STAT 371 (Winter 2013 - 1135)

Statistics for Businesses I

Prof. H. Fahmy
University of Waterloo

LaTeXer: W. KONG
http://wwkong.github.io
Last Revision: April 30, 2014

Table of Contents

These notes are currently a work in progress, and as such may be incomplete or contain errors.

ACKNOWLEDGMENTS:

Special thanks to *Michael Baker* and his LaTeX formatted notes. They were the inspiration for the structure of these notes.

Abstract

The purpose of these notes is to provide a secondary reference to the material covered in STAT 371. The official prerequisite for this course is STAT 231, but this author recommends that the student take a good course in linear algebra (such as MATH 136/146 and MATH 235/245) before enrolling in this course.

Personally, this author believes that this and STAT 443 are two of the most industry applicable courses at the University of Waterloo and would highly recommend this course to any mathematics student.

Errata

Dr. H. Fahmy

M3 2018

Office hours: T,Th @ 4-5pm

Midterm: Thursday, June 13th, 2013 @ 30% (2:30pm-4:30pm)

Assignments: 4 Assignments @ 5% each = 20%

Exam: Final Exam @ 50%

1 Review

Definition 1.1. Given a regression $Y_t = f(X_t)$, Y_t is called the *response variable* or *regressor*, and X_t is called the *explanatory variable* or *regressand*.

Definition 1.2. Here are the steps of model building:

1) Specification (define variables, gather data)

2) Estimation (MLE, Least Squares, GMM, Report/ANOVA)

3) Evaluation (inference)

4) Assessing the validity of your results

Definition 1.3. The error term is a random part of a regression model that accounts for all information that is not captured by the model. The presence of an error term indicates a stochastic formulation and the lack of one makes it a deterministic formulation.

Note 1. In the model $Y_t = \beta_0 + \beta_1 X_t + \mu_t$, Y_t and X_t are observed variables, β_0 and β_1 are true unknown parameters and μ_t is an unobserved error term.

1.1 Methods of Estimation

(Reviewed in the Tutorial and omitted here; the methods discussed were Least Squares and MLE for a simple linear regression).

2 The General Linear Regression Model (GLRM)

From this point forward, the author assumes that the reader has a good understanding of linear algebra.

Definition 2.1. We define the the GLRM as follows. Suppose that we have k explanatory variables (including the constant variable) and n equations (n is the number of observations) with $k < n$. Let X_{ab} be the b^{th} observation of the a^{th} variable, Y_t be the t^{th} observation, and β_t be the the t^{th} variable.

Define $Y = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_n \end{pmatrix}^t$, $U = \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_n \end{pmatrix}^t$, $\beta = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_n \end{pmatrix}^t$ and a matrix $X \in \mathbb{M}$ where the n^{th} row and m^{th} column entry is X_{mn} with $X_{1n} = 1$ for all n. That is, the l^{th} column is the vector of observations of the l^{th} explanatory variable.

The GLRM in a compactification is

1) "The true model": $Y = X\beta + U$

We also define:

2) "The estimated": $\hat{Y} = X\hat{\beta}$

3) "The residual": $\hat{U} = Y - \hat{Y}$

Note that $Y = X\hat{\beta} + \hat{U}$.

From the least squares method,

$$RSS = \sum_{i=1}^{n} \hat{\mu}_t^2 = \left\langle \hat{U}, \hat{U} \right\rangle$$

and we want to minimize RSS by changing $\hat{\beta}$ (ordinary least squares). Note that,

$$\begin{aligned} RSS &= \left\langle \hat{U}, \hat{U} \right\rangle \\ &= (Y - X\hat{\beta})^t (Y - X\hat{\beta}) \\ &= Y^t Y - Y' X\hat{\beta} - \hat{\beta}^t X^t Y + \hat{\beta}^t X^t X \hat{\beta} \\ &= Y^t Y - 2\hat{\beta}^t X^t Y + \hat{\beta}^t X^t X \hat{\beta} \end{aligned}$$

and using first order conditions, we want $\frac{\partial RSS}{\partial \hat{\beta}_{k\times 1}} = 0_{k\times 1}$ where

$$\frac{\partial RSS}{\partial \hat{\beta}_{k\times 1}} = -2X^t Y + 2X' X\hat{B} = 0_{k\times 1} \implies \hat{\beta}_{OLS} = (X^t X)^{-1} X^t Y$$

and note that the order of the variable in the denominator of the partial must match the result of the partial. The equation

$$\frac{\partial RSS}{\partial \hat{\beta}_{k\times 1}} = -2X^t Y + 2X' X\hat{B} = 0_{k\times 1}$$

is called the "normal equation".

Note that we assume that X is of rank n in order for $X^t X$ to be invertible since $null(A^t A) = null(A)$ by the *rank-nullity theorem*.

Example 2.1. In a simple regression, we have

$$X^t X = \begin{pmatrix} n & \sum x_{2t} \\ \sum x_{2t} & \sum x_{2t}^2 \end{pmatrix}, X^t Y = \begin{pmatrix} \sum Y_t & \sum X_{2t} Y_t \end{pmatrix}^t$$

Note that we also use the notation $x_t = (X_t - \bar{X})$, $y_t = (Y_t - \bar{Y})$ which we call deviation form.

Example 2.2. Consider the stochastic presentation of the Cobb-Douglas production function

$$Q_t = cL_t^{\alpha} K_t^{\beta} e^{\mu_t}$$

where μ_t is the error term. If we are given data from 2000 to 2010 per year, we are given 11 observation.

To model this we do the following:

1) (Estimation) Linearize the model [Log-log model]:

$$\ln Q_t = \ln c + \alpha \ln L_t + \beta \ln K_t + \mu_t$$

2) (Estimation) Re-parametrize: $Y_t = \ln Q_t$, $\beta_1 = \ln c$, $\beta_2 = \alpha$, $\ln L_t = X_{2t}$, $\beta_3 = \beta$, $X_{3t} = \ln K_t$ and so

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \mu_t$$

3) (Estimation) Calculate $\hat{B}_{OLS} = (X^t X)^{-1} X^t Y$ where $X \in \mathcal{M}^{11\times 3}(\mathbb{R})$ and $Y \in \mathcal{M}^{11\times 1}(\mathbb{R})$,

$$X^t X = \begin{pmatrix} n & \sum X_{2t} & \sum X_{3t} \\ \sum X_{2t} & \sum X_{2t}^2 & \sum X_{2t} X_{3t} \\ \sum X_{3t} & \sum X_{2t} X_{3t} & \sum X_{3t}^2 \end{pmatrix}_{3\times 3}$$

which is fairly difficult to invert. Instead we work with the deviation form

$$\begin{aligned} Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \mu_t \quad &\implies \quad \bar{Y} = \beta_1 + \beta_2 \bar{X}_{2t} + \beta_3 \bar{X}_{3t} + \bar{\mu} \\ &\implies \quad y_t = \beta_2 x_{2t} + \beta_3 x_{3t} + \mu_t \end{aligned}$$

This creates a new matrix form $y = x\beta' + U$ where y contains the $y_t's$, x contains the $x_t's$, U contains the $\mu_t's$ and β contains only β_2 and β_3. So we now have

$$x^t x = \begin{pmatrix} \sum x_{2t}^2 & \sum x_{2t}x_{3t} \\ \sum x_{2t}x_{3t} & \sum x_{3t}^2 \end{pmatrix}$$

which is easier to invert. Thus,

$$\hat{\beta}' = \begin{pmatrix} \sum x_{2t}^2 & \sum x_{2t}x_{3t} \\ \sum x_{2t}x_{3t} & \sum x_{3t}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum x_{2t}y_t \\ \sum x_{3t}y_t \end{pmatrix}$$

and we can deduce $\hat{\beta}_1$ using

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \beta_3 \bar{X}_3$$

Example 2.3. Let the true model be

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \mu_t, t = 1, 2, ..., 23$$

where Y_t is the log output, X_{2t} is the log labour output, and X_{3t} is the log capital output. The data given (in deviation form) is

$$\sum x_{2t}^2 = 12, \sum x_{3t}^2 = 12, \sum x_{2t}x_{3t} = 8, \sum x_{2t}y_t = 10, \sum x_{3t}y_t = 8, \sum y_t^2 = 10.$$

We want to estimate the model using the least squares estimate and explain the meaning using the estimated coefficient. The following is the solution.

$$y_t = \beta_2 x_{2t} + \beta_3 x_{3t} + \mu_t, n = 23, k = 2$$

where

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_2 & \hat{\beta}_3 \end{bmatrix}^t = \begin{pmatrix} 12 & 8 \\ 8 & 12 \end{pmatrix}^{-1} \begin{pmatrix} 10 \\ 8 \end{pmatrix} = \begin{pmatrix} 0.15 & -0.10 \\ -0.10 & 0.15 \end{pmatrix} \begin{pmatrix} 10 \\ 8 \end{pmatrix} = \begin{pmatrix} 0.7 \\ 0.2 \end{pmatrix}$$

and so $\hat{\beta}_2 = 0.7$, $\hat{\beta}_3 = 0.2$. Thus, our model is

$$\hat{Y}_t = \hat{\beta}_1 + 0.7X_{2t} + 0.2X_{3t}$$

and note that the betas are actually the X_t elasticities of Y_t. Here the A elasticity of B is given by

$$E_{AB} = \frac{dB}{dA} \cdot \frac{A}{B} = \frac{\%\triangle B}{\%\triangle A}$$

Summary 1. Here are a few relevant statistical models:

1) Log-log model: $\ln Y_t = \beta_1 + \beta_2 \ln X_t$

2) Semi-log model: $\ln Y_t = \beta_1 + \beta_2 X_t$

3) Linear model: $Y_t = \beta_1 + \beta_2 X_t$

4) Growth models: $\triangle \ln Y_t = \beta_1 + \beta_2 \triangle \ln X_t$

(For Midterm Review: p. 1-10 (Chapter 1- Simple linear regression), deviation form, p. 57-61 (Chapter 2: GLRM))

Summary 2. Recall the normal equations that are determined by the condition

$$\frac{\partial RSS}{\partial \hat{\beta}} = 0$$

which produces the (normal) equations

$$\begin{aligned} X^t X\hat{\beta} &= X^t Y \\ \hat{\beta}_{OLS} &= (X^t X)^{-1} X^t Y \end{aligned}$$

where $RSS = \hat{U}^t \hat{U}$.

Definition 2.2. We say that A and B are orthogonal ($A \perp B$) when

$$A^t B = B^t A = 0$$

Remark 2.1. Note that $X \perp \hat{U}$ since $\hat{\beta}X$ is a projection of $Y = \hat{\beta}X + \hat{U} = \beta X + U$ onto the column space of X with orthogonal component $\hat{U}$. This also can be shown using the above normal equations.

Corollary 2.1. *(The following are found in p. 61-69 in the course book)*

(1) $\hat{\beta}$ is unique.

(2) You can find $\hat{Y} = X\hat{\beta}$ by projecting Y onto the column space of X.

Remark 2.2. Any idempotent and symmetric matrix is a projection matrix and for any idempotent matrix, its rank is equal to its trace. Using this, note that the linear operator $M = (I - \text{Proj}_X) = (I - X^t(X^tX)^{-1}X)$ applied to U produces $\hat{U}$. [Called result # 11]

2.1 The Classicial Assumptions of the GLRM

Gauss-Markov Assumptions

1. The model is true as specified (below are some examples)
   (a) Over identification or adding irrelevant variables (too many variables)
   (b) Underfitting or omitting a relevant variable (too few variables)
   (c) Wrong functional form (e.g. linear model instead of log-linear model)
2. The X's are non-stochastic in repeated sampling; X is treated as constant; if not satisfied, this could indicate a sampling problem
3. The model is linear in the parameters and the error term; it is a linear function in the polynomial ring with coefficients span{Parameters,Error}
4. X^tX is of full rank; nullity$(A^t) = 0$; no multi-collinearity
5. Assumptions related to the disturbance term $U_{n\times 1}$; if satisfied, the error is said to be *white noise*:
   (a) If assumption 1. is satisfied, then $E[U] = 0_{n\times 1}$; $E[U|X] = 0_{n\times 1}$
   (b) Homoskedastic error term; $Var[U] = \sigma^2_{n\times 1}$
   (c) No serial correlation between the errors; $Cov(\mu_t, \mu_s) = 0$, $\forall t \neq s$

Notation 1. We first define notation for the assumptions for simple regression

- (5a) $E[u_t|x_{1t}, ..., x_{kt}] = E[u_t] = 0$
- (5b) $Var[u_t] = E[u_t - 0]^2 = E[u_t^2] = \sigma_u^2$
- (5c) $Cov[u_t, u_s] = E[(u_t - 0)(u_s - 0)] = E[u_t u_s] = 0, s \neq t$

And now for general regression (matrix form):

- (5a) $E[U_{n\times 1}] = 0_{n\times 1}$
- (5b) $Var[U] = E\left[(U - E[U])^2\right] = E\left[\left(U - \underbrace{E[U]}_{0_{n\times 1}}\right)\left(U - \underbrace{E[U]}_{0_{n\times 1}}\right)^t\right] = E[UU^t] = \sigma_u^2 I$ which is a diagonal matrix with diagonal entries equal to the error variance

3 Estimates and Estimators

Summary 3. In general, any estimator (formula) coming from any method of estimation should satisfy certain properties to ensure its reliability. These differ from large and small samples. For *small samples* ($n \leq 30$), it should have:

1) Unbiasedness: for $\hat{\beta}$, $E[\hat{\beta}] = \beta$

2) Minimum Variance/Efficiency: $Var(\hat{\beta})$ is small

For *large samples* ($n \to \infty$), it should have:

1) Consistency: $\lim_{n\to\infty} \hat{\beta} = \beta$

2) Asymptotic Normality: $n \to \infty \implies \hat{\beta} \sim N$

Summary 4. We investigate a few properties of $\hat{\beta}_{OLS} = (X^tX)^{-1}X^tY$. First, we find a few key facts:

a) First note that

$$\hat{\beta} = (X^tX)^{-1}X^t(X\beta + U) = \beta + (X^tX)^{-1}X^tU$$

and so

$$E[\hat{\beta}] = \beta + (X^tX)^{-1}X^tE[U] = \beta$$

by assumption 2 that says that X is non-stochastic and assumption 5a). So $\hat{\beta}$ is unbiased

b) Next, let's take a look at the variance

$$Var\left[\hat{\beta}_{k\times 1}\right] = \begin{bmatrix} Var(\beta_0) & Cov(\beta_0,\beta_1) \\ Cov(\beta_0,\beta_1) & Var(\beta_1) \end{bmatrix}$$

Writing this in an alternate form,

$$Var[\hat{\beta}] = E\left[\left(\hat{\beta} - E[\hat{\beta}]\right)\left(\hat{\beta} - E[\hat{\beta}]\right)^t\right]_{k\times k} = E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)^t\right]_{k\times k}$$

and recall from part a), the equations

$$\text{(1) } \hat{\beta} = (X^tX)^{-1}X^t(X\beta + U),\ \text{(2) } \hat{\beta} = \beta + (X^tX)^{-1}X^tU,\ \text{(3) } E[\hat{\beta}] = \beta$$

and from (2) we get

$$Var[\hat{\beta}] = E\left[(X^tX)^{-1}X^tUU^tX(X^tX)^{-1}\right] = (X^tX)^{-1}X^tX(X^tX)^{-1}E[UU^t] = (X^tX)^{-1}Var[U]$$

So using the form of $Var[U] = \sigma_U^2 I$, we get that

$$\text{(4) } Var[\hat{\beta}] = \sigma_u^2(X^tX)^{-1}$$

Thus, we need an estimator for σ_u^2. The first guess would be $RSS = \hat{U}^t\hat{U}$; that is $\hat{U}$ is a proxy for U and RSS could be a proxy for σ_u^2. However, note that this is slightly biased. To see this, first note that

$$\text{(5) } E[\hat{U}^t\hat{U}] = E\left[(MU)^t(MU)\right] = E\left[(U^tM^tMU)\right], M = I_n - X(X^tX)^{-1}X^t$$

and

$$\text{(6) } Rank(M) = tr(M) = n - tr((X^tX)^{-1}X^tX) = n - tr(I_k) = n - k$$

Continuing from (5), since M is idempotent, note that

$$\text{(7) } RSS = U^tM^tMU = U^tMU$$

and that for a general $n \times 1$ vector e, we have

$$\text{(8) } e^te = tr(ee^t)$$

So finally, using all equations,

$$E[RSS] = E[(MU)^t(MU)] = E[tr(MU(MU)^t)] = E[tr(UU^tM^tM)] = E[UU^ttr(M)]$$

and thus

$$(9)\ E[RSS] = (n-k)E[UU^t] = \sigma_u^2(n-k)$$

To create an unbiased estimate then, we use the estimate

$$(10)\ \hat{\sigma}^2 = \frac{RSS}{n-k}$$

We then have

$$(11)\ Var[\hat{\beta}] = \hat{\sigma}^2(X^tX)^{-1}$$

c) (Gauss-Markov Theorem) We now show that our estimate $\hat{\beta}_{OLS}$ is efficient and is the best linear unbiased estimator (BLUE).

See the course notes for the proof.

The formal statement of the theorem is that: *In the class of linear and unbiased estimators, it can be shown that $\hat{\beta}_{OLS}$ has the minimum variance. That is*

$$Var\left[\hat{\beta}_{OLS}\right] \leq Var\left[\hat{\beta}_M\right]$$

for any other method M that is linear and unbiased. Thus, $\hat{\beta}_{OLS}$ is the BLUE.

4 Analysis of Variance (ANOVA)

Remark 4.1. The mean of the residuals is 0. That is, $E[\hat{u_t}] = 0$. This can be seen from the normal equation

$$X^tX\hat{\beta} - X^tY = 0$$

or the fact that $X \perp \hat{U}$. This is because $X_{1l} = 1$ for all $l = 1...k$ and so $\sum \hat{u}_t = 0 \implies \bar{\hat{u}} = 0$.

Definition 4.1. Recall that

$$\underbrace{Y_t}_{Total} = \hat{Y}_t + \hat{u}_t = \underbrace{X\hat{\beta}}_{Explained} + \underbrace{\hat{u}_t}_{Residual}$$

We construct the ANOVA table as follows, where everything is expressed in deviation form. So summing and dividing by n on the above equation, we get

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1\bar{X}_1 + ... + \hat{\beta}_n\bar{X}_n + 0 = \bar{X}\hat{\beta}$$

and subtracting the above equation from the first, while squaring the result, we get

$$\sum y_t^2 = \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1,t} + ... + \hat{\beta}_n x_{n,t} + u_t\right)^2 = \sum\left(x\hat{\beta} + \hat{u}_t\right)^2$$

For simple regression (p. 21-24), we get

$$\begin{aligned}
\underbrace{\sum y_t^2}_{TSS} &= \sum\left(x_t\hat{\beta}_1 + \hat{u}_t\right)^2 \\
&= \beta_1^2\sum x_t^2 + 2\beta_1\underbrace{\sum x_t\hat{u}_t}_{X\perp\hat{U}} + \sum\hat{u}_t^2 \\
&= \underbrace{\beta_1^2\sum x_t^2}_{ESS} + \underbrace{\sum\hat{u}_t^2}_{RSS}
\end{aligned}$$

We use the notation that ESS is the explained sum of squares, RSS is the residual sum of squares, and TSS is the total sum of squares, all in deviation form. So $ESS = \beta_1^2\sum x_t^2$, $RSS = \sum u_t^2$ and $TSS = \sum y_t^2$. The actual ANOVA table looks like

Source	SS (Sum of Squares)	Df (Degrees of Freedom)	MSS (Mean SS)
Explained	$ESS = \beta_1^2 \sum x_t^2$	$k-1$	$ESS/(k-1)$
Residual	$RSS = \sum \hat{u}_t^2$	$n-k$	$RSS/(n-k)$
Total	$TSS = \sum y_t^2$	$n-1$	$TSS/(n-1)$

Note that for a simple regression, $k = 2$. The corresponding F statistics is

$$F_{Statistic} = \frac{ESS/(k-1)}{RSS/(n-k)}$$

and the coefficient of determination or R^2 (a crude estimate for the correlation coefficient) is

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{TSS - RSS}{TSS} = \frac{ESS}{TSS}$$

The interpretation for R^2 is that it is a measure of the goodness of fit. It shows how much, in percent, the variation of the dependent variable is being explained by the X's of the model.

4.1 Adjusted R^2 Statistic

Remark 4.2. The drawback of the coefficient of determination is that is only improves by adding more x's (explanatory variables). This might not always be feasible, so use the adjusted R^2, defined by

$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)}$$

and this is a better measure since it includes the number of observations; that is, it can be improved by increasing the number of observations. It can be shown (p. 24) that

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

4.2 Generalized ANOVA

The following can be found in p. 76 in the course notes.

Definition 4.2. Starting with the general RSS, we recall that

$$\begin{aligned} RSS &= \hat{U}^t\hat{U} \\ &= (Y - X\hat{B})^t(Y - X\hat{\beta}) \\ &= Y^tY - 2\hat{\beta}^tX^tY + \hat{\beta}^tX^tX \underbrace{\hat{\beta}}_{(X^tX)^{-1}X^tY} \\ &= Y^tY - 2\hat{\beta}^tX^tY + \hat{\beta}^tX^tY \\ &= Y^tY - \hat{\beta}^tX^tY \\ &= \sum Y_t^2 - \hat{\beta}^tX^tY \end{aligned}$$

and so $\sum \hat{u}_t^2 + \hat{\beta}^tX^tY = RSS + \hat{\beta}^tX^tY = \sum Y_t^2$. Subtracting $n\bar{Y}^2$ from both sides, we get

$$\underbrace{\sum \hat{u}_t^2}_{RSS} + \underbrace{\hat{\beta}^tX^tY - n\bar{Y}^2}_{ESS} = \underbrace{\sum y_t^2}_{TSS}$$

Thus, the general ANOVA table is

Source	SS (Sum of Squares)	Df (Degrees of Freedom)	MSS (Mean SS)
Explained	$ESS = \hat{\beta}^t X^t Y - n\bar{Y}^2$	$k-1$	$ESS/(k-1)$
Residual	$RSS = \sum \hat{u}_t^2$	$n-k$	$RSS/(n-k)$
Total	$TSS = \sum y_t^2$	$n-1$	$TSS/(n-1)$

The F statistic and the coefficient of determination are defined in the same way as in the previous section (in terms of TSS, ESS, RSS).

Summary 5. Let's recap all relevant information up to this point.

1. True model: $Y = X\beta + U$
2. Estimated Residual: $\hat{U} = Y - \hat{Y}$
3. ANOVA
   (a) The division
   $$\begin{aligned} Y &= X\hat{B} + \hat{U} \\ \underbrace{Y^tY - n\bar{Y}^2}_{TSS} &= \underbrace{\hat{\beta}^t X^t Y - n\bar{Y}^2}_{ESS} + \underbrace{\sum \hat{u}_t^2}_{RSS} \end{aligned}$$
   (b) $R^2 = 1 - \frac{RSS}{TSS} = \frac{TSS-RSS}{TSS} = \frac{ESS}{TSS}$
   (c) $F_{Statistic} = \frac{ESS/(k-1)}{RSS/(n-k)}$
   (d) $\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k}$
4. Regression
   (a) $\hat{\beta} = (X^tX)^{-1}X^tY$
   (b) $E[\hat{\beta}] = \beta$; is unbiased
   (c) $Var[\hat{\beta}] = \sigma_u^2(X^tX)^{-1}$; is unbiased
   (d) $\hat{\beta}$ is the BLUE [Guass-Markov]
   (e) $\hat{\sigma}_u^2 = \frac{RSS}{n-k}$

5 Statistical Inference and the GLRM

We start with some basic results from statistical theory.

(1) Recall that if $X \sim N(\mu, \sigma)$, then

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

(2) $Z = \frac{X-\mu}{\sigma}$ is distributed as $Z \sim N(0,1)$.

(3) The sum of squares of n independent standard normal variates is distributed as $\chi^2(n)$ with n degrees of freedom. That is,

$$Z_i \sim N(0,1), Z_i \perp Z_j, i \neq j \implies \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

(4) *(W. Gosset)* The ratio of a standard normal variable Z over the square root of a chi-square distributed r.v., V, over its degrees of freedom, r, is distributed as a $t-$distribution with r degrees of freedom, provided that $Z \perp V$. That is,

$$\frac{Z}{\sqrt{\frac{V}{r}}} \sim t(r), Z \perp V$$

(5) The ratio of two chi-square random variables over their corresponding degrees of freedom gives a Fisher $F-$distribution, provided that the r.v.s are statistically independent. That is, if $U \sim \chi(r_1)$ and $V \sim \chi(r_2)$ then

$$\tfrac{U}{r_1}/\tfrac{V}{r_2} \sim F(r_1, r_2), U \perp V$$

(6) Any linear combination of a set of normal random variables is also normal with different mean and different variance.

Example 5.1. Let $Y_t = \beta_0 + \beta_1 X_t + u_t$, $\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t$ and $\hat{u}_t = Y_t - \hat{Y}_t$. If u_t is normally distributed with mean 0 and variance σ_u^2, then $u_t \sim N(0, \sigma^2)$ for all t. Then, $Y_t = \beta_0 + \beta_1 X_t + u_t$ is $Y_t \sim N(\beta_0 + \beta_1 X_t, \sigma_u^2)$. So if u_t is normal, then Y_t is normal with the same variance.

5.1 Single Variable Inference

(7) Hypothesis Testing + Confidence Intervals / Inference:

Hypothesis Testing

1. Formulate the hypothesis: H_0, H_1
   (a) (Example 1) Suppose we are given a model and estimate:
   $$\begin{aligned} Cons_t &= \beta_0 + \beta_1 Income_t + u_t \\ \widehat{Cons_t} &= 10 + (0.8) Income_t \end{aligned}$$
   We claim that $\beta_1 = 0.9$. The null hypothesis is the claim ($H_0 : \beta_1 = 0.9$) and the alternate hypothesis is the objective, goal, or defense of the study ($H_1 : \beta_1 \neq 0.9$).
   (b) (Example 2) Testing if income is a significant variable in your model gives $H_0 : \beta_k = 0, H_1 : \beta_1 \neq 0$ (called the Test of Significance of One Parameter).
   (c) (Example 3) Using the model in Ex. 1, we claim that we expect the sign of β_1 is positive. We have $H_0 : \beta_1 \leq 0$, $H_1 : \beta > 0$ (called the Test of Expected Sign of a Coefficient)
2. Create the test statistic
   (a) (Example 1) From above, we need a distribution of the estimator of $\hat{\beta}_1$. To do this, we need some assumptions regarding the disturbance term u_t. Thus, we assume it to be standard normal for all t (Assumption 7) which is needed to perform inference. It can be shown that
   $$t = \frac{\hat{\beta}_1 - \beta_1}{sd(\hat{\beta}_1)} \sim t(n-k)$$
   (b) *Aside.* We do not need the normality assumption above to ensure that $\hat{\beta}_{OLS}$ is B.L.U.E.
3. Decision (critical value vs. statistic OR p-value)
   (a) (Example 1) From above, we must make a decision by comparing
   $$t_{Statistic} > t_{Critical} \equiv [t - table]$$

5.2 Inference in the GLRM

Suppose that

$$Y = X\beta + U, U_{n\times 1} \sim N(0_{n\times 1}, \sigma^2 I_n)$$

and using result (6), we have that

$$Y_{n\times 1} \sim N(X\beta, \sigma^2 I_n)$$

and since $\hat{\beta} = (X^tX)^{-1}X^tY$ which is a linear combination of a normal r.v., we have

$$\hat{\beta} \sim N(\beta, \sigma^2(X^tX)^{-1})$$

where we estimate σ^2 with $\hat{\sigma}^2 = \frac{RSS}{n-k}$. The following is the general framework for testing, called the $R\beta$ test.

5.3 $R\beta$ Framework

Example 5.2. Let $k = 5$, $Y_t = \beta_1 + \beta_2 X_{2t} + ... + \beta_5 X_{5t} + u_t$ and we are testing the hypothesis that $\beta_1 + \beta_2 + \beta_3 = 0$. That is

$$H_0 : \beta_1 + \beta_2 + \beta_3 = 0, H_1 : \beta_1 + \beta_2 + \beta_3 \neq 0$$

Suppose we have q restrictions, where

$$r_{q\times 1} = R_{q\times k}\beta_{k\times 1}$$

Then since $q = 1$ in this case,

$$r = 0_{1\times 1} = R\beta = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \end{bmatrix}\beta$$

and we can reform the hypothesis as $H_0 : R\beta = r$ and $H_1 : R\beta \neq r$.

Example 5.3. If the restriction is $\beta_1 + \beta_2 + \beta_4 = 1$ and $\beta_3 = 0$, then

$$r = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = R\beta = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}\beta$$

Definition 5.1. Let $\hat{\beta} = (X^tX)^{-1}X^tY$ be the unrestricted $\hat{\beta}_{OLS}$. Then let $\hat{\beta}_R$ be the restricted $\hat{\beta}_{OLS}$ under $H_0 : R\beta = r$. The derivation of the formula for $\hat{\beta}_R$ is as follows. We want to

$$\begin{aligned} \min_{\{\hat{\beta}_R\}} RSS_R &= \hat{U}_R^t\hat{U}_R \\ \text{subject to} \quad & R\beta = r \end{aligned}$$

and using Lagrange multipliers, define

$$\mathcal{L} = \underbrace{(Y - X\hat{\beta}_R)^t(Y - X\hat{\beta}_R)}_{1\times 1} + \underbrace{\lambda_{q\times 1}^t[r_{q\times 1} - R\hat{\beta}_R]}_{1\times 1}$$

and using the first order condition,

$$(1)\ \frac{\partial \mathcal{L}}{\partial\left(\hat{\beta}_R\right)_{k\times 1}} = 0 \implies -2X^tY + 2X^tX\hat{\beta}_R - R_{k\times q}^t\lambda_{q\times 1} = 0$$

and the second order condition

$$(2)\ \frac{\partial \mathcal{L}}{\partial \lambda_{q\times 1}} = 0 \implies r - R\hat{\beta}_R = 0$$

From (1), λ can not be defined because R is probably singular. To give a definition for λ, we multiply (1) by $R(X^tX)^{-1}$ to get

$$-2R(X^tX)^{-1}X^tY + 2R(X^tX)^{-1}X^tX\hat{\beta}_R - R(X^tX)^{-1}R^t\lambda = 0 \implies -2R\hat{\beta} + 2R\hat{\beta}_R - R(X^tX)^{-1}R^t\lambda = 0$$

and under H_0, we can rewrite this as

$$\begin{aligned} (3) - 2R\hat{\beta} + 2r - R(X^tX)^{-1}R^t\lambda = 0 \quad &\implies \quad \lambda R(X^tX)^{-1}R^t\lambda = -2R\hat{\beta} + 2r \\ &\implies \quad \lambda = 2\left[R(X^tX)^{-1}R^t\lambda\right]^{-1}(r - R\hat{\beta}) \end{aligned}$$

and plugging (3) back into (1), we can solve for $\hat{\beta}_R$ to get

$$(4)\ \hat{\beta}_R = \hat{\beta} + (X^tX)^{-1}R^t\left(R(X^tX)^{-1}R^t\right)^{-1}(r - R\hat{\beta})$$

where our restricted $\hat{\beta}$ is a function of the unrestricted $\hat{\beta}$. Note here that p. 86 and 87 in the textbook is just extra and not

testing material. It can also be shown that

$$\begin{aligned} E[\hat{\beta}_R] &= \hat{\beta} \\ Var[\hat{\beta}_R] &= \hat{\sigma}_u^2(I-AR)(X^tX)^{-1}(I-AR)^t \\ A &= (X^tX)^{-1}R^t\left(R(X^tX)^{-1}R^t\right)^{-1} \end{aligned}$$

which is used in the construction of confidence intervals.

Summary 6. In short, given

$$H_0 : R\beta = r \text{ and } H_1 : R\beta \neq r$$

then

$$\begin{aligned} \hat{\beta} &= (X^tX)^{-1}X^tY \\ \hat{\beta}_R &= \hat{\beta} + (X^tX)^{-1}R^t\left(R(X^tX)^{-1}R^t\right)^{-1}(r - R\hat{\beta}) \end{aligned}$$

Summary 7. There are 3 important tests that we use the above framework ($R\beta$ framework) for.

(1) Testing the significance of ONE coefficient [t-test] (two-sided)

(2) Testing the expected sign of ONE coefficient [t-test] (right/left sided)

(3) Testing the significance of the WHOLE relation [F-test] (based on the ANOVA table)

5.4 Single Restriction $R\beta$ Test

Let's test the validity of one single single restriction. The applications include (1) testing the significance of one parameter (2) testing the expected sign of the coefficient. Let $H_0 : R\beta = r, H_1 : R\beta \neq r$.

Given $U \sim N(0, \sigma_u^2 I_n)$ then

$$\frac{U-0}{\sigma_u} \sim N(0, I_n), \left(\frac{U}{\sigma_u}\right)^t\left(\frac{U}{\sigma_u}\right) = \frac{U^tU}{\sigma_u^2} \sim X(n)$$

and it also follows that $Y = X\beta + U$ is normal such that $Y \sim N(X\beta, \sigma^2 I_n)$ and since $\hat{\beta} = (X^tX)^{-1}X^tY$ then $\hat{\beta} \sim N(\beta, \sigma^2(X^tX)^{-1})$. We also use

$$\hat{\sigma}_u^2 = \frac{RSS}{n-k} = \frac{\hat{U}^t\hat{U}}{n-k} \implies \hat{U}^t\hat{U} \implies (n-k)\hat{\sigma}_u^2$$

and from the above,

$$R\hat{\beta} \sim N(R\beta, \sigma^2 R(X^tX)^{-1}R^t)$$

Define

$$Z = \frac{R\hat{\beta} - R\beta}{\sqrt{\sigma^2 R(X^tX)^{-1}R^t}} = \frac{r - R\beta}{\sqrt{\sigma^2 R(X^tX)^{-1}R^t}} \sim N(0,1)$$

by H_0. Since σ^2 is not known , we use the estimate for σ^2 above instead.

Claim 5.1. The statistic using $\hat{\sigma}_u^2$ instead of σ_u^2 to get $Z \sim t(n-k)$.

Proof. Let $M = I - P = I - X(X^tX)X^t$. From the above, note that

$$\frac{U^tMU}{\sigma_u^2} = \frac{U^tM^tMU}{\sigma_u^2} = \left(\frac{U^tM^t}{\sigma_u^2}\right)\cdot\left(\frac{MU}{\sigma_u^2}\right)_{n\times k} \sim \chi^2(n-k)$$

(which is a sum square of normal r.v.s) and since $MU = \hat{U}$ then

$$\left(\frac{\hat{U}^t}{\sigma_u^2}\right)\cdot\left(\frac{\hat{U}}{\sigma_u^2}\right) = \frac{\hat{U}^t\hat{U}}{\sigma_u^2} = \frac{(n-k)\hat{\sigma}_u^2}{\sigma_u^2} \sim \chi^2(n-k)$$

and so

$$(13)\ \frac{r - R\beta}{\sqrt{\sigma^2 R(X^tX)^{-1}R^t}} \cdot \frac{1}{\sqrt{\frac{(n-k)\hat{\sigma}_u^2}{(n-k)\sigma_u^2}}} = \frac{r - R\beta}{\sqrt{\hat{\sigma}^2 R(X^tX)^{-1}R^t}} \sim t(n-k)$$

if and only if Z and V are independent (proof in textbook). □

Applications

(1) Test the significance of one parameter

Equation (13) simplifies to $t = \frac{\hat{B}_j - Null}{sd(\hat{\beta}_j)}$.

Example 5.4. For a simple linear regression, we know that

$$Var(\hat{\beta}_1) = \frac{\hat{\sigma}_u^2}{\sum x_t^2}, \sum x_t^2 = \sum X_t^2 - n\bar{x}$$

and we want to test the significance of β_1. We are given $sd(\hat{\beta}_1) = 0.1$, $\hat{\beta}_1 = 0.8$. Here, let's say that Y_t is consumption and X_t is income. Testing the significance gives us

1) $H_0 : \beta_1 = 0 \implies r = R\beta, H_1 : \beta_2 \neq 0 \implies r \neq R\beta$.

2) $t = \frac{\hat{\beta}_1 - 0}{\sqrt{Var(\hat{\beta}_1)}} \sim t(n-k) \equiv t \sim \frac{r - R\hat{\beta}}{\sqrt{\hat{\sigma}_u^2 R(X^tX)^{-1}R^t}} \sim t(n-k)$. To see this, note that

$$R\hat{\beta} = \hat{\beta}_1, r = 0$$

and

$$(X^tX)^{-1} = \frac{1}{n\sum X_t^2 - (\sum X_t)^2}\begin{pmatrix} \sum X_t^2 & -\sum X_t \\ -\sum X_t & n \end{pmatrix} \implies R^t(X^tX)^{-1}R^t = \frac{n}{n\sum X_t^2 - (\sum X_t)^2} = \frac{1}{\sum X_t^2 - n\bar{x}}$$

which in deviation form, reduces to

$$R(X^tX)^{-1}R^t = \frac{1}{\sum x_t^2} \implies \hat{\sigma}_u^2 R(X^tX)^{-1}R^t = \frac{\hat{\sigma}_u^2}{\sum x_t^2}$$

and so

$$\sqrt{Var(\hat{\beta}_1)} = \sqrt{\hat{\sigma}_u^2 R^t(X^tX)^{-1}R^t} = \sqrt{\frac{\hat{\sigma}_u^2}{\sum x_t^2}} = sd(\hat{\beta}_1)$$

Thus our statistic is

$$t_{Statistic} = \frac{\hat{\beta}_1 - 0}{\sqrt{Var(\hat{\beta}_1)}} = \frac{0.8 - 1}{0.1} = 8$$

and we reject the hypothesis based on the $t_{Crtical} \equiv t_{Tabulated}$ defined by

$$2P(t_{Estimator} \geq t_{Critical}) = \alpha$$

for a level of significance α in a two tailed test. It is rejected when $t_{Statistic} > t_{Critical}$. Therefore this parameter $\hat{\beta}_1$, income, is significant with a 95% confidence level.

To construct the confidence interval, we want the interval determined by

$$\begin{aligned} Pr\left(\hat{\beta}_1 - sd(\hat{\beta}_1)t_{n-k}^{\frac{\alpha}{2}} < \beta_1 < \hat{\beta}_1 + sd(\hat{\beta}_1)t_{n-k}^{\frac{\alpha}{2}}\right) &= (1-\alpha) \\ Pr\left(0.064 < \beta_1 < 0.996\right) &= 0.95 \end{aligned}$$

where we interpret this as: we are 95% confident that the interval $[0.604, 0.996]$ contains β_1.

Proposition 5.1. *The following are equivalent (interchangeable amongst numbers)*

1) $H_0 : R\beta = r, H_1 : R\beta \neq r$

2) (H_0, H_1) *[Test of significance]* $\beta_j = 0, \beta_j \neq 0$ *{Two-sided}; [Expected sign]* $\beta_j \geq (\leq)0, \beta_j < (>)0$ *where this is the left (right) side {One-sided}; [Test of any claim]* $\beta_j =$ *any value*, $\beta_j \neq$ *same value {Two-sided}*

Remark that it can also be shown that, in the case of a single restriction, the quantity

$$t^2_{statistic} = \left(\frac{\hat{\beta}_j - Null}{sd(\hat{\beta}_j)} \right)^2 = F = \frac{(R\hat{\beta} - r)^t [R(X^tX)^{-1}R^t](R\hat{\beta} - r)/q}{(\hat{U}^t\hat{U})/(n-k)}$$

for the case of a single restriction.

Remark 5.1. In the two sided test, the significance level α is divided on both sides of the given distribution while in the one sided test, the significance level is entirely placed on one tail of the distribution.

5.5 Multiple Restriction $R\beta$ Test

It can be shown that

$$F = \frac{(R\hat{\beta} - r)^t [R(X^tX)^{-1}R^t](R\hat{\beta} - r)/q}{(\hat{U}^t\hat{U})/(n-k)} = \frac{(RSS_R - RSS_{UN})/q}{RSS/(n-k)}$$

Test of the Goodness of Fit

This test is based off of the ANOVA table and we sometimes refer to it as the *overall* significance of the *whole* relation. The statistic in question is

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} \sim F(k-1, n-k)$$

Example 5.5. (The Theory of the Allocation of Time, Gary Becker)

Suppose that we record $s_t =$ your test score, $T_t =$ study time measured in hours, $E_t =$ your consumption of energy drinks. Define $Y_t = \ln S_t$, $X_{2t} = \ln E_t$ and $X_{3t} = \ln T_t$. Our model will be

$$S_t = \psi E_t^{\beta_2} T_t^{\beta_3} e^{\mu_t} \implies \ln S_t = \ln \psi + \beta_2 \ln E_t + \beta_3 \ln T_t + \mu_t$$

which can be rewritten as $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \mu_t$ where $\beta_1 = \ln \psi$. Suppose that $n = 10$. If the data is in deviation form,

$$y_t = \beta_2 x_{2t} + \beta_3 x_{3t} + \mu_t$$

and we are given

$$x^t x = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}, x^t y = \begin{pmatrix} -1 \\ 8 \end{pmatrix}, \sum y_t^2 = 48.2$$

We want to:

1) Estimate $\beta_2, \beta_3, SE[\hat{\beta}_2], SE[\hat{\beta}_3]$ and explain the meaning of the coefficients

Use the least squares unrestricted strategy $\hat{\beta} = (x^tx)^{-1}x^ty$ with

$$\left(x^tx\right)^{-1} = \frac{1}{6-1}\begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} = \frac{1}{5}\begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 0.6 & 0.2 \\ 0.2 & 0.4 \end{pmatrix}$$

to get $\hat{\beta} = \begin{pmatrix} 0.6 & 02 \\ 0.2 & 0.4 \end{pmatrix}\begin{pmatrix} -1 \\ 8 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \implies \hat{\beta}_2 = 1, \hat{\beta}_3 = 3$ with $Y_t = \hat{\beta}_0 + 1X_{2t} + 3X_{3t}$. The coefficients represent the coefficient elasticity of test scores. That is, a 1% increase in energy drinks increases test scores by 1% (unit elastic). Similarly, a 1% increase in time studied increases test scores by 3%. To get the variance, covariance, calculate

$$Var[\hat{\beta}] = \hat{\sigma}_u^2 (x^tx)^{-1}$$

where $\sigma_u^2 = \frac{RSS}{n-k}$ and $RSS = TSS - ESS, TSS = \sum y_t^2 = 48.2, ESS = \beta^t x^t y = 23$. So

$$Var[\hat{\beta}] = \frac{25.2}{10-3}\begin{pmatrix} 0.6 & 02 \\ 0.2 & 0.4 \end{pmatrix} = 3.6\begin{pmatrix} 0.6 & 02 \\ 0.2 & 0.4 \end{pmatrix} = \begin{pmatrix} 2.16 & 0.72 \\ 0.72 & 1.44 \end{pmatrix}$$

and $Var[\hat{\beta}_2] = 2.16$, $Var[\hat{\beta}_3] = 1.44$, $Cov[\hat{\beta}_2, \hat{\beta}_3] = 0.72$.

2) Test the hypothesis that $\beta_2 = \beta_3$ using a *t-test*

Start with $H_0 : \beta_2 - \beta_3 = 0$ and $H_1 : \beta_2 - \beta_3 \neq 0$. The t-statistic is

$$t_{statistic} = \frac{(\hat{\beta}_2 - \hat{\beta}_3) - 0}{\sqrt{Var(\hat{\beta}_2 - \beta_3)}} = \frac{(\hat{\beta}_2 - \hat{\beta}_3) - 0}{\sqrt{Var(\hat{\beta}_2) + Var(\hat{\beta}_3) - 2Cov(\hat{\beta}_2, \hat{\beta}_3)}} = \frac{1 - 3 - 0}{\sqrt{2.16}} = -1.36$$

The $t_{critical} = 2.365$ is based on $\alpha = 5\%$, and $df = 10 - 3 = 7$. Hence, since $|t_{statistic}| < |t_{critical}| \implies$ We don't reject H_0. So we have a problem in our results.

3) Re-estimate the coefficients imposing the restriction $\hat{\beta}_R \equiv (\beta_1 = \beta_2)$.

The new model with the new restrictions is of the form

$$y_t = \beta_3(x_{2t} + x_{3t}) + u_t$$

We now estimate $\hat{\beta}_{2R}$ and $\hat{\beta}_{3R}$ unsigned

$$\hat{\beta}_R = \hat{\beta} + (x^t x)^{-1} \left[R(x^t x)^{-1} R^t\right]^{-1} (r - R\hat{\beta})$$

where

$$r = R\beta \implies [0] = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix}$$

$$\left[R(x^t x)^{-1} R^t\right] = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 0.6 & 0.2 \\ 0.2 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 0.6 \implies \left[R(x^t x)^{-1} R^t\right]^{-1} = \frac{5}{3}$$

$$r - R\hat{\beta} = 0 - \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = 2 \implies \hat{\beta}_R = \begin{bmatrix} 1 \\ 3 \end{bmatrix} + \begin{bmatrix} 0.6 & 0.2 \\ 0.2 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \cdot \frac{5}{3} \cdot 2 = \begin{bmatrix} \frac{7}{3} \\ \frac{7}{3} \end{bmatrix}$$

and so $\hat{\beta}_{2R} = \hat{\beta}_{3R} = \frac{7}{3}$

4) Construct the 95% C.I. for the restricted β_2

First note that

$$Var[\hat{\beta}_R] = \hat{\sigma}_u^2 (I - AR)(x^t x)^{-1} (I - AR)^t, A = (x^t x)^{-1} R^t \left[R(x^t x)^{-1} R^t\right]^{-1}$$

where using our calculation above gives us

$$A = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} \end{pmatrix}^t \implies I - AR = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

and bashing out some numbers and matrices gives us

$$Var[\hat{\beta}_R] = 3.6 \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix} \begin{pmatrix} 0.6 & 0.2 \\ 0.2 & 0.4 \end{pmatrix} \begin{pmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{2}{3} \end{pmatrix} = \begin{pmatrix} 0.56 & 1.12 \\ 1.12 & 2.24 \end{pmatrix}$$

so $Var[\hat{\beta}_{2R}] = 0.56 \implies SE[\hat{\beta}_{2R}] = \sqrt{0.56}$. Now since $t_{n-k=7}^{\alpha/2=2.5\%} = 2.365$ then our confidence interval becomes

$$\frac{7}{3} \pm 2.365 \cdot \sqrt{0.56} \equiv [0.56, 4.13]$$

and we say that we are 95% confident that this interval contains β_{2R}.

5) Test the same hypothesis using a general $R\beta$ test with an F distribution and conclude t^2 (from 2) $= F$ (from 5)

Use the hypotheses: $H_0 : r = R\beta$, $H_1 : r \neq R\beta$. We know that

$$F = \frac{(R\hat{\beta} - r)^t [R(X^t X)^{-1} R^t](R\hat{\beta} - r)/q}{(\hat{U}^t \hat{U})/(n - k)} = \frac{(R\hat{\beta} - r)^t [R(X^t X)^{-1} R^t](R\hat{\beta} - r)/q}{\hat{\sigma}_u^2} \sim F(q, n - k)$$

and calculating $R\hat{\beta} - r^t$ gives us $R\hat{\beta} - r = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} - 0 = -2$. Hence the statistic is

$$F = \frac{(-2)(\frac{5}{3})(-2)/1}{3.6} = 1.85 \implies \sqrt{F} = \sqrt{1.85} = t$$

So this is equivalent to the single restriction case. now the F critical value is

$$F_{Critical}(\alpha = 5\%, df_1 = q = 1, df_2 = n - k = 7) = 5.59$$

and so we do not reject H_0 which is the same conclusion as in 2).

6) Test the significance of the whole relation

Here, we use the hypotheses: $H_0 : \beta_2 = \beta_3 = 0, H_1 : \beta_2 \neq 0$ OR $\beta_3 \neq 0$ OR $\beta_2 \neq \beta_3 \neq 0$. The statistic in this case is

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} \sim F(k-1, n-k)$$

where $k = 3$, $n - k = 7$, $TSS = \sum y_t^2 = 48.2$, $ESS = \hat{\beta}^t x^t y = 23$ and $RSS = TSS - ESS = 25.2$. Thus, $F = \frac{23/2}{25.2/7} = 3.19$ and our $F_{Critical}$ is

$$F_{Critical}(\alpha = 5\%, df_1 = k - 1 = 2, df_2 = n - k = 7) = 4.74$$

Since $F < F_{Critical}$, we do not reject H_0 and our relation is insignificant. You will need to correct your specification or adjust your sample.

7) Test the significance of Energy drinks in the model

This is a simple 2-sided t-test so it will be left as an exercise to the reader.

8) Test you belief that time elasticity of test score is elastic

We want to test that $\beta_3 > 1$ so we use the hypotheses: $H_0 : \beta_3 \leq 1$ and $H_1 : \beta_3 > 1$. Our t-statistic is

$$t_{Statistic} = \frac{\hat{\beta}_3 - 1}{SE[\hat{\beta}_3]} = \frac{3 - 1}{\sqrt{1.44}} = 1.67$$

and our $t_{Critical}$ is

$$t_{Critical}(\alpha = 5\% \text{ (right tailed)}, n - k = 7) = 1.895$$

Since our $t_{Statistic} < t_{Critical}$ then we do not reject H_0.

Summary 8. Here are the relevant pages and content for the midterm

- **CH. 2**, P. 57 (Algebra, properties, geometry of LS, projection, residual matrices, derviations)
- P. 67, S. 2.4.3 (Read), P. 68 to beginning of 69
- P. 71 (LS Estimators; IMPORTANT)
- P. 75 (Everything but Gauss-Markov derivation; will need to known the Theorem, though)
- P. 76 (ANOVA Tables), Eq. 2.6.9., P. 77
- P. 78-79 is NOT required EXCEPT for the formulas for bottom of P. 79
- **CH. 3**, 83-86 (Proofs for P. 85 and P. 86 are not required), Eq. 3.20, Eq. 3.16 ONLY
- P. 87-89 (Hypothesis testing) where 88-89 is just reading
- P. 90-94 (Setting up $R\beta$, validity testing)
- P. 94-100 are VERY IMPORTANT
- P. 101, S. 3.5.

- P. 101-104 READ
- P. 104-106
- **Other**
- Use the $\bar{R}^2$, F test and significance tests (t-test) to check validity. Refer back to the theory from field that is being studied to check validity. These 4 methods help check and validate the model.

The midterm itself will be one question with 14 requirements, equally weighted. There is some computation and some theory. Some proofs will be included. t-Tables and F-tables will be provided.

Testing for Multiple Restrictions

In this section (p. 87), we work under the $R\beta$ framework under multiple restrictions:

$$H_0 : R\beta = r, H_1 : R\beta \neq r, q \neq 1$$

We claim that

$$F = \frac{(R\hat{\beta} - r)^t [R(X^tX)^{-1}R^t](R\hat{\beta} - r)/q}{(\hat{U}^t\hat{U})/(n-k)} \sim F(q, n-k)$$

Proof. Under H_0 and the normality of U,

$$(\hat{\beta} - \beta) \sim N(0, \sigma_u^2(X^tX)^{-1}) \implies (R\hat{\beta} - R\beta) \sim N(0, \sigma_u^2 R(X^tX)^{-1}R^t)$$

and with H_0 we get $(R\hat{\beta} - r) \sim N(0, \sigma_u^2 R(X^tX)^{-1}R^t)$. We then have

$$\frac{(R\hat{\beta} - r)}{\sqrt{\sigma_u^2[R(X^tX)^{-1}R^t]}} \sim N(0,1)$$

where this quantity is q standard normal variates. The sum of squares of this is

$$\text{(1) } (R\beta - r)^t \left(\sigma_u^2[R(X^tX)^{-1}R^t]\right)^{-1} (R\beta - r) \sim \chi^2(q)$$

We also know that

$$\text{(2) } \frac{\hat{\sigma}_u^2(n-k)}{\sigma_u^2} = \frac{U^tMU}{\sigma_u^2} \sim \chi(n-k)$$

and so

$$F = \frac{Eq(1)}{Eq(2)} = \frac{(R\beta - r)^t[R(X^tX)^{-1}R^t]^{-1}(R\beta - r)/q}{\hat{U}^t\hat{U}/(n-k)} \sim F(q, n-k)$$

provided that the r.v.s in Eq. (1) and Eq. (2) are independent. Note that this is also equivalent to

$$F = \frac{(RSS_R - RSS_{UN})/q}{RSS_{UN}/(n-k)}$$

□

6 Departure from Classical Assumptions

There are a few steps of running an experiment. Here, we list the basic outline:

1. Specification
   (a) Collect Data
      i. Sampling via a sample or experiment

ii. Historical time series data

2. Estimation
   (a) Report
3. Quick Valuation

6.1 Mis-specification and Errors of Measurements

Mis-specification of the model $Y = X\beta + U$ could be coming from:

1. Possible problems with the $X's$
   (a) Underfitting (omitting relevant variables)
   (b) Overfitting (adding irrelevant variables)
   (c) Incorrect functional form
2. Measurement errors (data problems)
3. Errors in U[Heteroskedasticity and Serial Correlation] ; Ch. 6
4. Problems related to β's ; Structural breaks ; Parameter constancy problem
   (a) This is when there are shocks in your observed data. That is, the regression model on two mutually exclusive time frames will have two different values for β.
      i. For example if we have a model estimated on data from 1900-1973 in the form of

$$\hat{Y}_t = 10 + 0.8P(oil)_t$$

      and another estimated model with data from 1974 to 2000 in the form of

$$\hat{Y}_t = 15 + 2.8P(oil)_t$$

      then we have a *structural break*.
      ii. The test for structure breaks is called the *Chow test*.

6.2 Problems with X

1. Suppose that we have an incorrect functional form (p. 112).
   (a) Consequences?
      i. It *could* be *unbiased* and *inefficient*
      ii. The t and F tests are invalid
   (b) Detection?
      i. The *informal test* would be to just plot the data.
      ii. The *formal test* is the *Ramsey RESET test*.
2. Suppose that we are underfitting.
   (a) Let the true model be

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \mu_t$$

   but you omitted X_{3t} in the specification of your model. So you mistakenly specified

$$Y_t = \phi_1 + \phi_2 X_{2t} + v_t, v_t = \beta_3 X_{3t} + \mu_t$$

   and you get $E[v_t] = \beta_3 X_{3t} \neq 0$ and $Var[v_t] = \beta_3^2 Var(X_t) + \sigma_u^2 \neq c$ for a constant c.

(b) Consequences?

i. On the least square estimators, the OLS estimators are *biased* iff the excluded variable X_{3t} is correlated with the included variable X_{2t} ($r_{23} \neq 0$)

A. Ex. $\text{Cons}_t = \gamma_0 + \gamma_1 \text{Income}_t + \gamma_2 \text{Interest}_t + \gamma_3 \text{Stock\_Investment}_t + \text{error}_t$ but we exclude the stock investment. and $r_{Inv,Interest} \neq 0 \implies$ Biased OLS estimate.

B. *Proof.* To see this, if we have

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \mu_t \text{ [Excluded]}$$

$$Y_t = \phi_1 + \phi_2 X_{2t} + v_t, v_t = \beta_3 X_{3t} + \mu_t \text{ [Included]}$$

We regress the excluded on the included as

$$\text{(3) } X_{3t} = \alpha_1 + \alpha_2 X_{2t} + \xi_t$$

and estimating $\hat{\alpha}_2$ gives us

$$\text{(4) } \hat{\alpha}_2 = \frac{\sum x_{3t} x_{2t}}{\sum x_{2t}^2}$$

Considering the excluded model in deviation form

$$\text{(5) } y_t = \beta_2 x_{2t} + \beta_3 x_{3t} + \mu_t$$

and multiplying by $\sum x_{2t} / \sum x_{2t}^2$ gives

$$\frac{\sum x_{2t} y_t}{\sum x_{2t}^2} = \beta_2 + \frac{\sum x_{3t} x_{2t}}{\sum x_{2t}^2} \beta_3 + \frac{\sum x_{3t} \mu_t}{\sum x_{2t}^2} \implies \hat{\phi}_2 = \beta_2 + \hat{\alpha}_2 \beta_3 + \frac{\sum x_{3t} \mu_t}{\sum x_{2t}^2}$$

and taking expectations gives us

$$E[\hat{\phi}_2] = \beta_2 + \hat{\alpha}_2 \beta_3 \neq \beta_2$$

Therefore, we conclude that $\hat{\phi}_2$ is a biased estimator of β_2 due to underfitting and the bias is $\beta_3 \cdot \hat{\alpha}_2$. Where β_3 is the coefficient of the excluded and $\hat{\alpha}_2$ is the coefficient of the regression of X_{3t} on X_{2t}. That is, to compute the bias, compute $(\hat{\beta}_2)_{OLS}$ from the excluded model and $(\hat{\alpha}_2)_{OLS}$ from (4).

C. Ex. If we have models

$$Y_t = 10 + 0.2 X_{2t} + 0.5 X_{3t} + \mu_t$$

$$X_{3t} = 2 + 0.25 X_{2t}$$

Then the bias is $\hat{\beta}_3 \cdot \hat{\alpha}_2 = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$ and the bias represents $(\frac{1}{8}/0.5)\%$ of the true coefficient.

ii. The t and F ratios are no longer valid.

(c) Detection?

i. An informal test is to add X_{3t} to your model and check if there is a change in R^2. If it goes up it is relevant.

ii. Another informal test is to add X_{3t} to the model and check the changes in the new estimated coefficients. If there is a significant change, then we have a relevant variable.

iii. The formal test is the Ramsey Reset test.

3. Suppose that we are overfitting.

(a) Let the true model be

$$Y_t = \beta_1 + \beta_2 X_{2t} + u_t$$

but the mis-specified model be

$$Y_t = \theta_1 + \theta_2 X_{2t} + \theta_3 X_{3t} + v_t$$

where X_{3t} is an irrelevant variable.

(b) Consequences?

i. The least squares estimator of the mis-specified model are *unbiased* and *consistent* but no longer *efficient*.

ii. The t and F ratios are no longer valid.

(c) Detection?

i. The informal tests are the same as above in the case of undefitting. However, $\bar{R}^2$ and the estimated coefficients are note expected to change very much.

ii. The more formal test is to test the restriction that $\theta_3 = 0$ using either the t test, the F test or the $t^2 = F$ statistic.

6.3 Ramsey RESET Test

This is used to test for an incorrect functional form or for underfitting.

1. Run OLS and obtain $\hat{Y}_t$ and $\hat{Y}_t$ will incorporate the true functional form or the underfitting (if any exists)
2. Take the unrestricted model
$$Y_t = \phi_0 + \phi_1 X_t + \phi_2 \hat{Y}_t^2 + \phi_3 \hat{Y}_t^3 + ... + \phi_k \hat{Y}_t^k$$
and use the hypotheses
$$H_0 : \forall k, \phi_k = 0, H_1 : \exists k, \phi_k \neq 0$$
Usually $k = 3$.
3. Compute
$$F = \frac{(RSS_R - RSS_{UN})/q}{RSS_{UN}/(n-k)} \sim F_{q,n-k}$$
and reject or don't reject H_0. If we don't reject then we have an incorrect functional form.

6.4 Errors in Y

Let $\tilde{Y}_t = \beta_1 + \beta_2 X_{2t} + u_t$ where $\tilde{Y}_t$ is the true variable and u_t is the population error as white noise. Let's say there are no data problems with X_{2t}. However, Y_t is observed such that $Y_t = \tilde{Y}_t + \xi_t$ where ξ_t is *measurement error* where we assume by simplicity that
$$E[\xi_t] = 0, Var[\xi_t] = \sigma_\xi^2$$
We then have the equation
$$Y_t = \beta_1 + \beta_2 X_{2t} + \left(\underbrace{u_t + \xi_t}_{\epsilon_t} \right)$$
where we call ϵ_t *composite error*.

Proposition 6.1. *The least squares estimators in Y_t from above will remain unbiased but no longer efficient.*

Proof. We first outline some conventional assumptions.

1) $E[\xi_t] = 0 \; \forall t$

2) $Cov(\xi_t, X_{2t}) = 0$

3) $Cov(\xi_t, u_t) = 0$

and these imply that $E[\epsilon_t] = 0$ and $Var(\epsilon_t) = \sigma_\xi^2 + \sigma_u^2$. Using our compact notation, $Y = X\beta + \epsilon$. Using our formula for $\hat{\beta}$ gives us
$$\hat{\beta} = (X^t X)^{-1} X^t (X\beta + \epsilon) = \beta + (X^t X)^{-1} \epsilon$$
$$E[\hat{\beta}] = \beta + (X^t X)^{-1} X^t E[\epsilon] = \beta$$
$$Var(\hat{\beta}) = \sigma_\epsilon^2 (X^t X)^{-1} = \sigma_\xi^2 (X^t X)^{-1} + \sigma_u^2 (X^t X)^{-1} > \sigma_u^2 (X^t X)^{-1}$$
so $\hat{\beta}$ is unbiased but no longer efficient. This makes the t and F ratios invalid (there are larger or smaller than the true values). □

6.5 Errors in X

Let $Y = \tilde{X}\beta + U$ where $U \sim N(0, \sigma^2 I_n)$, Y is true and observed and $\tilde{X}$ is the true matrix of explanatory variables. Suppose that $\tilde{X}_i = X_i - v_i$. In a compact notation, $X = \tilde{X} + V$ where

$$V = \begin{pmatrix} 0 & | & \cdots & | \\ \vdots & v_{2t} & \cdots & v_{kt} \\ 0 & | & \cdots & | \end{pmatrix}$$

So we then have

$$Y = (X - V)\beta + U = X\beta - V\beta + U = X\beta + \underbrace{(U - V\beta)}_{=\epsilon} = X\beta + \epsilon$$

Proposition 6.2. *The $\hat{\beta}_{OLS}$ from above is going to be biased in small samples and inconsistent in large samples.*

Proof. By direct evaluation,

$$\hat{\beta} = (X^tX)^{-1}X^tY = \beta + (X^tX)^{-1}X^t\epsilon$$

$$E[\hat{\beta}] = \beta + E[(X^tX)^{-1}X^t\epsilon] \neq \beta \text{ (because the } x's \text{are no longer fixed)}$$

We now check the inconsistency of the estimator. Using a weak form of the law of large numbers (LLN), that is as $n \to \infty$ then $\frac{1}{n}Q \xrightarrow{p} E[Q]$ or $P\lim_{n\to\infty} \frac{1}{n}Q = E[Q]$, we use some assumptions to show inconsistency. These are as follows

1) $E[\tilde{X}^tV] = 0$; $\tilde{X}$ are not correlated with V

2) $E[\tilde{X}^tU] = 0$; $\tilde{X}$ are not correlated with U

3) $E[U^tV] = 0$; U are not correlated with V

4) All random variables are i.i.d.

Suppose that $E[\tilde{X}^t\tilde{X}] = \Sigma$, $E[V^tV] = \Omega$ and by the law of large numbers, the expression,

$$\frac{X^tX}{n} = \frac{(\tilde{X} + V)^t(\tilde{X} + V)}{n} = \frac{\tilde{X}^t\tilde{X}}{n} + \frac{\tilde{X}^tV}{n} + \frac{V^t\tilde{X}}{n} + \frac{V^tV}{n}$$

will converge in probability to

$$P\lim_{n\to\infty} \frac{X^tX}{n} = E[\tilde{X}^t\tilde{X}] + E[V^tV] = \Sigma + \Omega$$

Similarly,

$$P\lim_{n\to\infty} \frac{X^t\epsilon}{n} = P\lim_{n\to\infty} \left(\frac{\tilde{X}^tU}{n} - \frac{\tilde{X}^tV\beta}{n} + \frac{V^tU}{n} - \frac{V^tV\beta}{n} \right) = 0 + 0 - 0 - \Omega\beta = -\Omega\beta$$

Thus, we have

$$\hat{\beta} = \beta + \left(\frac{X^tX}{n} \right)^{-1} \frac{X^t\epsilon}{n}$$

and so

$$\hat{\beta} = P\lim_{n\to\infty} \hat{\beta} = \lim_{n\to\infty} \left[\beta + \left(\frac{X^tX}{n} \right)^{-1} \frac{X^t\epsilon}{n} \right] = \beta - (\Sigma + \Omega)^{-1}\Omega\beta$$

□

6.6 Instrumental Variables

Theorem 6.1. *(Central Limit Theorem) Suppose that we have $X_1, ..., X_n$ i.i.d. r.v.s with mean μ and variance σ^2. Then*

$$\lim_{n\to\infty} \bar{X} \sim N(\mu, \sigma^2/n)$$

Example 6.1. How can we apply the above theorem? Suppose that we are given an estimator $\hat{\theta}$ for a true population parameter θ. A statement of asymptotic normality is

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, V)$$

for some variance V

Note 2. Recall the implications of having errors in X. We have

$$\hat{\beta} = \beta + (X^t X)^{-1} X^t \epsilon$$

where ϵ is a function of U the true population error and $-\beta V$ which is measurement error. Note that

$$E[\hat{\beta}] = \beta + E[(X^t X)^{-1} X^t \epsilon] \neq \beta \implies E[X^t \epsilon] \neq 0$$

which is a breakdown of our classical assumptions.

Summary 9. So we do the following:

1. Define $Z_{n\times l}$ a matrix of instruments. Note the number of explanatory variables in $Z_{n\times l}$ are not necessarily equal to those in $X_{n\times k}$.
2. Use ZX instead of X iff it satisfies certain properties:
   (a) $E[Z^t U] = 0 \iff p\lim_{n\to\infty}\left(\frac{Z^t U}{n}\right) = 0$
   (b) $E[Z^t X] \neq 0 \iff p\lim_{n\to\infty}\left(\frac{Z^t X}{n}\right) = \Sigma_{ZX}$; Z and X are correlated
   (c) Premultiply the GLRM by $Z_{n\times l}$ to get

$$Z^t Y = Z^t X\beta + Z^t U \implies \hat{\beta}_{IV} = (X^t Z Z^t X)^{-1} X^t Z Z^t Y = (Z^t X)^{-1} Z^t Y$$

   where $\hat{\beta}_{IV}$ is *consistent* and *asymptotically normal*.
   i. *Proof.* To show that it is consistent, we need to show that

$$p\lim_{n\to\infty} \hat{\beta} = \beta$$

   To do this, remark that

$$\begin{aligned}\hat{\beta}_{IV} &= (Z^t X)^{-1}(Z^t X\beta) + (Z^t X)^{-1} Z^t U \\ &= \beta + (Z^t X)^{-1} Z^t U \\ &= \beta + \left(\frac{Z^t X}{n}\right)^{-1}\left(\frac{Z^t U}{n}\right)\end{aligned}$$

   and taking limits gives us

$$p\lim_{n\to\infty} \hat{\beta}_{IV} = \beta + (\Sigma_{ZX})(0) = \beta$$

   ii. *Proof.* To show asymptotic convergence (normality), we need the CLT as defined above. We want to show that

$$\sqrt{n}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} N(0, V)$$

   Recall that

$$\begin{aligned}\hat{\beta}_{IV} &= (X^t Z Z^t X)^{-1}(X^t Z) Z^t Y \\ &= (X^t Z (Z^t Z)^{-1} Z^t X)^{-1}(X^t Z)(Z^t Z)^{-1} Z^t Y \\ &= \left(X^t \text{Proj}_Z X\right)^{-1} X^t \text{Proj}_Z Y \\ &= \left(X^t \text{Proj}_Z X\right)^{-1} X^t \text{Proj}_Z X\beta + \left(X^t \text{Proj}_Z X\right)^{-1} X^t \text{Proj}_Z U \\ &= \beta + \left(X^t \text{Proj}_Z X\right)^{-1} X^t \text{Proj}_Z U\end{aligned}$$

Next, by definition,

$$\begin{aligned} Var[\hat{\beta}_{IV}] &= E[(\beta_{IV}-\beta)(\beta_{IV}-\beta)^t] \\ &= E\left[\left(X^t \text{Proj}_Z X\right)^{-1} X^t \text{Proj}_Z U U^t \text{Proj}_Z X^t \left(X^t \text{Proj}_Z X\right)^{-1}\right] \\ &= \sigma_u^2 (X^t \text{Proj}_Z X)^{-1} \\ &= \sigma_u^2 (X^t Z (Z^t Z)^{-1} Z^t X)^{-1} \\ &= \frac{\hat{U}^t \hat{U}}{n} \left(\frac{X^t Z (Z^t Z)^{-1} Z^t X}{n}\right)^{-1} \end{aligned}$$

since $\text{Proj}_Z X$ is a non-stochastic term that minimizes the error between X and Z and hence annihilates the error in X (which allows us to advance the E operator). The idempotent and symmetric properties of the projection operator onto Z space also allow us to do the above. We now assume that

$$E[Z^t Z] = \Sigma_{ZZ}$$

and so

$$Var[\hat{\beta}_{IV}] = \frac{\hat{U}^t \hat{U}}{n} \left(\frac{X^t Z}{n} \cdot \frac{(Z^t Z)^{-1}}{n} \cdot \frac{Z^t X}{n}\right)^{-1} \xrightarrow{p} \hat{\sigma}_u \left(\Sigma_{ZX} \cdot \Sigma_{ZZ}^{-1} \cdot \Sigma_{ZX}^t\right)^{-1}$$

which when compared to $Var(\hat{\beta}_{OLS}) = \hat{\sigma}^2 (X^t X)^{-1}$ it is larger (BONUS question; show that the difference is positive definite).

Summary 10. Suppose that our true model is $\tilde{Y} = \tilde{X}\beta + U$ and we observe $Y = \tilde{Y} + \epsilon$, $X = \tilde{X} + V$ with the observed model

$$Y = X\beta + \epsilon$$

The problems here are:

1. The $X's$ are stochastic
2. $E[X^t \epsilon] \neq 0$
3. The errors $\epsilon' s$ are no longer white noise? (They are. See below)
   (a) $Var[\epsilon]$ is non-constant (heteroskedasticity) [NOT TRUE from below]
   (b) $Cor(\epsilon_t, \epsilon_s) \neq 0$ (serial correlation) [NOT TRUE from below]

Let's check out problem 3. First remark that

$$Y = (X - V)\beta + U = X\beta + (U - V\beta) = X\beta + \epsilon$$

where $\epsilon_t = u_t - \beta v_t$. Note that

$$Var[\epsilon_t] = \sigma_u^2 + \beta^2 \sigma_v^2 - 2\beta Cov(u_t, v_t) = \sigma_u^2 + \beta^2 \sigma_v^2$$

so we have homoskedasticity. Next we have

$$Cov(\epsilon_t, \epsilon_s) = Cov(u_t - \beta v_t, u_s - \beta v_s) = 0$$

so there is no serial correlation.

Conclusion 1. [ON THE FINAL] Note that problems 1, 2, 3a, 3b are violated in general in case of measurement errors. However if we impose the conventional assumptions that the measurement errors are i.i.d. with constant variance, then only 1 and 2 will be violated.

Summary 11. We have

- Errors in Y: $\hat{\beta}$ are still unbiased but inefficient
- Errors in X: $\hat{\beta}$ are biased, inconsistent and inefficient

- Errors in both: $\hat{\beta}$ are biased, inconsistent and inefficient

and to remedy this, we need to find a matrix $Z_{n\times l}$, $l \geq k$ such that it satisfies certain properties. These are

1. $E[Z^t\epsilon] = 0$
2. $E[Z^tX] = \Sigma_{ZX}$

We premultiply the observed model by Z^t to get:

- If $l = k$ then $p\lim_{n\to\infty}\hat{\beta} = \beta + \Sigma_{ZX}\cdot 0 = \beta$ (we need invertibility of Σ_{ZX})
- If $l > k$, we do a procedure called the two-stages least squares (2SLS):

1. Regress X on Z and obtain a matrix of fitted values $\hat{X}$ (Project X onto Z). That is
$$\hat{X} = Z(Z^tZ)^{-1}Z^tX$$
2. Regress Y on $\hat{X}$ and obtain $\hat{\beta}_{2SLS}$ where
$$\begin{aligned}\hat{\beta}_{2SLS} &= (\hat{X}^t\hat{X})^{-1}\hat{X}^tY \\ &= \left[X^tZ(Z^tZ)^{-1}Z^tZ(Z^tZ)^{-1}Z^tX\right]^{-1}\left[X^tZ(Z^tZ)Z^tY\right] \\ &= \left[X^tZ(Z^tZ)^{-1}Z^tX\right]^{-1}\left[X^tZ(Z^tZ)Z^tY\right] \\ &= (X^t\text{Proj}_ZX)^{-1}X^t\text{Proj}_ZY\end{aligned}$$

   We can show that $\hat{\beta}_{2SLS} = \hat{\beta}_{IV}$. To do this, multiply by $(Z^tZ)(Z^tZ)^{-1}$ in the equation for $\hat{\beta}_{IV}$ to get
$$\hat{\beta}_{IV} = (X^tZ(Z^tZ)^{-1}Z^tX)^{-1}X^tZ(Z^tZ)^{-1}Z^tY = (X^t\text{Proj}_ZX)^{-1}X^t\text{Proj}_ZY = \hat{\beta}_{2SLS}$$

Aside. Efficiency of the instrumental value estimates depends on the covariance of X and Z. To do this, we must show that as $r_{Z,X}\uparrow$ then $Var(\hat{\beta}_{IV})\downarrow$. Do the case for zero covariance and non-zero covariance. Compare this value to $Var(\hat{\beta}_{OLS})$.

7 Non-Spherical Disturbances

Definition 7.1. When we have serial correlation and heteroskedasticity on the error terms, we call these error terms *non-spherical disturbances.* This is when we have a covariance matrix that is not diagonalized and and has non-zero entries on the off-diagonal elements.

7.1 Heteroskedasticity

This is where the variance of u_t, σ_t^2, is a function of t or some other parameter, and is non-constant.

Example 7.1. To illustrate this, consider the income/savings model. Most people with low income will have very little savings (low variability) while there is large variability for people with larger income (it will depend on the individual) so the variance is non-constant.

Sources of Heteroskedasticity:

1. Nature of Y_t
2. Mis-specification
   (a) Suppose that the true model is $\tilde{Y}_t = \beta_1 + \beta_2X_{2t} + u_t$ but we observe $X_{2t} = \tilde{X}_{2t} + v_t$, $Y_t = \tilde{Y}_t$. The observed model is $Y_t + \beta_1 + \beta_2X_{2t} + \epsilon_t$ and the variance of ϵ_t can be shown to be non-constant (it is $\sigma_u^2 + \beta_2^2Var(v_t) - 2\beta_2Cov(u_t, v_t)$)

3. Transformations
4. Varying coefficients

Mathematical Representation of σ_t^2

Here we will describe some general presentations of heteroskedasticity:

1. $\sigma_t^2 = \sigma^2 Z_t^h$ for some $h \neq 0$
2. $\sigma_t^2 = \alpha_0 + \alpha_1 Z_t$
3. $\sigma_t = \alpha_0 + \alpha_1 Z_t$
4. $\sigma_t^2 = f(Z_1, Z_2, ..., Z_n)$

Testing for Heteroskedasticity

Here we describe the procedure for finding heteroskedasticity and what we do about it.

1. Park Test
   (a) Park specified $\sigma_t^2 = \sigma^2 X_t^\beta e^{v_t}$ for the model $Y_t = \beta_1 + \beta_2 X_t + u_t$.
   (b) From here, we linearize the above equation to get $\ln \sigma_t^2 = \ln \sigma^2 + \beta \ln X_t + v_t$. Since $\hat{u}_t$ is observed, it is a proxy for u_t and
   $$Var(\hat{u}_t) = E[(\hat{u}_t - 0)^2] = E[\hat{u}_t^2]$$
   we use $\ln \hat{u}_t$ as a proxy for $\ln u_t$. Our new equation is then
   $$\ln \hat{u}_t^2 = \ln \sigma^2 + \beta \ln X_t + v_t$$
   where we hope that v_t is white noise.
   (c) Test the hypothesis that $H_0 : \beta = 0$ using a t test and reject or not reject the null hypothesis. If we reject, then we have heteroskedasticity.
2. White Test
   (a) Let $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$ and regress Y on the X's to get a series of $\hat{u}_t$
   (b) Run the auxiliary regression (stated in R formula notation) $\hat{u}_t^2 \sim (X_{2t} + X_{3t})^2 + X_{2t}^2 + X_{3t}^2$
      i. This is $u_t = \phi_0 + \phi_1 X_{2t} + \phi_2 X_{3t} + \phi_3 X_{2t}^2 + \phi_4 X_{3t}^2 + \phi_5 X_{2t} X_{3t}$
   (c) Compute R^2 from the previous regression
   (d) White showed that asymptotically, the quantity $W = nR^2 \sim \chi^2(k-1)$ where k is the number of all the parameters in the auxiliary regression (here $k = 6$) If the test statistic is larger than the critical at $\alpha = 5\%, k-1$ then we have heteroskedasticity.
3. Of course we don't know which of the explanatory variables is causing this, but we have some remedies:
   (a) Test using the White procedure
   (b) Narrow it down to a specific variable (could be in the model) or outside the model (one unknown variable)
      i. If it is coming from one of the X's, we can:
         A. Try to replace it with a proxy
         B. Try to replace it with a combination of variables
         C. Drop it
         D. Do some transformations
      ii. It is due to Z (outside of the model)
         A. You could have underfitting
         B. Raise your specification and try to include that missing relevant variable

4. What if you know the exact form of heteroskedasticity?

   (a) Use General Least Squares

      i. *Example.* Suppose that heteroskedasticity is due to X_{2t} and it is taking the following form:

$$\sigma_t^2 = \sigma^2 X_{2t}^h, h = 2$$

      How can we correct for this problem? We use the method of Weighted Least Squares, also known as Generalized Least Squares (GLS)

      A. To do this, we want to "divide by the $\sqrt{}$ of whatever is causing the heteroskedasticity
      B. So let's transform our model as follows

$$\frac{Y_t}{\sqrt{X_{2t}^2}} = \frac{\beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t}{\sqrt{X_{2t}^2}} \equiv Y_t^* = \beta_1 \mathbf{1}^* + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + u_t^*$$

      We then get

$$Var[u_t^*] = \frac{1}{X_{2t}^2} Var[u_t] = \sigma^2$$

      and this new model is homoskedastic.

7.2 Serial Correlation

1. Problem: $Cov(u_t, u_s) \neq 0$ for $t \neq s$
2. Sources: P. 162-164 (more common in time series)
3. Mathematical Representation:

   (a) Let the true model be $Y_t = \beta_1 + \sum_{i=2}^n \beta_i X_{it} + u_t$ such that $E[u_t] = 0$, $Var(u_t) = \sigma^2$ and $Cov(u_s, u_t) \neq 0$

   (b) We will only consider the AR(1) (autoregressive 1) process given by

$$u_t = \rho u_{t-1} + \xi_t$$

   with $E[\xi_t] = 0$, $Var[\xi_t] = \sigma_\xi^2$, $Cov(\xi_t, \xi_s) = 0$ for $t \neq s$, and $|p| < 1$

   (c) Remark that the conversion of this form into a general linear process through the use of forward recursion gives

$$u_t = \xi_t + \sum_{k=1}^{\infty} \xi_{t-k} \rho^k$$

   This implies that $E[u_t] = 0$, $Var[u_t] = \frac{\sigma_\xi^2}{1-\rho^2}$. We also get that

$$Cov(u_t, u_{t-s}) = \frac{\rho^s \sigma_\xi^2}{1-\rho^2}$$

4. Test: Durbin-Watson (D-W) [applies only to AR(1)]:

(a) The $d-$statistic is

$$\begin{aligned} d &= \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n}\hat{u}_t^2} \\ &= \frac{\sum_{t=2}^{n}\left(\hat{u}_t^2 + \hat{u}_{t-1}^2 - 2\hat{u}_{t-1}\hat{u}_t\right)}{\sum_{t=1}^{n}\hat{u}_t^2} \\ &= \frac{\sum_{t=2}^{n}\hat{u}_t^2 + \sum_{t=2}^{n}\hat{u}_{t-1}^2 - 2\sum_{t=2}^{n}\hat{u}_{t-1}\hat{u}_t}{\sum_{t=1}^{n}\hat{u}_t^2} \\ &\approx \frac{2\sum_{t=2}^{n}\hat{u}_t^2 - 2\sum_{t=2}^{n}\hat{u}_{t-1}\hat{u}_t}{\sum_{t=1}^{n}\hat{u}_t^2} \\ &\approx 2(1-\hat{\rho}), \hat{\rho} = \frac{\sum_{t=2}^{n}\hat{u}_t\hat{u}_{t-1}}{\sum_{t=2}^{n}\hat{u}_{t-1}^2} \end{aligned}$$

due to the fact that $\sum \hat{u}_{t-1}^2 \approx \sum \hat{u}_t^2$.

(b) Remark that if:

i. $\rho = -1 \implies d = 4$
ii. $\rho = 1 \implies d = 0$
iii. $\rho = 0 \implies d = 2$

(c) According to Durbin and Watson, if $d \in (d_L, d_U)$ the test is inconclusive for $d_L, d_U \in (0, 2)$ and similarly for a symmetric reflection across $\rho = 2$ (this other interval is $(4 - d_U, 4 - d_L)$). Otherwise we make conclusions based on the proximity of d. Using this, we have several tests related to this.

i. Test for autocorrelation (p. 169):
   A. $H_0 : \rho = 0$; no autocorrelation, $H_1 : \rho \neq 0$; there exists autocorrelation
   B. Calculate $d \approx 2 - 2\hat{\rho}$ and use the d table to get d_L and d_U; use α and $df_1 = n$, $df_2 = k - 1$
   C. Reject, not reject, say the test is inconclusive

5. Remedies: GLS (Aitken 1936)

(a) Set up: $Y_t = \beta_1 + ... + \beta_k X_{kt} + u_t$, $u_t = \rho u_{t-1} + \xi_t$

(b) Apply D-W and if autocorrelation exists, correct using:

i. Use **GLS** if ρ is known:
   A. Take the second lag of

   (2) $Y_t = \beta_1 + ... + \beta_k X_{kt} + u_t$

   and pre-multiply by ρ to get

   (3) $\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{2,t-1} ... + \rho\beta_k X_{k,t-1} + \rho u_{t-1}$

   subtract from the original equation of Y_t to get

   (4) $Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_{2t} - \rho X_{2,t-1}) + ... + (u_t - \rho u_{t-1})$

   Reparamatrize to get

   (5) $Y_t^* = \beta_1^* + \beta_2 X_{2t}^* + ... + \xi_t$

   since

   (1) $u_t = \rho u_{t-1} + \xi_t$

   where ξ_t is white noise.

ii. **Cochrane-Orcutt Iterative Procedure** if ρ is not known:
   A. Run OLS on (2) and obtain a series of residuals $\hat{u}_t$
   B. Compute $\hat{\rho}_1 = \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_{t-1}^2}$
   C. Use $\hat{\rho}_1$ for autocorrelation by applying GLS to get the estimated version of (5)
   D. Apply D-W to (5)

E. If H_0 is accepted, then stop; if H_0 is rejected, go back to (2) using Y_t^* as the new proxy for Y_t

F. Keep iterating until $\hat{\rho}_s \approx \hat{\rho}_{s-1}$ and H_0 is accepted

iii. Remark that the Iterative Procedure doesn't also converge very well (it converges to a random walk) if $\rho \approx 1$

8 Maximum Likelihood Estimation

Suppose that Y is a random variable and $y_i's$ are the realizations of $Y = \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix}$ for $i = 1, 2, ..., n$. Let $\theta \in \Theta$, where θ is a vector of true unknown population parameters.

For example, in the standard GLRM, $Y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + U_{n\times 1}$,

$$\theta = \begin{bmatrix} \beta_1 & \dots & \beta_k & \sigma^2 \end{bmatrix}^t_{(k+1)\times 1}$$

We then do the following:

1. Assume a distribution for Y
2. Define the pdf of y_i as $f_i(y_i|\theta)$ for each i
3. Find the joint pdf of the n realizations, assuming independence, with $f(Y|\theta) = \prod_{i=1}^n f_i(y_i|\theta)$
4. Define the likelihood function $L(\theta|Y) = f(Y|\theta) = \prod_{i=1}^n f_i(y_i|\theta)$
5. Take the log of L as $l(\theta|Y) = \log L(\theta|Y)$
6. Find θ through $\hat{\theta} = \text{argmax}_{\{\theta\in\Theta\}} l(\theta|Y)$

8.1 MLE and the GLRM

Definition 8.1. We define a few matrices:

1) Score Matrix:

$$S(\theta) = \frac{\partial l}{\partial \theta}_{(k+1)\times 1} = 0_{(k+1)\times 1}$$

2) Hessian Matrix:

$$H(\theta) = \frac{\partial^2 l}{\partial\theta\partial\theta'} = \begin{bmatrix} \frac{\partial^2 l}{\partial\beta\partial\beta'} & \frac{\partial^2 l}{\partial\beta\partial\sigma^2} \\ \frac{\partial^2 l}{\partial\sigma^2\partial\beta} & \frac{\partial^2 l}{\partial(\sigma^2)^2} \end{bmatrix}_{(k+1)\times(k+1)}$$

3) Fisher Information Matrix:

$$I(\theta) = -E[H(\theta)]$$

Working in the GLRM framework (that is $Y = X\beta + U$), we will assume that $u_t \sim N(0, \sigma^2)$ for all t. Now the pdf is

$$f_u = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}u_t^2\right\}$$

and the joint pdf is

$$f_U = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}U^tU\right\}$$

Since $Y = X\beta + U$ then by the change of variable theorem, Y will also be normally distributed with a joint pdf of $f_Y = f_U$ which can be re-expressed as

$$f_Y = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}(Y - X\beta)^t(Y - X\beta)\right\} \implies L(\theta|Y) = f_Y$$

We thus define the log-likelihood function as

$$l(\theta|Y) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^t(Y - X\beta)$$

The first order conditions give us

(1) Result #1 [β]:

$$\frac{\partial l}{\partial \beta_{k\times 1}} = 0_{k\times 1} \implies -\frac{1}{2\sigma^2}\left(2X^tY + 2X^tX\beta\right) = 0 \implies \hat{\beta}_{ML} = (X^tX)^{-1}X^tY = \hat{\beta}_{OLS}$$

(2) Result #2 [σ^2]:

$$\frac{\partial l}{\partial \sigma^2} = 0 \implies -\frac{n}{2\sigma^2} + \frac{\underbrace{(Y - X\hat{\beta}_{ML})^t}_{\hat{U}_{ML}}\underbrace{(Y - X\hat{\beta}_{ML})}_{\hat{U}_{ML}}}{2\sigma^4} = 0 \implies \hat{\sigma}^2_{ML} = \frac{\hat{U}^t\hat{U}}{n}$$

In conclusion,

1. In terms of unbiased-ness,
   (a) $\hat{\beta}_{ML} = \hat{\beta}_{OLS} \implies$ the estimate is unbiased for β
   (b) $\hat{\sigma}_{ML} \neq \hat{\sigma}_{OLS} \implies \hat{\sigma}_{ML}$ is biased and $E[\hat{\sigma}_{ML}] = \left(\frac{n-k}{n}\right)\sigma^2$
2. In terms of efficiency,
   (a) $\hat{\beta}_{ML} = \hat{\beta}_{OLS} \implies Var[\hat{\beta}_{ML}] = Var[\hat{\beta}_{OLS}] = \sigma^2(X^tX)^{-1}$ and so our estimate is efficient
   (b) We need to find $Var[\hat{\beta}_{ML}]$
      i. First remark that
      $$H(\theta) = \begin{bmatrix} -\frac{1}{\sigma^2}X^tX & -\frac{1}{\sigma^4}(X^tY - X^tX\beta) \\ -\frac{1}{\sigma^4}(Y^tX - \beta^tX^tX) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(Y - X\beta)^t(Y - X\beta) \end{bmatrix}$$
      and we can define
      $$I(\theta) = -E[H(\theta)] = \begin{bmatrix} \frac{1}{\sigma^2}X^tX & 0 \\ 0 & -\frac{n}{2\sigma^4} + \frac{E[U^tU]}{\sigma^6} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}X^tX & 0 \\ 0 & -\frac{n}{2\sigma^4} + \frac{n}{\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}X^tX & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$
      since $\left(\frac{U}{\sigma}\right)^t\left(\frac{U}{\sigma}\right) \sim \chi^2(n)$ is a sum of squares of n standard normal random variables and $E[\frac{U^tU}{\sigma^2}] = n$. Note that the inverse of the information matrix is
      $$[I(\theta)]^{-1} = \begin{bmatrix} \sigma^2(X^tX)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$
      ii. Recall that
      $$\frac{(n-k)\hat{\sigma}_{OLS}}{\sigma^2} \sim \chi^2(n-k) \implies Var\left[\frac{(n-k)\hat{\sigma}_{OLS}}{\sigma^2}\right] = 2(n-k) \implies Var[\hat{\sigma}_{OLS}] = \frac{2\sigma^4}{n-k}$$
      But $\hat{\sigma}^2_{ML} = \frac{n-k}{n}\hat{\sigma}^2_{OLS}$ and hence
      $$Var(\hat{\sigma}^2_{ML}) = \frac{n-k}{n}\left(\frac{2\sigma^4}{n}\right) \neq \sigma^2$$
      which means that it is inefficient and biased.
3. In conclusion,
   (a) In small samples, $\hat{\beta}_{ML}$ is unbiased and efficient. $\hat{\sigma}_{ML}$ is biased and inefficient.
   (b) In large samples, it can be shown that both estimators are consistent ad asymptotically normal (not shown in this course); That is $\hat{\theta}_{ML}$ is a CAN (consistent and asymptotically normal) estimator.

(c) We can also show that they achieve the *Cramer-Rao lower bound* (WILL BE ON THE FINAL)

4. Let's describe the the *Cramer-Rao lower bound*

   (a) In the class of *consistent and asymptotically normally distributed* (CAN) estimators, $\hat{\theta}_{ML}$ achieves the *minimum variance*. This minimum variance is known as the Cramer-Rao lower bound. **It is the smallest variance of any CAN estimator** and it is equal to $I[\theta]^{-1}$. We can show that

$$Var[\hat{\theta}_{ML}] - [I[\theta]]^{-1} \geq 0$$

   and is in fact an equality. That is, $Var[\hat{\theta}_{ML}] = [I[\theta]]^{-1}$.

      i. To prove equality, we need to show that (A) $Var(\hat{\beta}_{ML}) = \sigma^2(X^tX)^{-1}$ and (B) $Var(\hat{\sigma}^2_{ML}) = \frac{2\sigma^4}{n}$ asymptotically for both.

         A. To follows from the fact that $\hat{\beta}_{ML} = \hat{\beta}_{OLS}$ and $Var(\hat{\beta}_{OLS})$

         B. Recall that

$$\begin{aligned} \frac{(n-k)\sigma^2_{OLS}}{\sigma^2} \sim \chi^2(n-k) \quad &\implies \quad \frac{(n-k)^2}{\sigma^4} Var(\hat{\sigma}_{OLS}) = 2(n-k) \\ &\implies \quad Var(\hat{\sigma}_{OLS}) = \frac{2\sigma^4}{n-k} \end{aligned}$$

         and as $n \to \infty$, $Var(\hat{\sigma}_{OLS}) \to \frac{2\sigma^4}{n}$ asymptotically. Now

$$\hat{\sigma}^2_{ML} = \frac{n-k}{n}\left(\sigma^2_{OLS}\right) \implies Var(\hat{\sigma}^2_{ML}) = \left(1 - \frac{k}{n}\right)^2 Var(\sigma^2_{OLS}) \to Var(\sigma^2_{OLS})$$

         asymptotically as $n \to \infty$. Hence $Var(\hat{\sigma}_{ML}) \to \frac{2\sigma^4}{n}$ as required

      ii. Remark that the OLS and ML methods are equivalent asymptotically.

   (b) We can also see that this is a generalization of the Gauss-Markov therorem relating to the BLUE classes except now generalized to large samples.

8.2 Asymptotic Test using ML (LR test)

Here LR test refers to the likelihood ratio test. The procedure is as follows:

1. Start with the unrestricted model:

   (a) $\hat{\theta}_{ML} = \begin{bmatrix} \hat{\beta}_{ML} = (X^tX)^{-1}X^tY \\ \hat{\sigma}^2_{ML} = \frac{\hat{U}^t\hat{U}}{n} \end{bmatrix}$ where $\hat{U}^t\hat{U} = y^ty - \hat{\beta}^tx^ty$

   (b) $L(\hat{\theta}_{ML}|Y) = \left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_{ML}}\right)^n \exp\left\{-\frac{1}{2\hat{\sigma}^2_{ML}} \underbrace{(Y - X\hat{\beta}_{ML})^t(Y - X\hat{\beta}_{ML})}_{\hat{U}^t\hat{U} = n\hat{\sigma}^2_{ML}}\right\} = (2\pi\hat{\sigma}^2_{ML})^{-\frac{n}{2}}e^{-\frac{n}{2}}$

2. Then do the same thing with the restricted model:

   (a) $\hat{\theta}_R = \begin{bmatrix} \hat{\beta}_R = \hat{\beta}_{ML} + (X^tX)^{-1}R^t[R(X^tX)^{-1}R^t]^{-1}(r - R\hat{\beta}_{ML}) \\ \hat{\sigma}^2_R = \frac{\hat{U}^t_R\hat{U}_R}{n} \end{bmatrix}$ where $H_0 : r = R\beta$

   (b) $L(\hat{\theta}_R|Y) = (2\pi\hat{\sigma}^2_R)^{-\frac{n}{2}}e^{-\frac{n}{2}}$

3. The Likelihood ratio test uses the fact that

$$LRT_{Statistic} = -2\left[\ln L(\hat{\theta}_R) - \ln(\hat{\theta}_{ML})\right] = -2\ln\left(\frac{L(\hat{\theta}_R)}{L(\hat{\theta}_{ML})}\right) \sim \chi^2(q)$$

   where $H_0 : r = R\beta$, $H_1 : r \neq R\beta$, $LRT_{Critical} = \chi^2(q)$ for $\alpha = 5\%$. If $LRT_{Stat} > LRT_{Crit}$ then reject H_0.

(a) Remark that $LRT_{Statistic}$ can also be re-written as

$$LRT_{Statistic} = -2\ln\left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}_{ML}^2}\right)^{-n/2} = n\ln\left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}_{ML}^2}\right) = -2\ln(\Lambda) \sim \chi^2(q)$$

with $\Lambda = \frac{L(\hat{\theta}_R)}{L(\hat{\theta}_{ML})}$

Example 8.1. Consider $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t}$ where

$$Y = \begin{bmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{bmatrix}, X = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 4 \\ 1 & 5 & 6 \\ 1 & 2 & 4 \\ 1 & 4 & 6 \end{bmatrix}, X^tX = \begin{bmatrix} 5 & 15 & 25 \\ 15 & 55 & 81 \\ 25 & 81 & 129 \end{bmatrix}, X^tY = \begin{bmatrix} 20 \\ 76 \\ 109 \end{bmatrix}$$

We want to test the hypothesis that $\beta_2 + \beta_3 = 0$ using LRT.

Solution. (Unrestricted) Working the deviation form $y_t = \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$,

$$(x^tx)^{-1} = \begin{bmatrix} 1 & -1.5 \\ -1.5 & 2.5 \end{bmatrix}, x^ty = \begin{bmatrix} 16 \\ 9 \end{bmatrix}, \hat{\beta}_{ML} = (x^tx)^{-1}x^ty = \begin{bmatrix} -2.5 \\ -1.5 \end{bmatrix}$$

This then gives

$$\hat{U}^t\hat{U} = y^ty - \beta^tx^ty = 28 - \begin{pmatrix} 2.5 & -1.5 \end{pmatrix}\begin{pmatrix} 16 \\ 9 \end{pmatrix} = 1.5 \implies \hat{\sigma}_{ML}^2 = \frac{1.5}{5} = 0.3$$

(Restricted) We then compute $\hat{\beta}_R$ as

$$\hat{\beta}_R = \hat{\beta}_{ML} + (X^tX)^{-1}R^t[R(X^tX)^{-1}R^t]^{-1}(r - R\hat{\beta}_{ML})$$

where

$$R(X^tX)^{-1}R^t = \begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1.5 \\ -1.5 & 2.5 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{2} \implies (R(X^tX)^{-1}R^t)^{-1} = 2$$

$$r - R\hat{\beta} = 0 - \begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 2.5 \\ -1.5 \end{pmatrix} = -1$$

So we then have

$$\hat{\beta}_R = \begin{bmatrix} 3.5 \\ -3.5 \end{bmatrix}, \hat{U}_R\hat{U}_R = 3.5 \implies \hat{\sigma}_R^2 = 0.7$$

Here, with $H_0 : r = R\beta, H_1 : r \neq R\beta$,

$$LRT = -2\ln\left(\frac{0.7}{0.3}\right)^{-5/2} = 4.24, LRT_{Critical}(\alpha = 5\%, df = 1) = 3.841$$

and so we reject H_0.

9 Basic Sampling Concepts

(Will be added by adapting online notes)

Let N be the number of members of the population, Y be the population variable, for $Y_1, ..., Y_N$, n be the size of the sample drawn from the population, y_i be the realization of the $Y_i's$ (i.e. they are values taken from the population Y_i for $i = 1, ..., n$).

We also define $f = \frac{n}{N}$ as the sample fraction.

In sampling, we care about 3 characteristics of the population:

1. Population Total $t = \sum_{i=1}^{N} Y_i$

2. Population Mean: $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{t}{N}$
3. Population Proportion: p

9.1 Simple Random Sampling (SRS)

In SRS,

1. We use $\bar{y}$ (the sample mean) to estimate $\bar{Y}$. That is, $\bar{y}$ is an estimator for $\bar{Y}$. Here, $\bar{y} = \frac{1}{n}\sum y_i$ and has the properties:
   (a) $E[\bar{y}] = \bar{Y}$
   (b) $Var[\bar{y}] = (1-f)\frac{S^2}{n}$ where S^2 is the true population variance. But S^2 is not known so we use the sample variance $s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$. Therefore, $\widehat{Var[\bar{y}]} = (1-f)\frac{s^2}{n}$.
2. Let's examine how we use the sample to estimate the population total. We know that $t = N\bar{Y}$ and since $\bar{y}$ is an estimator for $\bar{Y}$, we can use $\widehat{t} = N\bar{Y}$ which will be our estimator for t. It has the following properties:
   (a) $E[\hat{t}] = t$
   (b) $Var(\hat{t}) = N^2 Var(\bar{y}) = N^2(1-f)\frac{s^2}{2}$
3. We skip the estimator, $\hat{p}$, for p.

Example 9.1. (Assignment 4, Question 7) We are given $N = 6$, a population set $U_{\text{Index}} = \{1,2,3,4,5,6\}$ with $Y_i = \{3,4,3,4,2,2\}$.

a) We get that the population mean is $\bar{Y}_i = 3$ and the population variance is $s^2 = 0.8$.

b) The possible number of SRS's is $\binom{6}{3} = 20$

c) The probability of 1 SRS drawn is 1 over the number of possible SRS's. That is $\frac{1}{20}$.

d) The probability distribution of the sample mean is found as follows. We generate a list of all possible 3 element combinations from Y_i and the corresponding estimator values. Use this information to create the frequency distribution for the estimator. In this case, the mean has the following distribution:

$$P\left(\bar{y} = \frac{7}{3}\right) = \frac{2}{20}, P\left(\bar{y} = \frac{8}{3}\right) = \frac{4}{20}, P\left(\bar{y} = \frac{9}{3}\right) = \frac{8}{20}, P\left(\bar{y} = \frac{10}{3}\right) = \frac{4}{20}, P\left(\bar{y} = \frac{11}{3}\right) = \frac{2}{20}$$

and so $E[\bar{y}] = 3 = E[\bar{Y}]$ with $Var(\bar{y}) = \sum(y_i - \bar{y})^2 Pr_i = 0.133$.

9.2 Stratified Sampling

Example 9.2. (Assignment 4 Question 8) We are given that

$$U_{\text{index}} = \{1,2,3,4,5,6,7,8\}, Y_i = \{\underbrace{1,2,4,8}_{N_1}, \underbrace{4,7,7,7}_{N_2}\}$$

where N_1 and N_2 are the first and second stratums respectively. We want to take SRS's from from stratums:

a) $\underline{SRS_1 \text{ of size } n_1 = 2:}$

The number of possible SRS_1 is $\binom{4}{2} = 6$. We then have:

Sample No.	y_i	$P(s_i)$	$\bar{y}$	$\hat{t} = N_1\bar{y}$
1	$\{1,2\}$	$1/6$	1.5	$4 \times 1.5 = 6$
2	$\{1,4\}$	$1/6$	2.5	$4 \times 2.5 = 10$
3	$\{1,8\}$	$1/6$	4.5	$4 \times 4.5 = 18$
4	$\{2,4\}$	$1/6$	3	$4 \times 3 = 12$
5	$\{2,8\}$	$1/6$	5	$4 \times 5 = 20$
6	$\{4,8\}$	$1/6$	6	$4 \times 6 = 24$

b) $\underline{SRS_2 \text{ of size } n_2 = 2}$:

The number of possible SRS_2 is $\binom{4}{2} = 6$.

Sample No.	y_i	$P(s_i)$	$\bar{y}$	$\hat{t} = N_1\bar{y}$
1	$\{4,7\}$	$1/6$	5.5	$4 \times 5.5 = 22$
2	$\{4,7\}$	$1/6$	5.5	$4 \times 5.5 = 22$
3	$\{4,7\}$	$1/6$	5.5	$4 \times 5.5 = 22$
4	$\{7,7\}$	$1/6$	7	$4 \times 7 = 28$
5	$\{7,7\}$	$1/6$	7	$4 \times 7 = 28$
6	$\{7,7\}$	$1/6$	7	$4 \times 7 = 28$

c) The sampling distribution is $\hat{t}_{Str} = \hat{t}_1 + \hat{t}_2$. To do this, we construct the following table:

j	k	$j+k$	$P(\hat{t}_1 = j, \hat{t}_2 = k)$
6	22	28	$\frac{1}{6} \times \frac{1}{2} = 1/12$
6	28	34	$\frac{1}{6} \times \frac{1}{2} = 1/12$
10	22	32	$\vdots$
10	28	38	
12	22	34	
12	28	40	
18	22	$\vdots$	
18	28		
$\vdots$	$\vdots$		

and constructing the distribution table gives us:

$\hat{t}_{Str} = \hat{t}_1 + \hat{t}_2$	Probability
28	$1/12$
32	$1/12$
34	$2/12$
38	$\vdots$
40	
42	
46	
48	

Final Exam Review

- Will be 6 questions; 2 will be proofs; 4 will be problems
- Material will be on the following basic concepts:
  - Setting up the problem in the GLRM
  - Know that $\hat{\beta} = (X^tX)^{-1}X^tY$, $\hat{\sigma}_u = \frac{\hat{U}^t\hat{U}}{n-k}$, and the formula for $\hat{\beta}_R$
  - Know the F statistic for testing the restriction $R\beta = r$
  - Know that $TSS = RSS + ESS \implies Y^tY - n\bar{Y}^2 = \hat{U}^t\hat{U} + \hat{\beta}^tX^tY - n\bar{Y}^2 \implies y^ty = \hat{U}^t\hat{U} + \hat{\beta}^tx^ty$
  - Be able to work in deviation form
  - Explain the coefficients in the log-log, semi-log, and linear models
  - No proofs for pre-midterm material

- In Chapter 4 we have:
  - Eq. 4.14. to Eq. 4.17. are NOT required (p. 115-116)
  - Eq. 4.25. and Eq. 4.26. are NOT required but you should be able to state the result
  - Eq. 4.36. to Eq. 4.40. are NOT required but you should be able to state the consequences
  - P. 120-121, Eq. 4.44 to 4.51 are helpful for Assignment 4 Question 1 and for testing positive definiteness; should be able to state the consequences
  - P. 124-128 are NOT required
- We exclude Chapter 5 (not covered in lectures, unfortunately)
- In Chapter 6 we have:
  - P. 154 $\implies$ Be able to understand and list the sources
  - P.158-160 are NOT required
  - P. 162 $\implies$ Be able to understand and list the sources
  - P. 171 $\implies$ The Breusch-Godfrey test is NOT required
  - P. 170-179 are NOT required
  - P. 180-181 $\implies$ Know the feasible GLS procedure (square-root transformation)
  - P. 182-184 is NOT required
  - P. 185 $\implies$ Know the Cochrane-Orcutt Method
  - P. 186 $\implies$ The Hildreth-Lu procedure is NOT required
- For Chapter 7, you will be tested on what was covered in class:
  - Know the definition of $S(\theta)$, $H(\theta)$, $I(\theta)$
  - P. 204-207 $\implies$ Understand the LRT (likelihood ratio test), excluding the proof, based on Λ and the standard procedure
  - Know the asymptotic properties of MLE which are $\hat{\beta}_{ML}, \hat{\sigma}^2_{ML}$, and the Cramer-Rao lower bound
  - Know $\hat{\beta}_{IV}$, $\hat{\beta}_{2SLS}$
- Understand the sampling examples covered in class

Office hours are on August 1-7 during the usual office hours and all day on August 13, 14 (2013).

Assignment 4 is due by 4pm on Monday, July 29, 2013 in the drop off boxes in M3 or in class.

Index