56, respectively, that are significantly different ($F_{20,10}$=5.717, p-value=0.004). In such case, we would like to test whether the two sets of regression coefficients are different, i.e.

$$H_0 : \beta_1 = \beta_2 \text{ against } H_1 : \beta_1 \neq \beta_2,$$

where β_i is the vector of regression coefficients for the regime $i = 1, 2$. As the sample size of the first regime is small ($n_1 = 13$) and severe heteroscedasticity between two regimes exists, the classical asymptotic test and the Chow test are not appropriate for this case. We applied our proposed generalized test for testing the hypothesis, since the test can be efficiently used for small sample sizes and heteroscedastic case. The observed generalized test variable is 495406.5 with generalized p-value < 0.0001. Therefore, at 0.05 level of significance we reject the null hypothesis.

Then, we tested the equality of each of the regression coefficients of regime 1 with the corresponding regression coefficients of regime 2. We constructed the GCIs for the difference of the corresponding regression coefficients of the two regimes and then their GP-values. There are 3 regression coefficients. Therefore, to maintain the global level at least approximately to be 0.05, we construct GCIs each with confidence coefficients $(1 - 0.05/3) \cong 0.98$. Table 4.11 gives the GCIs for the difference of the regression coefficients with their GP-values. We observed that after the oil crisis in 1973, increase in gasoline price resulted in significant decrease in its consumption unlike prior 1973.

5.2.3 Conclusion

When we suspect regime change in the data, analysis should be done by regimes. The equality of the sets of regression parameters of the regimes should be tested. We applied generalized test that can efficiently test the equality of the sets of regression coefficients of two regimes for all sample sizes and heteroscedastic case.

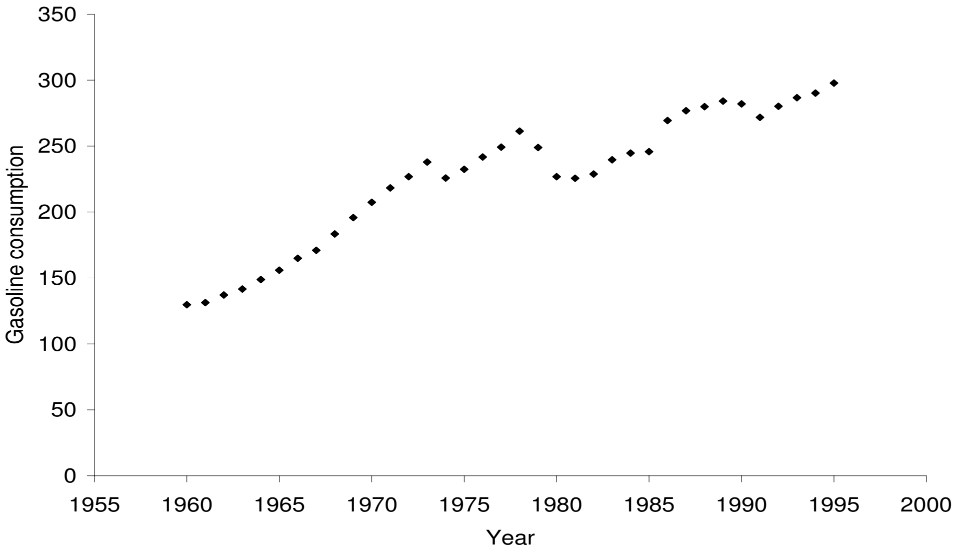

Figure 5.4: The plot of the consumption of gasoline by years

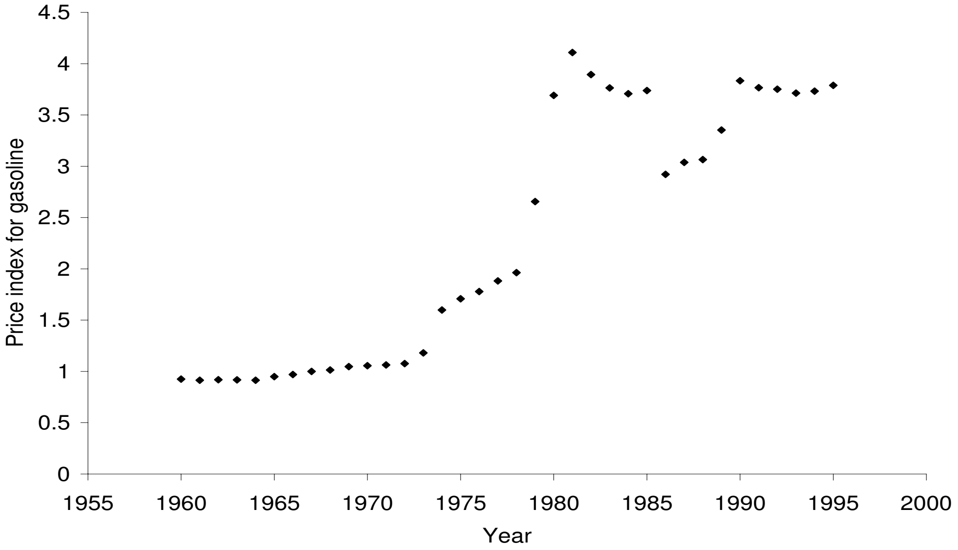

Figure 5.5: The plot of the price index for gasoline by years

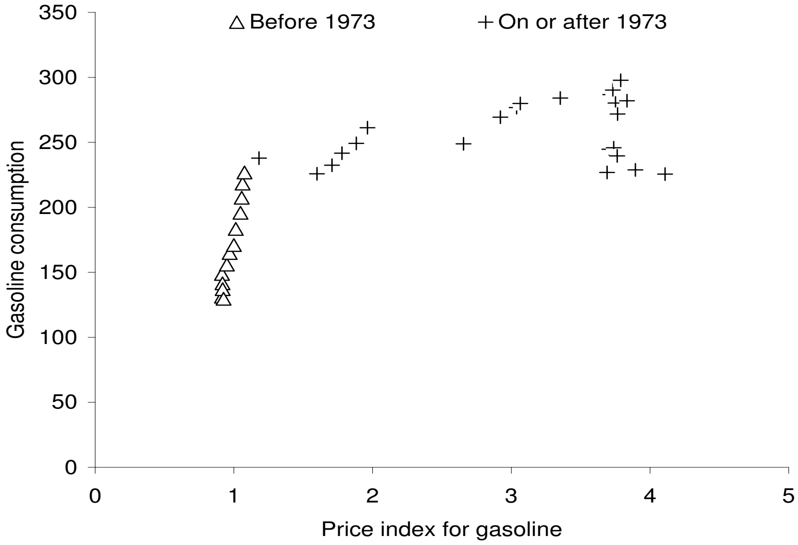

Figure 5.6: The scatter plot of the price and consumption of gasoline before and after the oil crisis

Table 5.8: Estimates of the parameters for the linear regression model of the total US gasoline consumption on price index for gasoline and price index for used cars during 1960 and 1995

Variables in the model	Estimate	95% GCI	GP-value	VIF
Intercept	154.42	(133.54, 173.31)	0.000	
Price index for gasoline	4.263	(-14.34, 22.861)	0.644	5.466
Price index for used cars	26.37	(9.86, 42.88)	0.003	5.466

Table 5.9: Collinearity diagnostics for the linear regression model of the total US gasoline consumption on price index for gasoline (PIG) and price index for used cars (PUC) during 1960 and 1995

			Variance proportion		
Dimension	Eigen value	Condition index	Constant	PIG	PUC
1	2.813	1.000	0.02	0.00	0.01
2	0.164	4.144	0.91	0.03	0.06
3	0.023	11.099	0.06	0.97	0.93

Table 5.10: Estimates of the parameters for the linear regression model of the total US gasoline consumption on price index for gasoline (PIG) and price index for used cars (PUC) before and after the oil crisis in 1973

Regime	Variables	Estimate	95% GCI	GP-value
1960-1972	Intercept	-337.03	(-378.68, -295.37)	0.000
	PIG	381.14	(307.10, 455.18)	0.000
	PUC	134.09	(74.04, 194.14)	0.001
1973-1995	Intercept	232.20	(216.73, 247.67)	0.000
	PIG	-18.66	(-26.36, -10.95)	0.000
	PUC	26.64	(20.80, 32.48)	0.000

Table 5.11: The Generalized confidence intervals for the difference of the corresponding regression coefficients of two regimes each with confidence coefficient 0.98

Difference of the parametrs	GCI	GP-value
$\beta_{10} - \beta_{20}$	(-623.6, -514.4)	0.000
$\beta_{11} - \beta_{21}$	(309.28, 491.9)	0.000
$\beta_{12} - \beta_{22}$	(31.02, 182.55)	0.003

Chapter 6

Conclusion and Future Research

Weerahandi (1989, 1993) developed the concept of generalized confidence intervals and generalized p-values for complex inference problems involving nuisance parameters. This generalized methodology of inference is suitable for all sample sizes and is found to be efficient even when the assumptions of optimal inference do not hold.

In this thesis we applied the concept to the case of simple and multiple linear regression models. Specifically, we constructed generalized confidence intervals for regression coefficients, dispersion parameters and the expected response for simple and multiple linear regression models. We also constructed generalized confidence regions for multiple linear regression parameters.

The regression data from two different regimes were considered. We considered a particular case when the regression coefficients remained the same for the two regimes and the error variances were assumed same within the regime but different between regimes. We proposed generalized confidence regions and generalized confidence intervals for the regression parameters in such a case. The global confidence level was maintained using the Bonferroni approach. The simulation study showed that the GCIs preserve the confidence levels close to the nominal level for overall sample size $n = n_1 + n_2 \geq 14$.

In testing the equality of the regression coefficients in the two regimes, we developed the generalized pivotal quantities of their differences and the generalized p-values. From the simulation study we observed that when the regimes' error variances are different, the generalized test is comparatively better than the classical test based on equal variance assumption in the case of unequal regimes' sample sizes. The test is found comparable with the test based on approximate t distribution in the heteroscedastic cases.

Generalized methods are especially useful in multiparameter cases where nontrivial tests are difficult to obtain. To test the equality of the sets regression coefficients of two regimes under heteroscedasticity, the Chow test was extended by Toyoda (1974). But the test's significance level is greatly affected if the regimes have small sample sizes. In classical inference, only asymptotic tests are available for this problem. In such a case, we proposed generalized test variables and generalized p-values that can be applied efficiently for all sample sizes and for homoscedastic as well as heteroscedastic cases. The simulation study showed that the proposed method preserves the nominal significance level and provides satisfactory power under heteroscedasticity and for small and moderate sample sizes.

We also constructed the generalized confidence region for the difference of the two sets of regression coefficients and then, the generalized confidence intervals for each elements of that vector.

We applied our proposed methodology on the two data sets: the community health study data of Sarnia in 2005 and the US gasoline consumption data before and after the 1973 oil crisis. These applications clearly showed that there is a regime change in the data. Accordingly, the analysis should be done by regimes.

In this thesis, regressor variables were considered to be non-stochastic. Also, the error distribution was assumed to be normally distributed. For future research, we plan to investigate

1. the general case where independent variables may be stochastic and

2. the complex cases where the normality assumption of the error is not satisfied.

Appendix A

A.1 Distributions of the estimators of the error variance in linear regression

Consider the linear regression model

$$Y_j = \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_p X_{pj} + \varepsilon_j,\ j = 1, 2, \ldots, n.$$

Here Y_j is the response variable for the j^{th} set of values of $(X_1, X_2, \ldots, X_p)$ and ε_j is the corresponding random error term. Also, $\beta_i,\ i = 1, 2, \ldots, p$ represent a total of p unknown parameters. Defining

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix},\ \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},\ X = \begin{pmatrix} X_{11} & X_{21} & \ldots & X_{p1} \\ X_{12} & X_{22} & \ldots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & \ldots & X_{pn} \end{pmatrix} \text{ and } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

the model can be written in matrix notation as

$$Y = X\beta + \varepsilon,$$

where X has full column rank. The ordinary least squares estimator for β is

$$\hat{\beta} = b = (X'X)^{-1}X'Y,\ \text{where } b = (b_1, b_2, \ldots, b_p)',$$

and given the normality assumption on $\varepsilon \sim N_n(0, \sigma^2 I_n)$,

$$b \sim N_p\left(\beta, \sigma^2 (X'X)^{-1}\right) \tag{A.1}$$

and

$$Y \sim N_n(X\beta, \sigma^2 I_n). \tag{A.2}$$

In this thesis we consider two estimators of the error variance σ^2:

$$S^2 = \frac{(Y-\hat{Y})'(Y-\hat{Y})}{(n-p)} = \frac{(Y-X\hat{\beta})'(Y-X\hat{\beta})}{(n-p)} \text{ and}$$

$$S_Y^2 = \frac{(Y-X\hat{\beta})'(Y-X\hat{\beta})}{n}.$$

The estimator S^2, the error mean square, is unbiased and the estimator S_Y^2, the maximum likelihood estimator, is biased. We are to find the distributions of these two estimators. From (A.2),

$$\frac{Y}{\sigma} \sim N_n\left(\frac{X\beta}{\sigma}, I_n\right).$$

Let us define

$$W = \frac{Y-X\hat{\beta}}{\sigma}.$$

Then,

$$\left(\frac{Y-X\hat{\beta}}{\sigma}\right)'\left(\frac{Y-X\hat{\beta}}{\sigma}\right) = \frac{1}{\sigma^2}(Y-X\hat{\beta})'(Y-X\hat{\beta}). \tag{A.3}$$

Since

$$Y-X\hat{\beta} = Y - X(X'X)^{-1}X'Y = \left[I - X(X'X)^{-1}X'\right]Y,$$

where

$$\left[I - X(X'X)^{-1}X'\right]Y$$

is an idempotent matrix, from (A.3) we have

$$\frac{1}{\sigma^2}(Y-X\hat{\beta})'(Y-X\hat{\beta}) = \frac{1}{\sigma^2}\, Y'\left[I - X(X'X)^{-1}X'\right]Y,$$

or,

$$\frac{1}{\sigma^2}(Y-X\hat{\beta})'(Y-X\hat{\beta}) = \left(\frac{Y}{\sigma}\right)'\left[I - X(X'X)^{-1}X'\right]\left(\frac{Y}{\sigma}\right). \tag{A.4}$$

Recall that if $X \sim N_n(\mu, \Sigma)$ where Σ is positive definite, then

$$X'AX \sim \chi^2_{\text{rank}(A)}\ (\mu' A \mu)$$

iff ΣA is idempotent. Further, rank$(A) =$ trace(ΣA).

Therefore, from (A.4),

$$\frac{1}{\sigma^2}(Y - X\hat{\beta})'(Y - X\hat{\beta}) \sim \chi^2$$

with degrees of freedom equal to

$$\text{trace}\left[I - X(X'X)^{-1}X'\right]$$

and non-centrality parameter

$$\left(\frac{X\beta}{\sigma}\right)' \left[I - X(X'X)^{-1}X'\right] \left(\frac{X\beta}{\sigma}\right).$$

One can verify that

$$\text{trace}\left[I - X(X'X)^{-1}X'\right] = n - p$$

and

$$\left(\frac{X\beta}{\sigma}\right)' \left[I - X(X'X)^{-1}X'\right] \left(\frac{X\beta}{\sigma}\right) = 0.$$

Thus,

$$\frac{1}{\sigma^2}(Y - X\hat{\beta})'(Y - X\hat{\beta}) \sim \chi^2_{(n-p)}.$$

Therefore, the error mean square $S^2 = \frac{(Y-X\hat{\beta})'(Y-X\hat{\beta})}{(n-p)}$ is distributed as

$$\frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{(n-p)}$$

and the maximum likelihood estimator $S^2_Y = \frac{(Y-X\hat{\beta})'(Y-X\hat{\beta})}{(n)}$ is distributed as

$$\frac{nS^2_Y}{\sigma^2} \sim \chi^2_{(n-p)}.$$

A.2 Ordinary least squares estimators $\hat{\beta}$ is independent of error mean square

Consider the regression model defined in appendix A.1

$$Y = X\beta + \varepsilon.$$

The ordinary least squares estimator for β is

$$\hat{\beta} = b = (X'X)^{-1}X'Y, \text{ where } b = (b_1, b_2, \ldots, b_p)',$$

and given the normality assumption on $\varepsilon \sim N_n(0, \sigma^2 I_n)$,

$$b \sim N_p\left(\beta, \sigma^2(X'X)^{-1}\right).$$

An unbiased estimator of σ^2 is $S^2 = \frac{(Y-\hat{Y})'(Y-\hat{Y})}{(n-p)} = \frac{(Y-X\hat{\beta})'(Y-X\hat{\beta})}{(n-p)}$. We are to show that $b = \hat{\beta}$ and S^2 are independent.

For σ fixed, $\hat{\beta}$ is a complete sufficient statistic for β while $(Y - X\hat{\beta})'(Y - X\hat{\beta})$ is an ancillary statistic for β.

Recall Basu's theorem (Basu; 1955, Lehmann; 1981) that states any complete sufficient statistic is independent of any ancillary statistic. Therefore, by Basu's theorem, we conclude that $\hat{\beta}$ and $(Y - X\hat{\beta})'(Y - X\hat{\beta})$ are independent. Since S^2 is a function of $(Y - X\hat{\beta})'(Y - X\hat{\beta})$, we have $\hat{\beta}$ and S^2 are independent.

A.3 U and V are independent

Consider the regression model defined in appendix A.1

$$Y = X\beta + \varepsilon.$$

The ordinary least squares estimator for β is

$$\hat{\beta} = b = (X'X)^{-1}X'Y, \text{ where } b = (b_1, b_2, \ldots, b_p)',$$

and given the normality assumption on $\varepsilon \sim N_n(0, \sigma^2 I_n)$,

$$b \sim N_p\left(\beta, \sigma^2(X'X)^{-1}\right)$$

and

$$Y \sim N_n(X\beta, \sigma^2 I_n).$$

An unbiased estimator of σ^2 is $S^2 = (Y - \hat{Y})'(Y - \hat{Y})/(n-p) = \frac{\Sigma(Y_j - \hat{Y}_j)^2}{n-p} =$ Error MS$, j = 1, 2, \ldots, n.$

Now, the total variability in the data $(Y - X\beta)'(Y - X\beta)$ can be decomposed into two parts as

$$\begin{aligned}(Y - X\beta)'(Y - X\beta) &= (Y - \hat{Y} + \hat{Y} - X\beta)'(Y - \hat{Y} + \hat{Y} - X\beta)\\ &= \Big[(Y - \hat{Y}) + X(b - \beta)\Big]' \Big[(Y - \hat{Y}) + X(b - \beta)\Big]\end{aligned}$$

$$= (Y - \hat{Y})'(Y - \hat{Y}) + (Y - \hat{Y})'X(b - \beta)$$

$$+(b - \beta)'X'(Y - \hat{Y}) + (b - \beta)'X'X(b - \beta)$$

$$= (Y - \hat{Y})'(Y - \hat{Y}) + (Y'X - \hat{Y}'X)(b - \beta)$$

$$+(b - \beta)(X'Y - X'\hat{Y}) + (b - \beta)'X'X(b - \beta).$$

If we put $\hat{Y}'X = Y'X$ and $X'\hat{Y} = X'Y$ in the above equation we get

$$(Y - X\beta)'(Y - X\beta) = (Y - \hat{Y})'(Y - \hat{Y}) + (b - \beta)'X'X(b - \beta).$$

It implies

$$\frac{1}{\sigma^2}(Y - X\beta)'(Y - X\beta) = \frac{1}{\sigma^2}(Y - \hat{Y})'(Y - \hat{Y}) + \frac{1}{\sigma^2}(b - \beta)'X'X(b - \beta).$$

Now,

$$\frac{1}{\sigma^2}(Y - X\beta)'(Y - X\beta) = (Y - X\beta)'(\sigma^2 I_n)^{-1}(Y - X\beta) \sim \chi^2_n,$$

$$\frac{1}{\sigma^2}(Y - \hat{Y})'(Y - \hat{Y}) = \frac{\Sigma(Y_j - \hat{Y}_j)^2}{\sigma^2} \sim \chi^2_{n-p} \text{ and}$$

$$\frac{1}{\sigma^2}(b - \beta)'X'X(b - \beta) = (b - \beta)'(\sigma^2(X'X)^{-1})^{-1}(b - \beta) \sim \chi^2_p.$$

Recall **Cochran's theorem** (Montgomery; 2006):

Let Z_i be NID(0,1) for $i = 1, 2, \ldots, v$ and

$$\sum Z_i^2 = Q_1 + Q_2 + \ldots + Q_s,$$

where $s \leq v$ and Q_j has v_j degrees of freedom $(j = 1, 2, \ldots, s)$. Then $Q_1, Q_2, \ldots, Q_s$ are independent chi-square random variables with $v_1, v_2, \ldots, v_s$ degrees of freedom, respectively, if and only if

$$v = v_1 + v_2 + \ldots + v_s.$$

Because the degrees of freedom of $\frac{1}{\sigma^2}(Y - \hat{Y})'(Y - \hat{Y})$ and $\frac{1}{\sigma^2}(b - \beta)'X'X(b - \beta)$ add to n, the total degrees of freedom, Cochran's theorem implies that they are independently distributed chi-square random variables.

Now, $V = \frac{S^2}{\sigma^2 s^2} = \frac{(Y - \hat{Y})'(Y - \hat{Y})}{(n-p)\sigma^2 s^2}$ is a function of $\frac{1}{\sigma^2}(Y - \hat{Y})'(Y - \hat{Y})$. Therefor, V and $U = \frac{1}{\sigma^2}(b - \beta)'X'X(b - \beta)$ are independent of each other.

Bibliography

[1] Alt, F.B. (1982). Bonferroni inequalities and intervals. *Encyclopedia of Statistical Sciences*, Vol. 1: 294-300.

[2] Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhy, Series A*, Vol. 15: 377-80.

[3] Bebu, I. and Mathew, T. (2007). Comparing the means and variances of a bivariate log-normal distribution. *Statistics in Medicine*, Vol. 27, No. 14: 2684-96.

[4] Chow, G.C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, Vol 28, No. 3: 591-605.

[5] Dougherty, C. (1992). *Introduction to Econometrics*. Oxford University Press.

[6] Fung, K.Y., Luginaah I.N. and Gorey, K.M. (2007). Impact of air pollution on hospital admissions in Southwestern Ontario, Canada: generating hypotheses in sentinel high-exposure places. *Environmental Health*, Vol. 6, No. 18: 1-7.

[7] Gamage, J., Mathew, T. and Weerahandi, S. (2004). Generalized p-values and generalized confidence regions for the multivariate Behrens-Fisher problem and MANOVA. *Journal of Multivariate Analysis*, Vol. 88, No. 1: 177-89.

[8] Gujarati, D.N. (1995). *Basic Econometrics*, (3rd ed.). McGraw-Hill, Inc.

[9] Hannig, J., Iyer, H. and Patterson, P. (2006). Fiducial generalized confidence intervals. *Journal of the American Statistical Association* , Vol. 101, No. 473: 254-69.

[10] Hogg, R.V., McKean, J.W. and Craig, A.T. (2007). *Introduction to Mathematical Statistics*, (6th ed., 2nd Impression). Pearson Education, Inc.

[11] Johnson, R.A. and Wichern, D.W. (1988). *Applied Multivariate Statistical Analysis*, (2nd ed.). Prentice Hall, Inc.

[12] Jordan, S.M. and Krishnamoorthy, K. (1995). Confidence regions for the common mean vector of several multivariate normal populations. *The Canadian Journal of Statistics*, Vol. 23, No. 3: 283-97.

[13] Krishnamoorthy, K., Mathew, T. and Ramachandran, G. (2006). Generalized p-values and confidence Intervals: a novel approach for analyzing lognormally distributed exposure data. *Journal of Occupational and Environmental Hygiene*, Vol. 3, No. 11: 642-50.

[14] Krishnamoorthy, K., Mathew, T. and Ramachandran, G. (2007). Upper limits for exceedance probabilities under the one-way random effects model. *Annals of Occupational Hygiene*, Vol. 51, No. 4: 397-406.

[15] Lehmann E.L. (1981). An Interpretation of Completeness and Basu's Theorem. *Journal of the American Statistical Association*, Vol. 76, No. 374: 335-40.

[16] Li, X., Xu, X and Li, G. (2007). A fiducial argument for generalized p -value . *Science in China, Series A: Mathematics, physics, astronomy and technological sciences*, Vol. 50, No. 7: 957-66.

[17] Lin, S. and Lee, J.C. (2005). Generalized inferences on the common mean of several normal populations. *Journal of Statistical Planning and Inference*, Vol. 134, No. 2: 568-82.

[18] Mathew, T. and Webb, D.W. (2005). Generalized p values and confidence intervals for variance components: applications to army test and evaluation. *Technometrics*, Vol. 47, No. 3: 312-22.

[19] McNally, R.J., Iyer, H. and Mathew, T. (2003). Tests for individual and population bioequivalence based on generalized p-values. *Statistics in Medicine*, Vol. 22, No. 1:31-53.

[20] Montgomery, D.C. (2006). *Design and Analysis of Experiments*, (5th ed.). John Wiley and Sons, Inc.

[21] Myers, R.H. (1990). *Classical and Modern Regression with Application*, (2nd ed.). Duxbury Thomson Learning.

[22] Nickerson, D.M. (1994). Construction of a conservative confidence region from projections of an exact confidence region in multiple linear regression. *The American Statistician*. Vol. 48, No. 2: 120-24.

[23] Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika*. Vol. 6, N0. 5: 309-16.

[24] Schechtman, E. and Sherman, M. (2007). The two-sample t-test with a known ratio of variances. *Statistical Methodology*. Vol. 4: 508-14.

[25] Scheffe, H. (1959). *The Analysis of Variance*. New York: John Wiley.

[26] Toyoda, T. (1974). Use of the Chow test under heteroscedasticity. *Econometrica*, Vol. 42, No. 3: 601-8.

[27] Tsui, K. and Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, Vol. 84, No. 406: 602-7.

[28] Wang, C.M. and Iyer, H.K. (2006). A generalized confidence interval for a measurand in the presence of type-A and type-B uncertainties. *Advanced Mathematical Tools for Measurement in Metrology and Testing* , Vol. 39, No. 9: 856-63.

[29] Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association*, Vol. 88, No. 423: 899-905.

[30] Welch, B.L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, Vol. 29, No. 3/4: 350-362.

[31] Working, H. and Hotteling, H. (1929). Application of the theory of error to the interpretation of trends. *Journal of the American Statistical Association*, Vol. 24, No. 165, Suppl.: 73-85.

Vita Auctoris

Quazi Imad Uddin Ibrahim was born in 1976 in Sylhet, a district town in Bangladesh. He completed his Higher Secondary School studies in 1993 from M.C. College. From there he went on to the University of Dhaka where he obtained a B.Sc. in Statistics in 1998 and M.Sc. in 2000. He is currently a candidate for the Master degree in Statistics at the University of Windsor and hopes to graduate in June 2009.