REGRESSION AND ANOVA UNDER HETEROGENEITY

by

Elmer Agudelo Rodríguez

Bachelor of Science, Mathematics and Physics Education, Antioquia University,

Colombia (May 1994)

Graduate Certificate in Pedagogy Contemporaneous Tendencies, Pontificia

Bolivariana University, Colombia (June 2000)

A Research Paper
Submitted in Partial Fulfillment of the Requirements for the
Master of Science Degree

Department of Mathematics
in the Graduate School
Southern Illinois University Carbondale
November, 2007

ABSTRACT

The aim in writing this paper is to present plots for Ordinary Least Squares (OLS), Generalized Least Squares (GLS) and Weighted Least Squares (WLS) and to study GLS and one way analysis of variance (ANOVA) testing under heterogeneity.

In particular, when we have heteroscedastic error, there is no guarantee that the OLS estimator is the most efficient within the class of linear unbiased (or the class of unbiased) estimators. Nonconstant error variance affects the properties of the OLS estimators and resulting test statistics. Recall the Gauss-Markov Theorem: Under the classical linear regression assumptions, the OLS estimator is the BLUE (Best Linear Unbiased Estimator).

Moreover, hypothesis testing based on the standard OLS estimator of the variance covariance matrix becomes invalid, and we need new estimation methods: GLS, WLS, FGLS (Feasible Generalized Least Squares).

Simulations for one way ANOVA compare the ANOVA F test (F), modified ANOVA F test (F_M), the Welch ANOVA F test (F_W) and ANOVA F rank test (F_R). Power was examined and the Welch test performed well.

KEY WORDS: OLS; GLS; WLS; FGLS; Heterogeneity; Linear Regression.

DEDICATION

To Elida.

ACKNOWLEDGMENTS

I would like to thank Dr. David Olive for his invaluable assistance and insights leading to the writing of this paper, from whose book many of the examples and definitions used in this sample research paper have been quoted. My sincere thanks also goes to the members of my graduate committee, Dr. Sakthivel Jeyaratnam and Dr. Walter Wallis for their patience and time dedicated to read and discuss this paper.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

For the classical linear regression model:

$$Y_i = \beta_1 + \beta_2 \cdot X_{i2} + ... + \beta_p \cdot X_{ip} + e_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i, \quad \text{where} \quad Var(e_i) = \sigma^2. \tag{1}$$

This is called the homoscedastic (constant variance) model.

However, consider as an example the case consumption and income relationship. The consumption variation of lower income groups is less than that of higher income groups. Thus, the assumption of constant variance may not be appropriate.

When we do not have constant variance, we have a heteroscedastic model. Often, the conditional variance is function of explanatory variables.

Then (1) is written:

$$Y_i = \beta_1 + \beta_2 \cdot X_{i2} + ... + \beta_p \cdot X_{ip} + e_i, \quad \text{and} \quad \mathrm{Var}(e_i) = \sigma_i^2. \tag{2}$$

When there is heteroscedastic error, $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ is no longer a scalar multiple of the identity matrix, and hence there is no guarantee that the OLS estimator is the most efficient within the class of linear unbiased (or the class of unbiased) estimators. The constant variance assumption is important. Recall the Gauss-Markov Theorem: Under the classical linear regression assumptions, the OLS estimator is the BLUE (Best Linear Unbiased Estimator).

Moreover, hypothesis testing based on the standard OLS estimator of the variance covariance matrix becomes invalid, and new estimation methods are needed: GLS and WLS.

The aim in writing this paper is to present plots for OLS, GLS and WLS and to study GLS and one way analysis of variance (ANOVA) testing under heterogeneity.

The approach will start in Chapter 1 by making some definitions and assumptions related to WLS and Generalized Least Squares

Chapter 2 deals with some definitions and examples of fixed effects one way ANOVA and provides a simulation of the ANOVA F test, modified ANOVA F test, the Welch ANOVA F test, and ANOVA F rank test. The notation and examples follow Olive (2007) closely.

CHAPTER 1

RANDOM VECTORS

1.1 OLS

The concepts of a random vector, the expected value of a random vector and the covariance of a random vector are needed before covering generalized least squares. Recall that for random variables Y_i and Y_j, the covariance of Y_i and Y_j is $\text{Cov}(Y_i, Y_j) \equiv \sigma_{i,j} = E[(Y_i - E(Y_i))(Y_j - E(Y_j)] = E(Y_iY_j) - E(Y_i)E(Y_j)$ provided the second moments of Y_i and Y_j exist.

Definition 1.1.1. $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ is an $n \times 1$ **random vector** if Y_i is a random variable for $i = 1, ..., n$. $\boldsymbol{Y}$ is a discrete random vector if each Y_i is discrete and $\boldsymbol{Y}$ is a continuous random vector if each Y_i is continuous. A random variable Y_1 is the special case of a random vector with $n = 1$.

Definition 1.1.2. The *population mean* of a random $n \times 1$ vector $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ is

$$E(\boldsymbol{Y}) = (E(Y_1), ..., E(Y_n))^T$$

provided that $E(Y_i)$ exists for $i = 1, ..., n$. Otherwise the expected value does not exist. The $n \times n$ *population covariance matrix*

$$\text{Cov}(\boldsymbol{Y}) = E[(\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T] = ((\sigma_{i,j}))$$

where the ij entry of $\text{Cov}(\boldsymbol{Y})$ is $\text{Cov}(Y_i, Y_j) = \sigma_{i,j}$ provided that each $\sigma_{i,j}$ exists. Otherwise $\text{Cov}(\boldsymbol{Y})$ does not exist.

Definition 1.1.3. The covariance matrix is also called the variance–covariance matrix and variance matrix. Sometimes the notation $\text{Var}(\boldsymbol{Y})$ is used. Note that $\text{Cov}(\boldsymbol{Y})$ is a symmetric positive semidefinite matrix. If $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are $n \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{a} + \boldsymbol{Y}) = \boldsymbol{a} + E(\boldsymbol{Y}) \quad \text{and} \quad E(\boldsymbol{Y} + \boldsymbol{Z}) = E(\boldsymbol{Y}) + E(\boldsymbol{Z}) \tag{1.1}$$

and

$$E(\boldsymbol{AY}) = \boldsymbol{A}E(\boldsymbol{Y}) \quad \text{and} \quad E(\boldsymbol{AYB}) = \boldsymbol{A}E(\boldsymbol{Y})\boldsymbol{B}. \tag{1.2}$$

So

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{AY}) = \text{Cov}(\boldsymbol{AY}) = \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T. \tag{1.3}$$

Example 1.1. Consider the OLS model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where the e_i are iid with mean 0 and variance σ^2. Then $\boldsymbol{Y}$ and $\boldsymbol{e}$ are random vectors while $\boldsymbol{a} = \boldsymbol{X\beta}$ is a constant vector. Notice that $E(\boldsymbol{e}) = \boldsymbol{0}$. Thus

$$E(\boldsymbol{Y}) = \boldsymbol{X\beta} + E(\boldsymbol{e}) = \boldsymbol{X\beta}.$$

Since the e_i are iid,

$$\text{Cov}(\boldsymbol{Y}) = \text{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}_n \tag{1.4}$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. This result makes sense because the Y_i are independent with $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$. Hence $\text{VAR}(Y_i) = \text{VAR}(e_i) = \sigma^2$.

Recall that $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Hence

$$E(\hat{\boldsymbol{\beta}}_{OLS}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{Y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X\beta} = \boldsymbol{\beta}.$$

That is, $\hat{\boldsymbol{\beta}}_{OLS}$ is an unbiased estimator of $\boldsymbol{\beta}$. Using (1.3) and (1.4),

$$\text{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Cov}(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$$= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Recall that $\hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$. Hence

$$E(\hat{\boldsymbol{Y}}_{OLS}) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{Y}) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} = E(\boldsymbol{Y}).$$

Using (1.3) and (1.4),

$$\text{Cov}(\hat{\boldsymbol{Y}}_{OLS}) = \boldsymbol{H}\text{Cov}(\boldsymbol{Y})\boldsymbol{H}^T = \sigma^2\boldsymbol{H}$$

since $\boldsymbol{H}^T = \boldsymbol{H}$ and $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$.

Recall that the vector of residuals $\boldsymbol{r}_{OLS} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_{OLS}$. Hence $E(\boldsymbol{r}_{OLS}) = E(\boldsymbol{Y}) - E(\hat{\boldsymbol{Y}}_{OLS}) = E(\boldsymbol{Y}) - E(\boldsymbol{Y}) = \boldsymbol{0}$. Using (1.3) and (1.4),

$$\text{Cov}(\hat{\boldsymbol{r}}_{OLS}) = (\boldsymbol{I} - \boldsymbol{H})\text{Cov}(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{H})^T = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

since $\boldsymbol{I} - \boldsymbol{H}$ is symmetric and idempotent: $(\boldsymbol{I} - \boldsymbol{H})^T = \boldsymbol{I} - \boldsymbol{H}$ and $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{I} - \boldsymbol{H}$.

1.2 THE NO INTERCEPT OLS MODEL

The following results hold for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, but output tends to give this result if the "no intercept" option is selected. The no intercept model is also know as regression through the origin, and assumes that $\boldsymbol{X}$ does not contain a column of ones. For the no intercept model, the assumption $E(\boldsymbol{e}) = \boldsymbol{0}$ is important, and this assumption is rather strong.

Many of the multiple linear regression (MLR) results still hold: $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, the vector of *predicted fitted values* $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ provided the inverse exists, and the vector of residuals is $\boldsymbol{r} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$. The forward response plot of $\widehat{\boldsymbol{Y}}$ vs $\boldsymbol{Y}$ and residual plot of $\widehat{\boldsymbol{Y}}$ vs $\boldsymbol{r}$ are made in the same way and should be made before performing inference.

The main difference in the output is the ANOVA table. The ANOVA F tests $Ho : \beta_2 = \cdots = \beta_p = 0$. The test in this section tests $Ho : \beta_1 = \cdots = \beta_p = 0 \equiv Ho : \boldsymbol{\beta} = \boldsymbol{0}$. The following definition and test follows Guttman (1982, p. 147) closely.

Definition 1.2.1. Assume that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where the e_i are iid. Assume that it is desired to test $Ho : \boldsymbol{\beta} = \boldsymbol{0}$ versus $Ha : \boldsymbol{\beta} \neq \boldsymbol{0}$.

a) The *uncorrected total sum of squares*

$$SST = \sum_{i=1}^{n} Y_i^2. \tag{1.5}$$

b) The *model sum of squares*

$$SSM = \sum_{i=1}^{n} \hat{Y}_i^2. \tag{1.6}$$

c) The residual sum of squares or *error sum of squares* is

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{1.7}$$

d) The degrees of freedom (df) for SST is p, the df for SSM is $n - p$ and the df for SST is n. The mean squares are MSE $=$ SSE$/(n-p)$ and MSE $=$ SSE$/p$.

The ANOVA table given for the “no intercept” or “intercept = F” option is below.

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Model	p	SSM	MSM	Fo=MSM/MSE	for Ho:
Residual	n-p	SSE	MSE		$\boldsymbol{\beta} = \mathbf{0}$

The 4 step ANOVA F test for $\boldsymbol{\beta} = \mathbf{0}$:

i) State the hypotheses Ho: $\boldsymbol{\beta} = \mathbf{0}$, Ha: $\boldsymbol{\beta} \neq \mathbf{0}$.

ii) Find the test statistic $F_o = MSM/MSE$ or obtain it from output.

iii) Find the p–value from output or use the F–table: p–value =

$$P(F_{p,n-p} > F_o).$$

iv) State whether you reject Ho or fail to reject Ho. If Ho is rejected, conclude that there is an MLR relationship between Y and the predictors x_1, ..., x_p. If you fail to reject Ho, conclude that there is not a MLR relationship between Y and the predictors x_1, ..., x_p.

Warning: Several important models can be cast in the no intercept OLS form, but often a different test than $Ho : \boldsymbol{\beta} = \mathbf{0}$ is desired. For example, when the generalized or weighted least squares models are transformed into no intercept OLS form, the test of interest is Ho: $\beta_2 = \cdots = \beta_p = 0$. The one way ANOVA model is equivalent to the cell means model, which is in no intercept OLS form, but the test of interest is $Ho : \beta_1 = \cdots = \beta_p$.

Proposition 1.2.2. Suppose $Y = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ may or may not contain a column of ones. Then the change in Sum of Squares (SS) test can be used for inference. See section 1.4.

Example 1.2.1. Consider the Gladstone (1905-6) data described in Olive (2007). If the file of datasets `robdata` is down loaded into *R/Splus*, then the ANOVA F statistic for testing $\beta 2 = \cdots = \beta 4 = 0$ can be found with the following commands. The command *lsfit* adds a column of ones to x which contains the variables *size*, *sex*, *breadth* and *circumference*. Three of these predictor variables are head measurements. Then response Y is *brain weight*.

```
 > y <- cbrainy
 > x <- cbrainx[,c(11,10,3,6)]
 > ls.print(lsfit(x,y))
F-statistic (df=4, 262)=196.2433
```

The ANOVA F test can also be found with the no intercept model by adding a column of ones to x and then performing the change in SS test with the full model and the reduced model that only uses the column of ones. Notice that the "intercept=F" option needs to be used to fit both models. The residual standard error = RSE = $\sqrt{MSE}$. Thus SSE = $(n-k)(RSE)^2$ where $n-k$ is the denominator degrees of freedom for the F test and k is the numerator degrees of freedom = number of variables in the model. The column of ones *xone* is counted as a variable. The last line of output computes the change in SS F statistic and is again ≈ 196.24.

```
 > xone <- 1 + 0*1:267

 > x <- cbind(xone,x)

 > ls.print(lsfit(x,y,intercept=F))

Residual Standard Error=82.9175

F-statistic (df=5, 262)=12551.02


         Estimate  Std.Err t-value Pr(>|t|)

xone     99.8495 171.6189  0.5818   0.5612

size      0.2209   0.0358  6.1733   0.0000

sex      22.5491  11.2372  2.0066   0.0458

breadth  -1.2464   1.5139 -0.8233   0.4111

circum    1.0255   0.4719  2.1733   0.0307


 > ls.print(lsfit(x[,1],y,intercept=F))

Residual Standard Error=164.5028

F-statistic (df=1, 266)=15744.48


   Estimate Std.Err t-value Pr(>|t|)

X 1263.228 10.0674 125.477        0


 > ((266*(164.5028)^2 - 262*(82.9175)^2)/4)/(82.9175)^2

[1] 196.2435
```

1.3 GLS, WLS AND FGLS

Definition 1.3.1. Suppose that the response variable and at least one of the predictor variables is quantitative. Then the generalized least squares (GLS) model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.8}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Also $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{V}$ where $\boldsymbol{V}$ is a known $n \times n$ positive definite matrix.

Definition 1.3.2. The GLS estimator

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{Y}. \tag{1.9}$$

The fitted values are $\hat{\boldsymbol{Y}}_{GLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS}$.

Definition 1.3.3. Suppose that the response variable and at least one of the predictor variables is quantitative. Then the weighted least squares (WLS) model with weights $w_1, ..., w_n$ is the special case of the GLS model where $\boldsymbol{V}$ is diagonal: $\boldsymbol{V} = \mathrm{diag}(\mathrm{v}_1, ..., \mathrm{v}_\mathrm{n})$ and $w_i = 1/v_i$. Hence

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.10}$$

$E(\boldsymbol{e}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \mathrm{diag}(\mathrm{v}_1, ..., \mathrm{v}_\mathrm{n}) = \sigma^2 \mathrm{diag}(1/\mathrm{w}_1, ..., 1/\mathrm{w}_\mathrm{n})$.

Definition 1.3.4. The WLS estimator

$$\hat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{Y}. \tag{1.11}$$

The fitted values are $\hat{\boldsymbol{Y}}_{WLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{WLS}$.

Definition 1.3.5. The feasible generalized least squares (FGLS) model is the same as the GLS estimator except that $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is a function of an unknown $q \times 1$ vector of parameters $\boldsymbol{\theta}$. Let the estimator of $\boldsymbol{V}$ be $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$. Then the FGLS estimator

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\boldsymbol{X}^T\hat{\boldsymbol{V}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y}. \tag{1.12}$$

The fitted values are $\hat{\boldsymbol{Y}}_{FGLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{FGLS}$. The feasible weighted least squares (FWGLS) estimator is the special case of the FGLS estimator where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is diagonal. Hence the estimated weights $\hat{w}_i = 1/\hat{v}_i = 1/v_i(\hat{\boldsymbol{\theta}})$. The FWLS estimator and fitted values will be denoted by $\hat{\boldsymbol{\beta}}_{FWLS}$ and $\hat{\boldsymbol{Y}}_{FWLS}$, respectively.

Notice that the ordinary least squares (OLS) model is a special case of GLS with $\boldsymbol{V} = \boldsymbol{I}_n$, the $n \times n$ identity matrix. It can be shown that the GLS estimator minimizes the GLS criterion

$$Q_{GLS}(\boldsymbol{\eta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta})^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}).$$

Notice that the FGLS and FWLS estimators have $p + q + 1$ unknown parameters. These estimators can perform very poorly if $n < 10(p + q + 1)$.

The GLS and WLS estimators can be found from the OLS regression (without an intercept) of a transformed model. Typically there will be a constant in the model: the first column of $\boldsymbol{X}$ is a vector of ones. Following Seber and Lee (2003, p. 66-68), there is a nonsingular $n \times n$ matrix $\boldsymbol{K}$ such that $\boldsymbol{V} = \boldsymbol{K}\boldsymbol{K}^T$. Let $\boldsymbol{Z} = \boldsymbol{K}^{-1}\boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{K}^{-1}\boldsymbol{X}$ and $\boldsymbol{\epsilon} = \boldsymbol{K}^{-1}\boldsymbol{e}$.

Proposition 1.3.1. a)

$$\boldsymbol{Z} = \boldsymbol{U\beta} + \boldsymbol{\epsilon} \tag{1.13}$$

follows the OLS model since $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$.

b) The GLS estimator $\hat{\boldsymbol{\beta}}_{GLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$.

c) For WLS, $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$. The corresponding OLS model $\boldsymbol{Z} = \boldsymbol{U\beta} + \boldsymbol{\epsilon}$ is equivalent to $Z_i = \boldsymbol{u}_i^T \boldsymbol{\beta} + \epsilon_i$ for $i = 1, ..., n$ where $\boldsymbol{u}_i^T$ is the ith row of $\boldsymbol{U}$. Then $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. Hence $\hat{\boldsymbol{\beta}}_{GLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}_i = \sqrt{w_i}\, Y_i$ on $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$.

Proof. a) $E(\boldsymbol{\epsilon}) = \boldsymbol{K}^{-1} E(\boldsymbol{e}) = \boldsymbol{0}$ and

$$\text{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{K}^{-1} \text{Cov}(\boldsymbol{e}) (\boldsymbol{K}^{-1})^T = \sigma^2 \boldsymbol{K}^{-1} \boldsymbol{V} (\boldsymbol{K}^{-1})^T$$

$$= \sigma^2 \boldsymbol{K}^{-1} \boldsymbol{K} \boldsymbol{K}^T (\boldsymbol{K}^{-1})^T = \sigma^2 \boldsymbol{I}_n.$$

Notice that OLS without an intercept needs to be used since $\boldsymbol{U}$ does not contain a vector of ones. The first column of $\boldsymbol{U}$ is $\boldsymbol{K}^{-1}\boldsymbol{1} \neq \boldsymbol{1}$.

b) Let $\hat{\boldsymbol{\beta}}_{ZU}$ denote the OLS estimator obtained by regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. Then

$$\hat{\boldsymbol{\beta}}_{ZU} = (\boldsymbol{U}^T \boldsymbol{U})^{-1} \boldsymbol{U}^T \boldsymbol{Z} = (\boldsymbol{X}^T (\boldsymbol{K}^{-1})^T \boldsymbol{K}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T (\boldsymbol{K}^{-1})^T \boldsymbol{K}^{-1} \boldsymbol{Y}$$

and the result follows since $\boldsymbol{V}^{-1} = (\boldsymbol{K}\boldsymbol{K}^T)^{-1} = (\boldsymbol{K}^T)^{-1} \boldsymbol{K}^{-1} = (\boldsymbol{K}^{-1})^T \boldsymbol{K}^{-1}$.

c) The result follows from b) if $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. But for WLS, $\boldsymbol{V} = \text{diag}(\text{v}_1, ..., \text{v}_\text{n})$ and hence $\boldsymbol{K} = \boldsymbol{K}^T = \text{diag}(\sqrt{\text{v}_1}, ..., \sqrt{\text{v}_\text{n}})$. Hence

$$\boldsymbol{K}^{-1} = \text{diag}(1/\sqrt{\text{v}_1}, ..., 1/\sqrt{\text{v}_\text{n}}) = \text{diag}(\sqrt{\text{w}_1}, ..., \sqrt{\text{w}_\text{n}})$$

and $\boldsymbol{Z} = \boldsymbol{K}^{-1}\boldsymbol{Y}$ has ith element $Z_i = \sqrt{w_i}\ Y_i$. Similarly, $\boldsymbol{U} = \boldsymbol{K}^{-1}\boldsymbol{X}$ has ith row $\boldsymbol{u}_i^T = \sqrt{w_i}\ \boldsymbol{x}_i^T$. QED

Remark 1.3.1. Standard software produces WLS output and the ANOVA F test and Wald t tests are performed using this output.

Remark 1.3.2. The FGLS estimator can also be found from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$. Similarly the FWLS estimator can be found from the OLS regression (without an intercept) of $Z_i = \sqrt{\hat{w}_i} Y_i$ on $\boldsymbol{u}_i = \sqrt{\hat{w}_i}\boldsymbol{x}_i$. But now $\boldsymbol{U}$ is a random matrix instead of a constant matrix. Hence these estimators are highly nonlinear. OLS output can be used for exploratory purposes, but the p–values are generally not correct.

Under regularity conditions, the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}$ when the GLS model holds, but $\hat{\boldsymbol{\beta}}_{GLS}$ should be used because it generally has higher efficiency.

Definition 1.3.6. Let $\hat{\boldsymbol{\beta}}_{ZU}$ be the OLS estimator from regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. The vector of fitted values is $\hat{\boldsymbol{Z}} = \boldsymbol{U}\hat{\boldsymbol{\beta}}_{ZU}$ and the vector of residuals is $\boldsymbol{r}_{ZU} = \boldsymbol{Z} - \hat{\boldsymbol{Z}}$. Then $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{GLS}$ for GLS, $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{FGLS}$ for FGLS, $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{WLS}$ for WLS and $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{FWLS}$ for FWLS. For GLS, FGLS, WLS and FWLS, a *forward response plot* is a plot of $\hat{Z}_i$ versus Z_i and a *residual plot* is a plot of $\hat{Z}_i$ versus $r_{ZU,i}$.

Notice that the residual and forward response plots are based on the OLS output from the OLS regression without intercept of $\boldsymbol{Y}$ on $\boldsymbol{U}$. If the model is good, then the plotted points in the forward response plot should follow the identity line

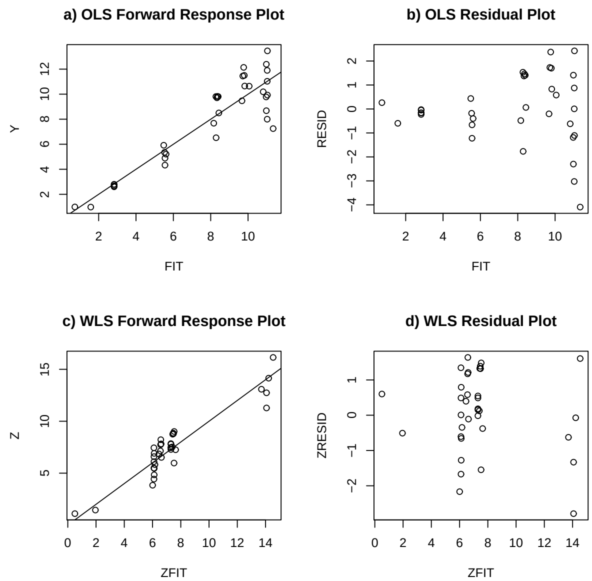

Figure 1.1. Plots for Draper and Smith Data

in an evenly populated band while the plotted points in the residual plot should follow the line $r_{ZU,i} = 0$ in an evenly populated band.

Plots based on $\hat{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{ZU}$ and on $r_i = Y_i - \hat{Y}_i$ should not be used for several reasons. First, often $\boldsymbol{V}$ depends on $\boldsymbol{X}$. Although the plot of $\hat{Y}_i$ versus Y_i should be linear, the plotted points will not scatter about the identity line in an evenly populated band. Hence this plot can not be used to check whether the GLS model with $\boldsymbol{V}$ is a good approximation to the data. Moreover, the r_i and $\hat{Y}_i$ may be correlated and usually do not scatter about the $r = 0$ line in an evenly populated band. The plots in Definition 1.3.6 are both a check on linearity and on whether the model using $\boldsymbol{V}$ (or $\hat{\boldsymbol{V}}$) gives a good approximation of the data, provided that $n > k(p + q + 1)$ where $k \geq 5$ and preferably $k \geq 10$.

For GLS and WLS (and for exploratory purposes for FGLS and FWLS), plots and model building and variable selection should be based on $\boldsymbol{Z}$ and $\boldsymbol{U}$. Form $\boldsymbol{Z}$ and $\boldsymbol{U}$ and then use OLS software for model selection and variable selection. If the columns of $\boldsymbol{X}$ are $\boldsymbol{v}_1, ..., \boldsymbol{v}_p$, then the columns of $\boldsymbol{U}$ are $\boldsymbol{U}_1, ..., \boldsymbol{U}_p$ where $\boldsymbol{U}_j = \boldsymbol{K}^{-1}\boldsymbol{v}_j$ corresponds to the jth predictor x_j. For example, the analog of the OLS residual plot of jth predictor versus the residuals is the plot of the jth predictor u_j versus $r_{ZU,i}$. The notation is confusing but the idea is simple: form $\boldsymbol{Z}$ and $\boldsymbol{U}$, then use OLS software and the OLS techniques to build the model.

Example 1.3.2. Draper and Smith (1981, p. 112-114) presents a FWLS example with $n = 35$ and $p = 2$. Hence $Y = \beta_1 + \beta_2 x + e$. Let $\hat{v}_i = v_i(\hat{\boldsymbol{\theta}}) = 1.5329 - 0.7334x_i + 0.0883x_i^2$. Thus $\hat{\boldsymbol{\theta}} = (1.5329, -0.7334, 0.0883)^T$. Figure 1.1a and b show the forward response and residual plots based on the OLS regression on Y on x. The residual plot has the shape of the right opening megaphone, suggesting that the variance is not constant. Figure 1.1c and d show the forward response and residual plots based on FWLS with weights $\hat{w}_i = 1/\hat{v}_i$.. Software meant for WLS needs the weights. Hence FWLS can be computed using WLS software with the estimated weights, but the software may print WLS instead of FWLS, as in Figure 1.1c and d.

1.4 INFERENCE FOR GLS

Inference for the GLS model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ can be performed by using the change in SS F test for the equivalent no intercept OLS model $\boldsymbol{Z} = \boldsymbol{U\beta} + \boldsymbol{\epsilon}$. Following Section 1.3, create $\boldsymbol{Z}$ and $\boldsymbol{U}$, fit the full and reduced model using the "no intercept" or "intercept = F" option.

The 4 step change in SS F test of hypotheses: i) State the hypotheses Ho: the reduced model is good Ha: use the full model

ii) Find the test statistic $F_R =$

$$\left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F)$$

iii) Find the p–value = $\mathrm{P}(F_{df_R - df_F, df_F} > F_R)$. (Here $df_R - df_F = p - q =$ number of parameters set to 0, and $df_F = n - p$.)

iv) State whether you reject Ho or fail to reject Ho. Reject Ho if the p–value $< \delta$ and conclude that the full model should be used. Otherwise, fail to reject Ho and conclude that the reduced model is good.

Assume that the GLS model contains a constant β_1. The GLS ANOVA F test of $Ho : \beta_2 = \cdots = \beta_p$ versus Ha: not Ho uses the reduced model that contains the first column of $\boldsymbol{U}$. The GLS ANOVA F test of $Ho : \beta_i = 0$ versus $Ho : \beta_i \neq 0$ uses the reduced model with the ith column of $\boldsymbol{U}$ deleted. For the special case of WLS, the software will often have a `weights` option that will also give correct output for inference.

Example 1.4.1. Suppose that the data from Example 1.3.2 has valid weights, so that WLS can be used instead of FWLS. The *R/Splus* commands below perform WLS.

```
 > ls.print(lsfit(dsx,dsy,wt=dsw))
Residual Standard Error=1.137
```

```
R-Square=0.9209

F-statistic (df=1, 33)=384.4139

p-value=0


            Estimate Std.Err t-value Pr(>|t|)

Intercept  -0.8891  0.3004 -2.9602   0.0057

X           1.1648  0.0594 19.6065   0.0000
```

The F statistic 886.4982 tests $Ho : \boldsymbol{\beta} = \boldsymbol{0}$ and is not of interest. The WLS ANOVA F test for $Ho : \beta_2 = 0$ can also be found with the no intercept model by adding a column of ones to x, form $\boldsymbol{U}$ and $\boldsymbol{Z}$ and compute the change in SS F test where the reduced model uses the first column of $\boldsymbol{U}$. Notice that the "intercept=F" option needs to be used to fit both models. The residual standard error = RSE = $\sqrt{MSE}$. Thus SSE = $(n-k)(RSE)^2$ where $n-k$ is the denominator degrees of freedom for the F test and k is the numerator degrees of freedom = number of variables in the model. The column of ones *xone* is counted as a variable. The last line of output computes the change in SS F statistic and is again ≈ 384.4.

```
 > xone <- 1 + 0*1:35

 >  x <- cbind(xone,dsx)

 > z <- as.vector(diag(sqrt(dsw))%*%dsy)

 >  u <- diag(sqrt(dsw))%*%x

 > ls.print(lsfit(u,z,intercept=F))

Residual Standard Error=1.137
```

```
R-Square=0.9817

F-statistic (df=2, 33)=886.4982

p-value=0


      Estimate Std.Err t-value Pr(>|t|)

xone  -0.8891  0.3004 -2.9602   0.0057

dsx    1.1648  0.0594 19.6065   0.0000

 > ls.print(lsfit(u[,1],z,intercept=F))

Residual Standard Error=3.9838

R-Square=0.7689

F-statistic (df=1, 34)=113.1055

p-value=0


   Estimate Std.Err t-value Pr(>|t|)

X   4.5024  0.4234 10.6351        0


 > ((34*(3.9838)^2-33*(1.137)^2)/1)/(1.137)^2

[1] 384.4006
```

The WLS t-test for this data has $t = 19.6065$ which corresponds to $F = t^2 = 384.4$ since this test is equivalent to the WLS ANOVA F test when there is only one predictor. The WLS t-test for the intercept has $F = t^2 = 8.76$. This test statistic can be found from the no intercept OLS model by leaving the first column of $\boldsymbol{U}$ out of the model, then perform the change in SS F test as shown below.

```
 > ls.print(lsfit(u[,2],z,intercept=F))

Residual Standard Error=1.2601

F-statistic (df=1, 34)=1436.300


   Estimate Std.Err t-value Pr(>|t|)

X   1.0038  0.0265 37.8985        0


 > ((34*(1.2601)^2-33*(1.137)^2)/1)/(1.137)^2

[1] 8.760723
```

A problem with GLS and FGLS is that the weigths can cause outliers. Weighted least squares (WLS) regression compensates for violation of the homoscedasticity assumption by weighting cases differentially: cases with large variances count less and those with small variances count more in estimating the regression coefficients. The result is that the estimated coefficients are usually very close to what they would be in OLS regression, but under WLS regression their standard errors are much smaller. Weighted predicted/residual plots can be used to assess the goodness of fit of the weighted model. That is, the WLS fit is plotted on the x axis and the WLS residual on the y axis. When there is good fit, the residuals will no longer form a funnel shape but instead be uniformly distributed around the 0.0 line of the y axis.

Figures 1.2 and 1.3 provides evidence of a Good Plot (Figure 1.2), and a Bad plot (Figure 1.3) for FWLS. Observe that for Figure 1.2a and b the residual plot has the shape of the right opening megaphone, suggesting that the variance is not constant. Figure 1.2.c and d show the forward response and residual plots based on FWLS, and the plots looks good. Software meant for WLS needs the weights. Hence FWLS was computed using WLS software with the stimated weights. In Figure 1.3.a the OLS plots look better than the FWLS plots. Notice the FWLS plot has a left opening megaphone shape with one outlier.

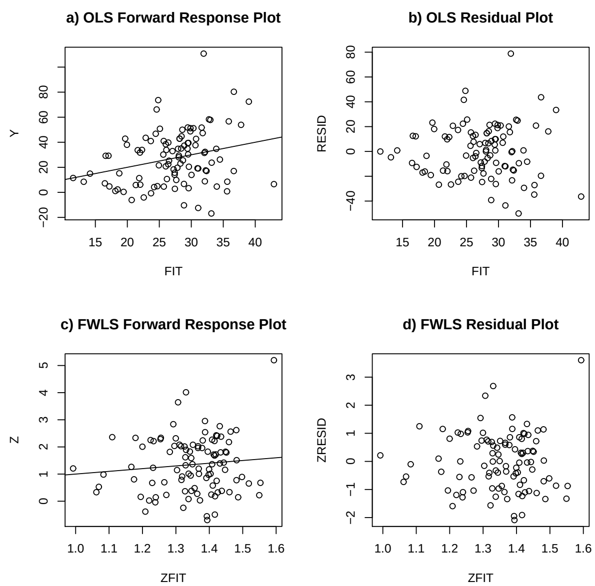

Figure 1.2. Good Plots for WLS

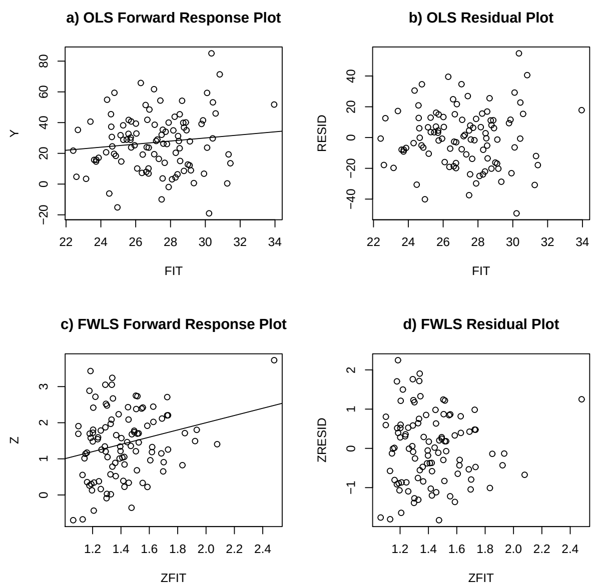

Figure 1.3. Bad Plots for WLS

CHAPTER 2

ONE WAY ANOVA

2.1 FIXED EFFECTS ONE WAY ANOVA

Definition 2.1.1. Models in which the response variable Y is quantitative, but all of the predictor variables are qualitative are called *analysis of variance* (ANOVA) models. Each combination of the levels of the predictors gives a different distribution for Y. A predictor variable W is often called a factor and a factor level a_i is one of the categories W can take.

Definition 2.1.2. Let $f_Z(z)$ be the pdf of Z. Then the family of pdf's $f_Y(y) = f_Z(y - \mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is the *location family* for the random variable $Y = \mu + Z$ with *standard pdf* $f_Z(z)$.

Definition 2.1.3. A one way fixed effects ANOVA model has a single qualitative predictor variable W with p categories a_1, ..., a_p. There are p different distributions for Y, one for each category a_i. The distribution of

$$Y|W = a_j \sim f_Z(y - \mu_j)$$

where the location family has second moments. Hence all p distributions come from the same location family with different location parameter μ_j and the same variance σ^2.

Definition 2.1.4. The one way fixed effects normal ANOVA model is the special case where

$$Y|W = a_j \sim N(\mu_j, \sigma^2).$$

Example 2.1.1. The pooled 2 sample t–test is a special case of a one way ANOVA model with $p = 2$. For example, one population could be ACT scores for men and the second population ACT scores for women. Then W = *gender* and Y = *score.*

Notation. It is convenient to relabel the response variable $Y_1, ..., Y_n$ as the vector $\boldsymbol{Y} = (Y_{11}, ..., Y_{1,n_1}, Y_{21}, ..., Y_{2,n_2}, ..., Y_{p1}, ..., Y_{p,n_p})^T$ where the Y_{ij} are independent and $Y_{i1}, ..., Y_{i,n_i}$ are iid. Here $j = 1, ..., n_i$ where n_i is the number of cases from the ith level where $i = 1, ..., p$. Thus $n_1 + \cdots + n_p = n$. Similarly use double subscripts on the errors. Then there will be many equivalent parameterizations of the one way fixed effects ANOVA model.

Definition 2.1.5. The *cell means model* is the parameterization of the one way fixed effects ANOVA model such that

$$Y_{ij} = \mu_i + e_{ij}$$

where Y_{ij} is the value of the response variable for the jth trial of the ith factor level. The μ_i are the unknown means and $E(Y_{ij}) = \mu_j$. The e_{ij} are iid from the location family with pdf $f_Z(z)$ and unknown variance $\sigma^2 = \text{VAR}(Y_{ij}) = \text{VAR}(e_{ij})$. For the normal cell means model, the e_{ij} are iid $N(0, \sigma^2)$ for $i = 1, ..., p$ and $j = 1, ..., n_i$.

The cell means model is an OLS model (without intercept) of the form:

$\boldsymbol{Y} = \boldsymbol{X}_c\boldsymbol{\beta}_c + \boldsymbol{e} =$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1,n_1} \\ e_{21} \\ \vdots \\ e_{2,n_2} \\ \vdots \\ e_{p,1} \\ \vdots \\ e_{p,n_p} \end{bmatrix}. \qquad (2.1)$$

Notation. Let $Y_{i0} = \sum_{j=1}^{n_i} Y_{ij}$ and let

$$\hat{\mu}_i = \overline{Y}_{io} = Y_{i0}/n_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}. \qquad (2.2)$$

Hence the "dot notation" means sum over the subscript corresponding to the 0, eg j. Similarly, $Y_{00} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}.

Notice that the indicator variables used in the OLS cell means model (2.1) are $x_i = 1$ if $W = a_i$, and $x_i = 0$, otherwise, for $j = 1, ..., p$. The model can use p indicator variables for the factor instead of $p - 1$ indicator variables because the model does not contain an intercept. Also notice that

$$E(\boldsymbol{Y}) = \boldsymbol{X}_c\boldsymbol{\beta}_c = (\mu_1, ..., \mu_1, \mu_2, ..., \mu_2, ..., \mu_p, ..., \mu_p)^T,$$

$(\boldsymbol{X}_c^T\boldsymbol{X}_c) = \mathrm{diag(n_1, ..., n_p)}$ and $\boldsymbol{X}_c^T\boldsymbol{Y} = (Y_{10}, ..., Y_{10}, Y_{20}, ..., Y_{20}, ..., Y_{p0}, ..., Y_{p0})^T$.

Hence $(\boldsymbol{X}_c^T\boldsymbol{X}_c)^{-1} = \text{diag}(1/\text{n}_1, ..., 1/\text{n}_\text{p})$ and

$$\hat{\boldsymbol{\beta}}_c = (\boldsymbol{X}_c^T\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c^T\boldsymbol{Y} = (\overline{Y}_{10}, ..., \overline{Y}_{p0})^T = (\hat{\mu}_1, ..., \hat{\mu}_p)^T.$$

Thus $\hat{\boldsymbol{Y}} = \boldsymbol{X}_c\hat{\boldsymbol{\beta}}_c = (\overline{Y}_{10}, ..., \overline{Y}_{10}, ...\overline{Y}_{p0}, ..., \overline{Y}_{p0})^T$. Hence the ijth fitted value is

$$\hat{Y}_{ij} = \overline{Y}_{i0} = \hat{\mu}_i \tag{2.3}$$

and the ijth residual is

$$r_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i. \tag{2.4}$$

Since the cell means model is an OLS model, there is an associated forward response plot and residual plot and ANOVA F test. However, many of the interpretations of the OLS quantities for ANOVA models differ from the interpretations for multiple linear regresion (MLR) models. First, for MLR models with continuous predictors $\boldsymbol{x}$, the conditional distribution $Y|\boldsymbol{x}$ makes sense even if $\boldsymbol{x}$ is not one of the observed $\boldsymbol{x}_i$ if $\boldsymbol{x}$ is not far from the $\boldsymbol{x}_i$. This fact makes MLR very powerful. But for the one way fixed effects ANOVA model, the only distributions that make sense are $Y|\boldsymbol{x}_i$ where $\boldsymbol{x}_i$ corresponds to observed combinations of levels. Hence for one way fixed effects ANOVA, there are p $\boldsymbol{x}_i$ that make sense where $\boldsymbol{x}_i$ corresponds to level a_i.

All of the parameterizations of the one way fixed effects ANOVA model yield the same predicted values, residuals and ANOVA F test, but the interpretations of the parameters differ.

Definition 2.1.6. Consider the one way fixed effects ANOVA model. The *forward response plot* is a plot of $\hat{\mu}_i$ versus Y_{ij} and the *residual plot* is a plot of $\hat{\mu}_i$

versus r_{ij}.

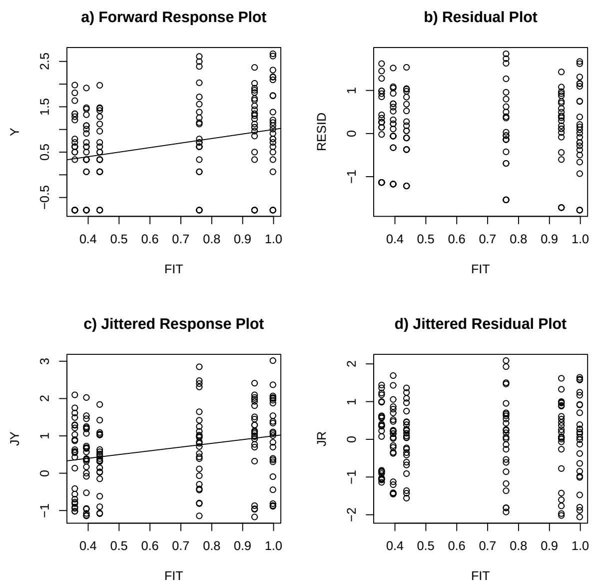

Figure 2.1. Plots for Crab Data

The points in the forward response plot scatter about the identity line and the points in the residual plot scatter about the $r = 0$ line, but the scatter need not be in an evenly populated band. A *dot plot* of $Z_1, ..., Z_m$ consists of an axis and m points each corresponding to the value of Z_i. The forward response plot consists of p dot plots, one for each value of $\hat{\mu}_i$. The dot plot corresponding to $\hat{\mu}_i$ is the dot plot of $Y_{i1}, ..., Y_{i,n_i}$. Similarly, the residual plot consists of p dot plots, and the plot corresponding to $\hat{\mu}_i$ is the dot plot of $r_{i1}, ..., r_{i,n_i}$. Assume that each $n_{ij} \geq 10$. Under the assumption that the Y_{ij} are from the same location scale family with different parameters μ_i, each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot. If the forward response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there is not much difference in the μ_i. If the identity line is clearly superior to any horizontal line, then at least some of the means differ.

The assumption of the Y_{ij} coming from the same location scale family with different location parameters μ_i and the same constant variance σ^2 is a big assumption and often does not hold. Another way to check this assumption is to make a box plot of the Y_{ij} for each i. The box in the box plot corresponds to the lower, middle and upper quartiles of the Y_{ij}. The middle quartile is just the sample median of the data m_{ij}: at least half of the $Y_{ij} \geq m_{ij}$ and at least half of the $Y_{ij} \leq m_{ij}$. The p boxes should be roughly the same length and the median should occur in roughly the same position. The "whiskers" in each plot should also be roughly similar. Histograms for each of the p samples could also be made. All of the histograms should look similar in shape.

Example 2.1.2. Kuehl (1994, p. 128) gives data for counts of hermit crabs on 25 different transects in each of six different coastline habitats. Let Z be the count. Then the response variable $Y = \log_{10}(Z + 1/6)$. Although the counts Z varied greatly, each habitat had several counts of 0 and often there were several counts of 1, 2 or 3. Hence Y is not continuous. The cell means model was fit with $n_j = 25$ for j = 1, ..., 6. Each of the six habitats was a level. Figure 2.1a and b shows the forward response plot and residual plot. There are 6 dot plots in each plot. Because several of the smallest values in each plot are identical, it does not always look like the identity line is passing through the six sample means $\overline{Y}_{i0}$ for

$i = 1, ..., 6$. In particular, examine the dot plot for the smallest mean (look at the 25 dots furthest to the left that fall on the vertical line FIT ≈ 0.36). Random noise (jitter) has been added to the response and residuals in Figure 2.1c and d. Now it is easier to compare the six dot plots. They seem to have roughly the same spread.

Definition 2.1.7. a) The *total sum of squares*

$$SSTO = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{00})^2.$$

b) The treatment sum of squares

$$SSTR = \sum_{i=1}^{p} n_i (\overline{Y}_{i0} - \overline{Y}_{00})^2.$$

c) The residual sum of squares or *error sum of squares*

$$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{io})^2.$$

Definition 2.1.8. Associated with each SS in Definition 2.1.7 is a degrees of freedom (df) and a mean square $= SS/df$. For SSTO, df $= n - 1$ and $MSTO = SSTO/(n-1)$. For SSTR, df $= p-1$ and $MSTR = SSR/(p-1)$. For SSE, $df = n-p$ and $MSE = SSE/(n-p)$.

Let $S_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i0})^2/(n_i - 1)$ be the sample variance of the ith group. Then the MSE is a weighted sum of the S_i^2:

$$\hat{\sigma}^2 = MSE = \frac{1}{n-p} \sum_{i=1}^{p} \sum_{j=1}^{n_i} e_{ij}^2 = \frac{1}{n-p} \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i0})^2 =$$

$$\frac{1}{n-p}\sum_{i=1}^{p}(n_i-1)S_i^2 = S_{pool}^2$$

where S_{pool}^2 is known as the pooled variance estimator.

The ANOVA table is the same as that for MLR, except that SSTR replaces the regression sum of squares. The MSE is again an estimator of σ^2. The ANOVA F test tests whether all p means μ_i are equal. Shown below is an ANOVA table given in symbols. Sometimes "Treatment" is replaced by "Between treatments," "Between Groups," "Model," "Factor" or "Groups." Sometimes "Error" is replaced by "Residual," or "Within Groups."

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Treatment	p-1	SSTR	MSTR	Fo=MSTR/MSE	for Ho:
Error	n-p	SSE	MSE		$\mu_1 = \cdots = \mu_p$

The 4 step fixed effects one way ANOVA F test of hypotheses:

i) State the hypotheses Ho: $\mu_1 = \mu_2 = \cdots = \mu_p$ and Ha: not Ho.

ii) Find the test statistic $F_o = MSTR/MSE$ or obtain it from output.

iii) Find the p–value from output or use the F–table: p–value =

$$P(F_{p-1,n-p} > F_o).$$

iv) State whether you reject Ho or fail to reject Ho. If the p–value $< \delta$, reject Ho and conclude that the mean response depends on the factor. Otherwise fail to reject Ho and conclude that the response does not depend on the factor. Give a nontechnical sentence.

Rule of thumb 2.1.2. Moore (1999, p. 512). If

$$\max(S_1, ..., S_p) \leq 2 \min(S_1, ..., S_p),$$

then the one way ANOVA F test results will be approximately correct.

Remark 2.1.1. When the assumption that the p groups come from the same location family with finite variance σ^2 is violated, the one way ANOVA F test may not make much sense because unequal means may not imply the superiority of one category over another. Suppose Y is the time in minutes until relief from a headache and that $Y_{1j} \sim N(60, 1)$ while $Y_{2j} \sim N(65, \sigma^2)$. If $\sigma^2 = 1$, then the type 1 medicine gives headache relief 5 minutes faster, on average, and is superior, all other things being equal. But if $\sigma^2 = 100$, then many patients taking medicine 2 experience much faster pain relief than those taking medicine 1, and many experience much longer time until pain relief. In this situation, predictor variables that would identify which medicine is faster for a given patient would be very useful.

```
fat1  fat2  fat3  fat4 One way Anova for Fat1 Fat2 Fat3 Fat4
64    78    75    55   Source     DF    SS     MS     F    P
72    91    93    66   treatment  3    1636.5 545.5  5.41 0.0069
68    97    78    49   error      20   2018.0 100.9
77    82    71    64
56    85    63    70
95    77    76    68
```

Example 2.1.3. The output above represents grams of fat (minus 100 grams) absorbed by doughnuts using 4 types of fat. See Snedecor and Cochran (1967, p. 259). Let μ_i denote the mean amount of fat i absorbed by doughnuts, $i = 1, 2, 3$ and 4. a) Find $\hat{\mu}_1$. b) Perform a 4 step Anova F test.

Solution: a) $\hat{\boldsymbol{\beta}}_{1c} = \hat{\mu}_1 = \overline{Y}_{10} = Y_{10}/n_1 = \sum_{j=1}^{n_1} Y_{1j}/n_1 =$

$(64 + 72 + 68 + 77 + 56 + 95)/6 = 432/6 = 72.$

b) i) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ H_a: not H_0

ii) F = 5.41

iii) pvalue = 0.0069

iv) Reject H_0, the mean amount of fat absorbed by doughnuts depends on the type of fat.

Three tests for $Ho : \mu_1 = \cdots = \mu_p$ can be used if Rule of Thumb 2.1.2: $\max(S_1, ..., S_p) \leq 2\min(S_1, ..., S_p)$ fails. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, and let $Y_{(1)} \leq Y_{(2)} \cdots \leq Y_{(n)}$ be the order statistics. Then the rank transformation of the response is $\boldsymbol{Z} = rank(\boldsymbol{Y})$ where $Z_i = j$ if $Y_i = Y_{(j)}$ is the jth largest order statistic. For example, if $\boldsymbol{Y} = (7.7, 4.9, 33.3, 6.6)^T$, then $\boldsymbol{Z} = (3, 1, 4, 2)^T$. The first test performs the one way ANOVA F test with $\boldsymbol{Z}$ replacing $\boldsymbol{Y}$. See Montgomery (1984, p. 117-118). The remaining two tests are described in Brown and Forsythe (1974b). Let $\lceil x \rceil$ be the smallest integer $\geq x$, eg $\lceil 7.7 \rceil = 8$. Then the Welch ANOVA F test uses test statistic

$$F_W = \frac{\sum_{i=1}^p w_i(\overline{Y}_{i0} - \tilde{Y}_{00})^2/(p-1)}{1 + \frac{2(p-2)}{p^2-1} \sum_{i=1}^p (1 - \frac{w_i}{u})^2/(n_i - 1)}$$

where $w_i = n_i/S_i^2, u = \sum_{i=1}^p w_i$ and $\tilde{Y}_{00} = \sum_{i=1}^p w_i \overline{Y}_{i0}/u$. The test statistic is compared to an F_{p-1,d_W} distribution where $d_W = \lceil f \rceil$ and

$$1/f = \frac{3}{p^2-1} \sum_{i=1}^{p} (1 - \frac{w_i}{u})^2/(n_i - 1).$$

The modified ANOVA F test uses test statistic

$$F_M = \frac{\sum_{i=1}^p n_i(\overline{Y}_{i0} - \overline{Y}_{00})^2}{\sum_{i=1}^p (1 - \frac{n_i}{n})S_i^2}$$

The test statistic is compared to an F_{p-1,d_M} distribution where $d_M = \lceil f \rceil$ and

$$1/f = \sum_{i=1}^{p} c_i^2/(n_i - 1)$$

where

$$c_i = (1 - \frac{n_i}{n})S_i^2/[\sum_{i=1}^{p} (1 - \frac{n_i}{n})S_i^2].$$

The `regpack` function *anovasim* can be used to compare the four test statistics.

Simulation.

Simulations were done in R. The simulation used different values of sample size (n_i), populations means (μ_i) , and standard deviation (σ_i), 5,000 runs were used at the nominal 5% level for the ANOVA F (F)test, modified ANOVA F (F_M) test, the Welch ANOVA F (F_W)test and ANOVA F residual (F_R) test. For Table 2.1, want level near .05 and for Table 2.3 want level near 1 unless $\mu_1 = \mu_2 = \mu_3 = \mu_4$ where power=level=0.05 is wanted. If $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$ the classical ANOVA F test should work.

As expected, in Table 2.1, the ANOVA F statistic shows marked deviations from from its nominal size when the variances of the groups are unequal. In the examples given at the nominal 5% size, the simulated size varies from 0.1428 (14.28%)

when the sample has larger variances to 0.031 (3.1%) when the same sample has smaller variances. Also in this case, the other three statistics show relatively little fluctuation in size. F_W had size closest to .05 except for small sample sixes where F_M had a slight advantage. F_R was not as good as F_W or F_M. F_W performed well and F_R was OK, but F_M was often worse than the classical test.

Table 2.1. $\mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$ F is the proportion of times the ANOVA F test rejected Ho with nominal level 0.05, 5000 runs

n_1, n_2, n_3, n_4	$\sigma_1, \sigma_2, \sigma_3, \sigma_4$	F	F_M	F_W	F_R
4, 4, 4, 4	1, 1, 1, 1	0.0516	0.0412	0.0554	0.0642
	1, 2, 2, 3	0.0656	0.0506	0.0606	0.0734
	3, 2, 2, 1	0.0628	0.0472	0.0628	0.064
4, 8, 10, 12	1, 1, 1, 1	0.0464	0.0474	0.0604	0.0494
	1, 2, 2, 3	0.0246	0.0482	0.047	0.0304
	3, 2, 2, 1	0.133	0.0604	0.0604	0.0888
8, 16, 20, 24	1, 1, 1, 1	0.0568	0.0548	0.0584	0.055
	1, 2, 2, 3	0.031	0.065	0.057	0.04
	3, 2, 2, 1	0.1428	0.064	0.0568	0.0972
11, 11, 11, 11	1, 1, 1, 1	0.0518	0.051	0.0534	0.0526
	1, 2, 2, 3	0.0622	0.0576	0.0496	0.0582
11, 16, 16, 21	1, 1, 1, 1	0.054	0.0546	0.051	0.0518
	1, 2, 2, 3	0.038	0.06	0.0484	0.0416
	3, 2, 2, 1	0.0988	0.057	0.043	0.0752
22, 32, 32, 42	1, 1, 1, 1	0.0518	0.0504	0.0542	0.0494
	1, 2, 2, 3	0.0344	0.0592	0.0516	0.0426
	3, 2, 2, 1	0.0986	0.054	0.0492	0.0764
5, 5, 5, 5	1, 1, 1, 1	0.0532	0.0438	0.048	0.0556
	1, 2, 2, 3	0.0684	0.054	0.0596	0.0684
	3, 2, 2, 1	0.0644	0.0488	0.0592	0.0674
10, 10, 10, 10	1, 1, 1, 1	0.0542	0.051	0.053	0.0548
	1, 2, 2, 3	0.068	0.0614	0.0588	0.0632
	3, 2, 2, 1	0.067	0.0612	0.0546	0.0632
20, 20, 20, 20	1, 1, 1, 1	0.0498	0.0498	0.0478	0.0532
	1, 2, 2, 3	0.0688	0.0642	0.0558	0.0666
	3, 2, 2, 1	0.063	0.0612	0.0506	0.0616

Table 2.2. (Table 2.1 continued)

n_1, n_2, n_3, n_4	$\sigma_1, \sigma_2, \sigma_3, \sigma_4$	F	F_M	F_W	F_R
50, 50, 50, 50	1, 1, 1, 1	0.0506	0.0502	0.0488	0.0496
	1, 2, 2, 3	0.0612	0.0602	0.0472	0.0536
	3, 2, 2, 1	0.0636	0.0626	0.0488	0.0586
100, 100, 100, 100	1, 1, 1, 1	0.0438	0.0438	0.046	0.045
	1, 2, 2, 3	0.0638	0.063	0.0458	0.061
	3, 2, 2, 1	0.0592	0.0586	0.0518	0.0564

Table 2.3. F is the proportion of times the ANOVA F test rejected Ho with nominal level 0.05, 5000 runs

n_1, n_2, n_3, n_4	$\sigma_1, \sigma_2, \sigma_3, \sigma_4$	$\mu_1, \mu_2, \mu_3, \mu_4$	F	F_M	F_W	F_R
4, 4, 4, 4	1, 1, 1, 1	0, 0, 0, 0	0.0544	0.0416	0.0624	0.0634
		1, 0, 0, 0	0.2078	0.1696	0.178	0.2088
		1, 0, 0, 0.7	0.211	0.1738	0.1882	0.2222
		5, 0, 0, 0.5	1	1	0.9982	1
5, 5, 5, 5	1, 1, 1, 1	0, 0, 0, 0	0.0532	0.0444	0.0532	0.0584
		1, 0, 0, 0	0.2656	0.241	0.2344	0.2656
		1, 0, 0, 0.7	0.1502	0.1388	0.1424	0.1572
		5, 0, 0, 0.5	1	1	1	1
10, 10, 10, 10	1, 1, 1, 1	0, 0, 0, 0	0.0512	0.0492	0.0538	0.0544
		1, 0, 0, 0	0.5736	0.5656	0.533	0.5524
		1, 0, 0, 0.7	0.2992	0.2914	0.2732	0.2912
		5, 0, 0, 0.5	1	1	1	1
20, 20, 20, 20	1, 1, 1, 1	0, 0, 0, 0	0.052	0.0514	0.0544	0.0544
		1, 0, 0, 0	0.8998	0.8996	0.8856	0.884
		1, 0, 0, 0.7	0.5786	0.5774	0.561	0.557
		5, 0, 0, 0.5	1	1	1	1
100, 100, 100, 100	1, 1, 1, 1	0, 0, 0, 0	0.0446	0.0444	0.045	0.048
		1, 0, 0, 0	1	1	1	1
		1, 0, 0, 0.7	1	1	1	0.9998
		5, 0, 0, 0.5	1	1	1	1

Table 2.4. (Table 2.3 continued)

n_1, n_2, n_3, n_4	$\sigma_1, \sigma_2, \sigma_3, \sigma_4$	$\mu_1, \mu_2, \mu_3, \mu_4$	F	F_M	F_W	F_R
4, 4, 4, 4	3, 2, 2, 1	0, 0, 0, 0	0.0718	0.0544	0.0636	0.076
		1.5, 0, 0, 0	0.1476	0.11138	0.1046	0.1412
		0, 0, 0, 1	0.092	0.0662	0.1202	0.1046
		1.3, 0, 0, 1.3	0.131	0.098	0.1534	0.1476
5, 5, 5, 5	3, 2, 2, 1	0, 0, 0, 0	0.0692	0.0544	0.0566	0.0704
		1.5, 0, 0, 0	0.1818	0.1508	0.1172	0.151
		0, 0, 0, 1	0.104	0.0838	0.1366	0.1214
		1.3, 0, 0, 1.3	0.166	0.1352	0.1988	0.183
10, 10, 10, 10	3, 2, 2, 1	0, 0, 0, 0	0.066	0.0598	0.0558	0.0634
		1.5, 0, 0, 0	0.3126	0.2928	0.1946	0.2502
		0, 0, 0, 1	0.1464	0.132	0.279	0.1938
		1.3, 0, 0, 1.3	0.2946	0.2772	0.3542	0.4176
20, 20, 20, 20	3, 2, 2, 1	0, 0, 0, 0	0.0606	0.0578	0.0506	0.0572
		1.5, 0, 0, 0	0.5312	0.524	0.3676	0.4394
		0, 0, 0, 1	0.2718	0.2592	0.5774	0.3998
		1.3, 0, 0, 1.3	0.592	0.5792	0.7866	0.6852
100, 100, 100, 100	3, 2, 2, 1	0, 0, 0, 0	0.058	0.0566	0.0464	0.057
		1.5, 0, 0, 0	0.9976	0.9976	0.9862	0.9888
		0, 0, 0, 1	0.9834	0.9832	1	0.997
		1.3, 0, 0, 1.3	1	1	1	1

Table 2.5. (Table 2.3 continued)

n_1, n_2, n_3, n_4	$\sigma_1, \sigma_2, \sigma_3, \sigma_4$	$\mu_1, \mu_2, \mu_3, \mu_4$	F	F_M	F_W	F_R
4, 4, 4, 4	1, 2, 2, 3	0, 0, 0, 0	0.0704	0.0526	0.0678	0.0714
		1.3, 0, 0, 0	0.1166	0.0828	0.1652	0.1398
		0, 0, 0, 1	0.1026	0.077	0.0802	0.1038
		1, 0, 0, 1	0.1108	0.085	0.1176	0.1228
5, 5, 5, 5	1, 2, 2, 3	0, 0, 0, 0	0.0574	0.0452	0.055	0.058
		1.3, 0, 0, 0	0.2214	0.1058	0.2012	0.1614
		0, 0, 0, 1	0.1232	0.1006	0.0856	0.1104
		1, 0, 0, 1	0.1252	0.1016	0.1376	0.1374
10, 10, 10, 10	1, 2, 2, 3	0, 0, 0, 0	0.059	0.052	0.0522	0.0564
		1.3, 0, 0, 0	0.2214	0.2024	0.4548	0.3158
		0, 0, 0, 1	0.1732	0.1606	0.106	0.1466
		1, 0, 0, 1	0.1878	0.1712	0.2486	0.2116
20, 20, 20, 20	1, 2, 2, 3	0, 0, 0, 0	0.0662	0.0626	0.0492	0.059
		1.3, 0, 0, 0	0.4904	0.4756	0.8282	0.6628
		0, 0, 0, 1	0.281	0.2738	0.175	0.2268
		1, 0, 0, 1	0.3658	0.3562	0.5376	0.434
100, 100, 100, 100	1, 2, 2, 3	0, 0, 0, 0	0.0618	0.061	0.0516	0.0578
		1.3, 0, 0, 0	1	1	1	1
		0, 0, 0, 1	0.8758	0.8756	0.7636	0.7926
		1, 0, 0, 1	0.9904	0.9904	0.9992	0.9972

Table 2.6. (Table 2.3 continued)

n_1, n_2, n_3, n_4	$\sigma_1, \sigma_2, \sigma_3, \sigma_4$	$\mu_1, \mu_2, \mu_3, \mu_4$	F	F_M	F_W	F_R
4, 4, 4, 4	1, 1, 1, 9	0, 0, 0, 0	0.1486	0.0744	0.0588	0.1072
		1.3, 0, 0, 0	0.1586	0.0848	0.2478	0.11602
		0, 0, 0, 1	0.1566	0.0858	0.0626	0.1138
		1, 0, 0, 1	0.1384	0.078	0.1706	0.126
5, 5, 5, 5	1, 1, 1, 9	0, 0, 0, 0	0.1396	0.0764	0.0584	0.081
		1.3, 0, 0, 0	0.1596	0.088	0.3278	0.2076
		0, 0, 0, 1	0.1524	0.0902	0.059	0.0892
		1, 0, 0, 1	0.1422	0.0778	0.2142	0.1532
10, 10, 10, 10	1, 1, 1, 9	0, 0, 0, 0	0.122	0.08	0.0504	0.089
		1.3, 0, 0, 0	0.1378	0.0988	0.7194	0.489
		0, 0, 0, 1	0.132	0.0938	0.059	0.1024
		1, 0, 0, 1	0.1332	0.0956	0.475	0.3086
20, 20, 20, 20	1, 1, 1, 9	0, 0, 0, 0	0.1102	0.0924	0.048	0.0834
		1.3, 0, 0, 0	0.1642	0.1362	0.9784	0.9014
		0, 0, 0, 1	0.1536	0.1284	0.0642	0.1076
		1, 0, 0, 1	0.1516	0.1226	0.8568	0.6634
100, 100, 100, 100	1, 1, 1, 9	0, 0, 0, 0	0.1064	0.1034	0.054	0.0868
		1.3, 0, 0, 0	0.4568	0.4372	1	1
		0, 0, 0, 1	0.303	0.2984	0.1292	0.1928
		1, 0, 0, 1	0.3328	0.323	1	1

REFERENCES

[1] Brown, M. B. and Forsythe, A. B. (1974), "The Small Sample Behavior of Some Statistics Which Test the Equality of Several Means," *Biometrika*, **27**, 129-132.

[2] Draper, N.R., and Smith, H. (1981), *Applied Regression Analysis*, 2nd ed., Wiley, NY. p.112-114.

[3] Gladstone, R.J. (1905-1906), "A study of the Relation of the Brain to the Size of the Head," *Biometrika*, **4**, 105-123.

[4] Guttman, I. (1982), *Linear Models: an Introduction*, Wiley, NY.

[5] Kachigan, S. (1982), *Multivariate Statistical Analysis. A Conceptual introduction*, Radius, NY.

[6] Kuehl, R. (1994), *Statistical Principles of Research Design and Analysis*, Duxbury Press, Belmont, California.

[7] Montgomery, D.C. (1984), *Design and Analysis of Experiments,* 2nd ed., Wiley, NY.

[8] Moore, D.S. (2000), *The Basic Practice of Statistics,* 2nd ed., W.H. Freeman, NY.

[9] Olive. D.J. (2007), *Multiple Linear and 1D Regression Models*, Unpublished Online Text available from (www.math.siu.edu/olive/regbk.htm).

[10] Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, NY.

[11] Snedecor, G.W. and Cochran, W.G. (1967), *Statistical Methods*, Iowa State University Press, Ames, IA.

VITA

Graduate School
Southern Illinois University

Elmer Agudelo Rodríguez Date of Birth: September 2, 1970

1945 Evergreen Terrace Dr 5, Carbondale, IL 62901

Bachelor of Science, Mathematics and Physics Education, Antioquia University, Colombia (May 1994)

Graduate Certificate in Pedagogy Contemporaneous Tendencies, Pontificia Bolivariana University, Colombia (June 2000)

Special Honors and Awards: Three times Part of the Deans list undergraduate studies, and winner in 2002/2, Antioquia University, Colombia

Research paper title: Regression and ANOVA under Heterogeneity

Adviser: David Olive