

Graduate Econometrics Recitations: Fall 2022

Alex Houtz*

March 24, 2023

Contents

1	Basics and Review	5
1.1	Matlab	5
1.2	Math-Stats Review	9
2	Classical Linear Regression	13
2.1	Building the Model	13
2.2	OLS Estimator	15
3	Bias and Consistency	21
3.1	Practice Problem 1: Hansen 2.16	21
3.2	Asymptotic Theorems	24
3.3	Practice Problem 2: Hansen 4.23 and 7.2	25
4	Applying OLS Fundamentals	29
4.1	Practice Problem: Hansen 7.7	29
4.2	Delta Method	31
4.3	Practice Problem: Intro Hansen 8.8	31
4.4	Practice Problem: Hansen 3.13	33
5	More Regression	37
5.1	Practice Problem: Hansen 3.13	37
5.2	Previous Problem: Question 2	40
5.3	Previous Problem: Question 5	42
5.4	Matlab Functions	43
6	Algebra Review	47
6.1	Previous Problem: Homework 3 Question 1	47
6.2	Previous Question: Homework 3 Question 2	50
6.3	Violating the Exogeneity Assumption	53
7	Non-Linear Least Squares	55
7.1	Theory	55
7.2	Example	58
7.3	Homework 4 F-Statistics	60

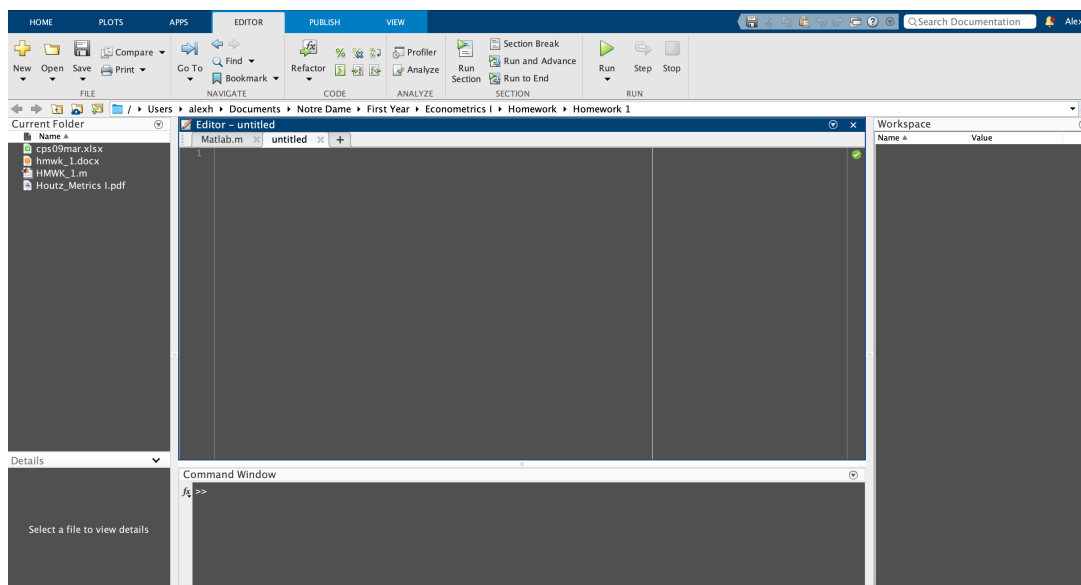
8	Midterm Review	63
8.1	Frisch-Waugh-Lovell Theorem	63
8.2	Wald Test	65
8.3	Lagrange Multiplier Test	66
8.4	Chow Test	67
8.5	Strict vs. Weak Exogeneity	68
9	Maximum Likelihood Estimation	71
9.1	Maximum Likelihood Theory	71
9.2	Practice Problem 1: Hansen 10.4	74
9.3	Practice Problem 2: Hansen 10.7	75
9.4	Practice Problem 3	76
9.5	Practice Problem 4: Hansen 13.3 Extended	79
9.6	Practice Problem 5: Hansen 13.1 Extended	82
9.7	Graphing MLE	85
10	Generalized Least Squares	87
10.1	GLS Theory	87
10.2	Heteroskedasticity	91
10.3	Example	92
11	Generalized Method of Moments	95
11.1	Previous Problem: PS 7, Question 2	95
11.2	Generalized Method of Moments Theory	99
12	Time Series	103
12.1	ARMA(1,1)	103
12.2	AR(2) Process	106
12.3	Deriving the MA(∞) Form for an AR(1)	110
13	Final Review	113
13.1	Asymptotic Properties of Non-Linear Least Squares	113
13.2	Cochrane-Orcutt Procedure	116
13.3	Reverse Regression	118

Chapter 1

Basics and Review

1.1 Matlab

We need to begin with the structure of Matlab.

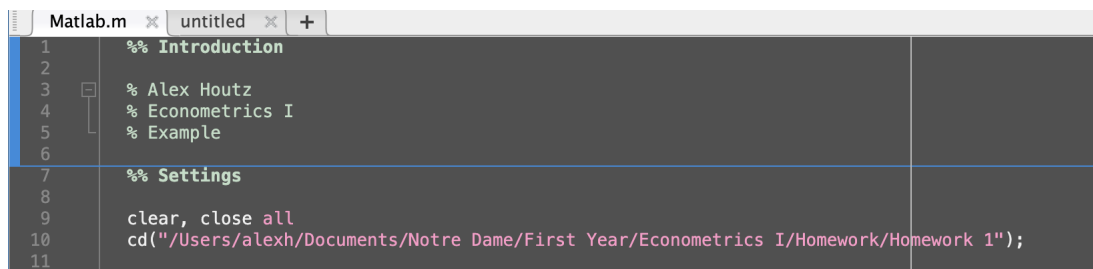


- The “Editor” window (upper middle) is where we create the code file. Your saved work will come from here.
- The “Command” window (bottom middle) contains output from your code. You can also directly input code here that you don’t want saved.
- The “Current Folder” window (upper left) contains the files specified in the path I have declared (see below in Matlab Set-up).

- The “Workspace” window (right) contains saved variables, matrices, arrays, etc.

1.1.1 Matlab Set-up

There will be many Matlab assignments throughout the year. Establishing a simple, clean set-up to your code will make all of our lives easier. Feel free to experiment and create your own, but here is an example:



```

Matlab.m x untitle d x +
1 %% Introduction
2
3 % Alex Houtz
4 % Econometrics I
5 % Example
6
7 %% Settings
8
9 clear, close all
10 cd("/Users/alexh/Documents/Notre Dame/First Year/Econometrics I/Homework/Homework 1");
11

```

There are a couple “tricks” to note:

- The double percent signs, “%%”, create new sections. So in the example, I have the section “Introduction” and the section “Settings”.
- The single percent sign, “%”, tells Matlab that the following line is a comment and not code.
- The “clear, close all” commands clear the workspace, ensuring I’m not overwriting variables in other files, and close all figures I have open.
- The “cd(...)” command specifies the path on which Matlab will look for data and function files.
- The semi-colon, “;” suppresses output in the command window.

1.1.2 Some Useful Commands

```
%% Examples

% Defining Variables

x = 10000;
y = sin(x);

% Import Data (can also use 'readtable')

data = readmatrix("cps09mar.xlsx");
ex = ones(size(data(:,1)));

% For loop

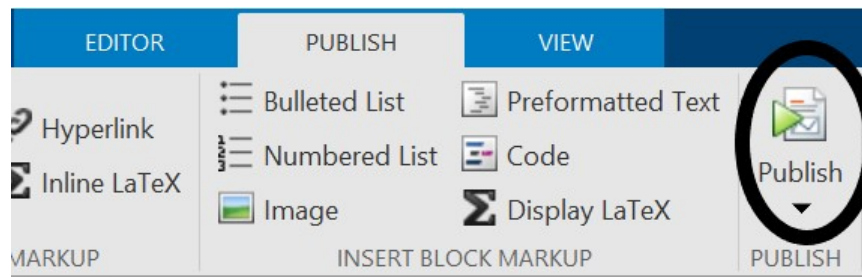
mu = [0 0];
sigma = [1 .9; .9 1];
rng(121);

for g = 1:10000
    R = mvnrnd(mu, sigma, 100);
    u(:,g) = R(:,1);
    v(:,g) = R(:,2);
end
```

- Variables must be defined for values, function outputs, and data if you want to reference them later in your code.
- To import data, declare your path then read the data into Matlab using “readmatrix” or “readtable.” See documentation for more details on these commands.
- For loops will be useful in most Matlab assignments. In the example above, I’m creating two error term vectors with a multivariate normal distribution. Essentially, Matlab is pulling two random numbers from the distribution 10,000 times. NOTE: Try to avoid triple loops if possible. These can take a long time to run.

1.1.3 Publishing

Your code needs to be able to run from the beginning to the end with no problems. One easy way to check this, and also to turn in your code and output, is to publish your code.



- Usually we work in the Editor tab. To publish, click the Publish tab and then click on Publish.
- When the preview pops up, print to PDF. The end result should look similar to the picture below.
- Make sure the necessary output is visible in the published PDF. Remember that by using a semi-colon you suppress the output.

Contents

- [Introduction](#)
- [Settings](#)
- [Examples](#)

Introduction

```
% Alex Houtz
% Econometrics I
% Example
```

Settings

```
clear, close all
cd("/Users/alexh/Documents/Notre Dame/First Year/Econometrics I/Homework/Homework 1");
```

1.1.4 General Advice

- Coding is hard. Do not worry if you are struggling. Find a friend to code with and work through these together (do make sure you understand what the code is doing).
- Try to reduce run times for your code. Most assignments should be able to run in less than a minute.
- The internet is your best friend for Matlab (and code in general). Check out places like Matlab Answers first.
- Make sure to save your work.

1.2 Math-Stats Review

1.2.1 The Analogy Principle

Suppose we have a function $\beta = h(\theta)$ where $\theta = \mathbb{E}[g(y_i)]$. We want to estimate β . We do not know $\mathbb{E}[g(y_i)]$. Therefore, we replace θ with $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n y_i$. So $\hat{\beta} = h(\hat{\theta})$.

Let's go through an example using variance. Variance is given by :

$$\sigma^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

In practice, replace the expectations with the sample mean. Therefore:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2$$

We now have a viable estimate for variance.

1.2.2 Bias

The definition of bias is:

$$\mathbb{E}[\hat{\theta}] - \theta = \text{bias}$$

If $\text{bias} = 0$, we say that the estimator is unbiased. How does this apply to the analogy principle? Ideally, we'd like to create unbiased estimators for the parameters we are looking for. Let's look at the estimator for the variance that we just found.

First, we should check our two plug-ins to see if they are unbiased.

$$\begin{aligned} \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n x_i^2 \right] &= \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n x_i^2 \right] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[x_i^2] \\ &= \frac{1}{n} \cdot n \mathbb{E}[x_i^2] \\ &= \mathbb{E}[x_i^2] \end{aligned}$$

So that part is unbiased. Let's check the standard mean:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n x_i\right] &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n x_i\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[x_i] \\
&= \frac{1}{n}\sum_{i=1}^n \mu \\
&= \frac{1}{n} \cdot n\mu \\
&= \mu
\end{aligned}$$

So this is also unbiased. Let's check them together in the variance estimate:

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n x_i^2\right] - \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n x_i^2\right] - \mathbb{E}\left[\left(\mu - \mu + \frac{1}{n}\sum_{i=1}^n x_i\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n x_i^2\right] - \mathbb{E}\left[\left(\mu + \frac{1}{n}\sum_{i=1}^n (x_i - \mu)\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n x_i^2\right] - \mathbb{E}\left[\mu^2 + 2\mu\frac{1}{n}\sum_{i=1}^n (x_i - \mu) + \left(\frac{1}{n}\sum_{i=1}^n (x_i - \mu)\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n x_i^2\right] - \mathbb{E}\left[\mu^2 + 2\mu(\bar{x} - \mu) + (\bar{x} - \mu)^2\right] \\
&= \mathbb{E}[x_i^2] - \mu^2 - 2\mu(\mu - \mu) - \mathbb{E}[(\bar{x} - \mu)^2] \\
&= \sigma^2 - \text{Var}(\bar{x}) \\
&= \sigma^2 - \text{Var}\left(\frac{1}{n}\sum_{i=1}^n x_i\right) \\
&= \sigma^2 - \frac{1}{n^2}\sum_{i=1}^n \text{Var}(x_i) \\
&= \sigma^2 - \frac{1}{n^2} \cdot n\sigma^2 \\
&= \sigma^2 - \frac{\sigma^2}{n}
\end{aligned}$$

$$= \left(1 - \frac{1}{n}\right) \sigma^2$$

Unfortunately, our estimator is biased, where the bias is $-\frac{1}{n}\sigma^2$. Consider the following estimator:

$$\hat{s}^2 = \frac{1}{n-1} \sum_{i=1}^n [x_i - \bar{x}]^2$$

You can prove that this is unbiased on your own. To correct our estimator, we simply add $\frac{\hat{s}^2}{n}$:

$$\begin{aligned} \hat{\sigma}_{unbiased}^2 &= \hat{\sigma}^2 + \frac{\hat{s}^2}{n} \\ \mathbb{E}[\hat{\sigma}_{unbiased}^2] &= \mathbb{E}\left[\hat{\sigma}^2 + \frac{\hat{s}^2}{n}\right] \\ &= \mathbb{E}[\hat{\sigma}^2] + \mathbb{E}\left[\frac{\hat{s}^2}{n}\right] \\ &= \sigma^2 - \frac{\sigma^2}{n} + \frac{\sigma^2}{n} \\ &= \sigma^2 \end{aligned}$$

1.2.3 Deriving the OLS Estimator

Consider the $k \times 1$ vector of population moment conditions

$$\mathbb{E}[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = 0$$

where $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters, y_i is a scalar and \mathbf{x}_i is a $k \times 1$ vector.

- (a) Solve for the method of moments population parameter $\boldsymbol{\beta}_0$.
- (b) In the solution to (a), the population method of moments parameter $\boldsymbol{\beta}_0$ is only identified if what condition holds?
- (c) Suppose you observe an *i.i.d.* sample $\{y_i, \mathbf{x}_i\}_{i=1}^n$ from some unknown "true" joint distribution $f(y_i, \mathbf{x}_i)$. Using the analogy principle, propose estimators for the following population moments:

$$\begin{aligned} &\mathbb{E}[\mathbf{x}_i \mathbf{x}_i'] \\ &\mathbb{E}[\mathbf{x}_i \mathbf{x}_i']^{-1} \\ &\mathbb{E}[\mathbf{x}_i y_i] \end{aligned}$$

- (d) Consider again the *i.i.d.* sample $\{y_i, \mathbf{x}_i\}_{i=1}^n$ from $f(y_i, \mathbf{x}_i)$. Given your solution to part (a), use the analogy principle to propose a method of moments estimator $\hat{\boldsymbol{\beta}}_n$ for the population parameter $\boldsymbol{\beta}_0$.

Part a

We first note the moment conditions:

$$\mathbb{E}[\mathbf{x}_i(y_i - \mathbf{x}'_i\boldsymbol{\beta})] = 0$$

Now, solve for $\boldsymbol{\beta}_0$:

$$\begin{aligned}\mathbb{E}[\mathbf{x}_i(y_i - \mathbf{x}'_i\boldsymbol{\beta})] &= 0 \\ \mathbb{E}[\mathbf{x}_iy_i] - \mathbb{E}[\mathbf{x}_i\mathbf{x}'_i]\boldsymbol{\beta} &= 0 \\ \mathbb{E}[\mathbf{x}_iy_i] &= \mathbb{E}[\mathbf{x}_i\mathbf{x}'_i]\boldsymbol{\beta} \\ \mathbb{E}[\mathbf{x}_i\mathbf{x}'_i]^{-1}\mathbb{E}[\mathbf{x}_iy_i] &= \boldsymbol{\beta}_0\end{aligned}$$

Part b

For there to be one unique solution for $\boldsymbol{\beta}_0$, we need to be able to solve the equation. $\mathbb{E}[\mathbf{x}_iy_i]$ shouldn't be a problem, but $\mathbb{E}[\mathbf{x}_i\mathbf{x}'_i]^{-1}$ may cause issues. If $\mathbb{E}[\mathbf{x}_i\mathbf{x}'_i]$ is singular, then we will not be able to find $\boldsymbol{\beta}_0$. For a matrix to be invertible, it needs to be full rank. Therefore, we need $\mathbb{E}[\mathbf{x}_i\mathbf{x}'_i]$ to be full rank. This requirement is known as the rank condition.

Part c

Simply apply the analogy principle:

$$\mathbb{E}[\mathbf{x}_i\mathbf{x}'_i] \longrightarrow \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i\mathbf{x}'_i \tag{1.1}$$

$$\mathbb{E}[\mathbf{x}_i\mathbf{x}'_i]^{-1} \longrightarrow \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i\mathbf{x}'_i \right)^{-1} \tag{1.2}$$

$$\mathbb{E}[\mathbf{x}_iy_i] \longrightarrow \frac{1}{n} \sum_{i=1}^n \mathbf{x}_iy_i \tag{1.3}$$

Part d

Simply plug in our estimators from (c):

$$\hat{\boldsymbol{\beta}}_n = \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i\mathbf{x}'_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_iy_i \right)$$

Chapter 2

Classical Linear Regression

2.1 Building the Model

We start with the basic structure of the linear regression model:

$$y = x\beta + \varepsilon \qquad \text{—or—} \qquad y_i = x_i'\beta + \varepsilon_i$$

On the left, y is an $n \times 1$ vector, x is an $n \times k$ matrix, β is a $k \times 1$ vector, and ε is an $n \times 1$ vector. On the right, y_i and ε_i are scalars, x_i is a $k \times 1$ vector (so x_i' is a $1 \times k$ vector) and β is a $k \times 1$ vector.

For our coefficient estimators to be the best, unbiased linear estimators (BLUE), we need five assumptions:

Assumption 1 (Linearity of the Model). *All coefficients in the model must be linear.*

This assumption is straightforward. β must be linear. It cannot be in an exponent, logged, etc. Notice that this assumption does not restrict transformations of our variables. For example:

$$\ln(y) = \ln(x)\beta + \varepsilon$$

This equation is fine. But:

$$y = x\ln(\beta) + \varepsilon$$

is not fine.

Assumption 2 (Rank Condition). *The right hand side variables, x , must have full column rank:*

$$\text{rank}(x) = k$$

If x does not have full column rank, then later on we will not be able to invert $x'x$ to identify β .

Assumption 3 (Exogeneity of x). *The right-hand side variables, x , are exogenously related to the*

error term, ε :

$$\mathbb{E}[\varepsilon_i|x] = 0$$

This assumption of strict exogeneity is strong, but necessary for the coefficient estimators to be unbiased. Later, when we cover asymptotics, we will need a weaker assumption called “orthogonality,” or $\mathbb{E}[x'\varepsilon] = 0$. Strict exogeneity gives us orthogonality automatically:

$$\begin{aligned}\mathbb{E}[x'\varepsilon] &= \mathbb{E}_x [x\mathbb{E}[\varepsilon|x]] && \text{(Using LIE)} \\ &= 0\end{aligned}$$

In addition, strict exogeneity tells us that the errors have mean zero:

$$\begin{aligned}\mathbb{E}[\varepsilon_i] &= \mathbb{E}_x [\mathbb{E}[\varepsilon_i|x]] \\ &= 0\end{aligned}$$

Therefore, the covariance between ε_i and x is zero:

$$\begin{aligned}\text{Cov}(\varepsilon_i, x) &= \mathbb{E}[x\varepsilon_i] - \mathbb{E}[\varepsilon_i]\mathbb{E}[x] \\ &= 0 - \mathbb{E}_x [\mathbb{E}[\varepsilon_i|x]] \mathbb{E}[x] \\ &= 0\end{aligned}$$

So then the conditional expectation of the model is:

$$\begin{aligned}\mathbb{E}[y|x] &= \mathbb{E}[x\beta|x] + \mathbb{E}[\varepsilon|x] \\ &= x\beta\end{aligned}$$

Assumption 4 (Spherical Errors). *The error terms are homoskedastic and are not cross-correlated:*

$$\begin{aligned}\text{Var}(\varepsilon_i|x) &= \sigma^2 && \text{(Homoskedasticity)} \\ \text{Cov}(\varepsilon_i, \varepsilon_j) &= 0 \quad \forall i \neq j && \text{(No Cross-Correlation)}\end{aligned}$$

This assumption implies that $\text{Var}(\varepsilon|x) = \sigma^2 I$. We need to assume this for the coefficient estimators to be considered “best.”

Assumption 5 (Exogenous Data Generating Process). *The right hand side variables, x , are exogenously generated.*

We treat x as a randomly created variable.

Assumption 6 (Normally Distributed Errors). *The error term, ε , is normally distributed after conditioning on the right hand variables, x :*

$$\varepsilon|x \sim N(0, \sigma^2 I)$$

This assumption is made for convenience and is not necessary for the coefficient estimator to be BLUE.

2.2 OLS Estimator

In class, you derived the OLS estimator by minimizing the sum of squared residuals. Last week in recitation, we derived the same estimator by starting with a moment condition. Today, let's use algebra.

Starting with the regression model:

$$\begin{aligned} y &= x\beta + \varepsilon \\ x'y &= x'x\beta + x'\varepsilon \\ \mathbb{E}[x'y] &= \mathbb{E}[x'x\beta] + \mathbb{E}[x'\varepsilon] \\ \mathbb{E}[x'y] &= \mathbb{E}[x'x]\beta + \mathbb{E}[x'\varepsilon] \end{aligned} \quad \text{(Use exogeneity)}$$

$$\mathbb{E}[x'y] = \mathbb{E}[x'x]\beta + 0 \quad \text{(Use rank condition)}$$

$$\mathbb{E}[x'x]^{-1}\mathbb{E}[x'y] = \beta \quad \text{(Use analogy principle)}$$

$$\left(\frac{1}{n} \sum_{i=1}^n x_i x_i'\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i y_i\right) = \hat{\beta}$$

2.2.1 Linear Projection

Define the linear projection as:

$$p(x) = x'\gamma$$

The best linear projection minimizes the mean squared error:

$$\begin{aligned} mse(p(x)) &= \mathbb{E}[\varepsilon^2] \\ &= \mathbb{E}[(y - x'\gamma)^2] \\ &= \mathbb{E}\left[\left((y - \mathbb{E}[y|x]) + (\mathbb{E}[y|x] - x'\gamma)\right)^2\right] \\ &= \mathbb{E}[(y - \mathbb{E}[y|x])^2] + 2\mathbb{E}[(y - \mathbb{E}[y|x])(\mathbb{E}[y|x] - x'\gamma)] + \mathbb{E}[(\mathbb{E}[y|x] - x'\gamma)^2] \end{aligned}$$

Let's look at just the middle term first:

$$\begin{aligned} 2\mathbb{E}[(y - \mathbb{E}[y|x])(\mathbb{E}[y|x] - x'\gamma)] &= 2\mathbb{E}[y\mathbb{E}[y|x] - yx'\gamma - \mathbb{E}[y|x]^2 + \mathbb{E}[y|x]x'\gamma] \\ &= 2(\mathbb{E}[y]\mathbb{E}[y|x] - \mathbb{E}[yx']\gamma - \mathbb{E}[y|x]^2 + \mathbb{E}[yx']\gamma) \\ &= 2(\mathbb{E}_x[\mathbb{E}[y|x]^2] - \mathbb{E}_x[\mathbb{E}[y|x]x']\gamma - \mathbb{E}_x[\mathbb{E}[y|x]^2] + \mathbb{E}_x[\mathbb{E}[y|x]x']\gamma) \\ &= 0 \end{aligned}$$

Back to the mean-squared error:

$$mse(p(x)) = \mathbb{E}[(y - \mathbb{E}[y|x])^2] + \mathbb{E}[(\mathbb{E}[y|x] - x'\gamma)^2]$$

We want to minimize, so take the derivative with respect to γ :

$$\begin{aligned} \frac{\partial mse(p(x))}{\partial \gamma} &= 2\mathbb{E}[x(\mathbb{E}[y|x] - x'\gamma)] = 0 \\ 2\mathbb{E}[x\mathbb{E}[y|x]] - 2\mathbb{E}[xx']\gamma &= 0 \\ \mathbb{E}[xx']^{-1}\mathbb{E}[xy] &= \gamma \end{aligned}$$

Now we know that, if the true process isn't in the space generated by the right-hand side variables, the best linear predictor is still the OLS estimator.

Projection of y

Suppose we want the best linear prediction of the dependent variable y . Then we would use our right-hand side variables and the estimated β :

$$\hat{y} = x\hat{\beta}$$

If we rearrange this equation:

$$\begin{aligned} \hat{y} &= x(x'x)^{-1}x'y \\ &= Py \end{aligned}$$

where $P = x(x'x)^{-1}x'$. We call P the projection matrix.

Projection Matrix

Let x be an $n \times k$ matrix that is full rank. Then the projection matrix P is an $n \times n$ that results from:

$$P = x(x'x)^{-1}x'$$

The projection matrix can be shown to have the following properties:

- (i) $Px = x$
- (ii) $P = P'$
- (iii) $PP = P$
- (iv) $tr(P) = k$ and $rank(P) = k$

We can use the projection matrix to find estimated \hat{y} values in our regressions.

Annihilator Matrix

Let x be an $n \times k$ matrix that is full rank. Then the annihilator matrix M is an $n \times n$ that results from:

$$\begin{aligned} M &= I - P \\ &= I - x(x'x)^{-1}x' \end{aligned}$$

The annihilator matrix can be shown to have the following properties:

- (i) $Mx = 0$
- (ii) $MP = 0$
- (iii) $M = M'$
- (iv) $MM = M$
- (v) $tr(M) = n - k$ and $rank(M) = n - k$

We can use the annihilator matrix to remove parts of the regression that we are not interested in estimating.

2.2.2 Partitioned Regression

Let x be composed of x_1 and x_2 . Then we can write the regression model as:

$$\begin{aligned} y &= x\beta + \varepsilon \\ &= x_1\beta_1 + x_2\beta_2 + \varepsilon \end{aligned}$$

Rearranging the original equation above:

$$\begin{aligned} x'y &= x'x\beta + x'\varepsilon \\ \mathbb{E}[x'y] &= \mathbb{E}[x'x]\beta \end{aligned}$$

Applying the analogy principle and stacking vectors, let's put this in matrix form:

$$\begin{bmatrix} x'_1x_1 & x'_1x_2 \\ x'_2x_1 & x'_2x_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} x'_1y \\ x'_2y \end{bmatrix}$$

Take the first equation (the top line). Let's solve for $\hat{\beta}_1$:

$$\begin{aligned} (x'_1x_1)\hat{\beta}_1 + (x'_1x_2)\hat{\beta}_2 &= x'_1y \\ \hat{\beta}_1 &= (x'_1x_1)^{-1}x'_1y - (x'_1x_1)^{-1}(x'_1x_2)\hat{\beta}_2 \end{aligned} \quad (*)$$

Now take equation two:

$$\begin{aligned}
(x'_2x_1)\hat{\beta}_1 + (x'_2x_2)\hat{\beta}_2 &= x'_2y && \text{(Plug in } \hat{\beta}_1\text{:)} \\
(x'_2x_1) \left((x'_1x_1)^{-1}x'_1y - (x'_1x_1)^{-1}x'_1x_2\hat{\beta}_2 \right) + (x'_2x_2)\hat{\beta}_2 &= x'_2y \\
(x'_2x_1)(x'_1x_1)^{-1}x'_1y - (x'_2x_1)(x'_1x_1)^{-1}x'_1x_2\hat{\beta}_2 + (x'_2x_2)\hat{\beta}_2 &= x'_2y \\
(x'_2x_2)\hat{\beta}_2 - (x'_2x_1)(x'_1x_1)^{-1}x'_1x_2\hat{\beta}_2 &= x'_2y - (x'_2x_1)(x'_1x_1)^{-1}x'_1y \\
x'_2(I - x_1(x'_1x_1)^{-1}x'_1)x_2\hat{\beta}_2 &= x'_2(I - x_1(x'_1x_1)^{-1}x'_1)y \\
x'_2(I - P_1)x_2\hat{\beta}_2 &= x'_2(I - P_1)y \\
x'_2M_1x_2\hat{\beta}_2 &= x'_2M_1y \\
\hat{\beta}_2 &= (x'_2M_1x_2)^{-1}(x'_2M_1y)
\end{aligned}$$

Theorem 1 (Frisch-Waugh-Lovell). *Let the regression model be as follows:*

$$y = x_1\beta_1 + x_2\beta_2 + \varepsilon$$

Then the estimate for β_2 will be the same as the estimate from the following model:

$$M_1y = M_1x_2\beta_2 + M_1u$$

This is the estimator we derived above. What is the variance of the FWL estimator?

$$\begin{aligned}
\text{Var}(\hat{\beta}_2|x) &= \text{Var} \left((x'_2M_1x_2)^{-1}(x'_2M_1y) | x \right) \\
&= \text{Var} \left((x'_2M_1x_2)^{-1}(x'_2M_1x_2)\beta + (x'_2M_1x_2)^{-1}(x'_2M_1\varepsilon) | x \right) \\
&= \text{Var} \left((x'_2M_1x_2)^{-1}(x'_2M_1\varepsilon) | x \right) \\
&= \mathbb{E} \left[\left((x'_2M_1x_2)^{-1}(x'_2M_1\varepsilon) \right) \left((x'_2M_1x_2)^{-1}(x'_2M_1\varepsilon) \right)' \middle| x \right] \\
&= (x'_2M_1x_2)^{-1}x_2M_1\mathbb{E}[\varepsilon\varepsilon'|x]M_1x_2(x'_2M_1x_2)^{-1} \\
&= (x'_2M_1x_2)^{-1}x_2M_1\sigma^2IM_1x_2(x'_2M_1x_2)^{-1} \\
&= \sigma^2(x'_2M_1x_2)^{-1}x_2M_1M_1x_2(x'_2M_1x_2)^{-1} \\
&= \sigma^2(x'_2M_1x_2)^{-1}x_2M_1x_2(x'_2M_1x_2)^{-1} \\
&= \sigma^2(x'_2M_1x_2)^{-1}
\end{aligned}$$

What would happen if we just ran a regression of y on x_1 , ignoring x_2 ? Would this estimator be biased

or unbiased? We start by going back to equation (*):

$$\begin{aligned}\hat{\beta}_1 &= (x_1'x_1)^{-1}x_1'y - (x_1'x_1)^{-1}(x_1'x_2)\hat{\beta}_2 \\ \hat{\beta}_1 &= \tilde{\beta}_1 - (x_1'x_1)^{-1}(x_1'x_2)\hat{\beta}_2 \\ \tilde{\beta}_1 &= \hat{\beta}_1 + (x_1'x_1)^{-1}(x_1'x_2)\hat{\beta}_2 \\ \mathbb{E}[\tilde{\beta}_1|x] &= \mathbb{E}[\hat{\beta}_1|x] + (x_1'x_1)^{-1}(x_1'x_2)\mathbb{E}[\hat{\beta}_2|x] \\ \mathbb{E}[\tilde{\beta}_1|x] &= \beta_1 + (x_1'x_1)^{-1}(x_1'x_2)\beta_2\end{aligned}\tag{*}$$

So $\mathbb{E}[\tilde{\beta}_1] \neq \beta_1$ unless $\beta_2 = 0$ or $(x_1'x_1)^{-1}(x_1'x_2) = 0$.

Chapter 3

Bias and Consistency

3.1 Practice Problem 1: Hansen 2.16

Let X and Y have the joint density $f(x, y) = \frac{3}{2}(x^2 + y^2)$ on $0 \leq x \leq 1$ and $0 \leq y \leq 1$.

- Compute the coefficients of the best linear predictor of $Y = \alpha + \beta X + \varepsilon$.
- Compute the conditional expectation $m(x) = \mathbb{E}[Y|X = x]$. Is the BLP different from the conditional expectation?

3.1.1 Solution

Part a

As in class, the BLP is defined as:

$$P(y|x) = \mathbf{x}'\beta^*$$

So we need to calculate β^* . Return to your notes to find that:

$$\beta^* = \mathbb{E}[\mathbf{x}\mathbf{x}']^{-1}\mathbb{E}[\mathbf{x}y]$$

where $\mathbf{x} = [1 \ x]'$. Let's multiply out the two parts of β^* . First, $\mathbb{E}[\mathbf{x}\mathbf{x}']$:

$$\mathbb{E} \left[\begin{bmatrix} 1 \\ x \end{bmatrix} \begin{bmatrix} 1 & x \end{bmatrix} \right] = \begin{bmatrix} 1 & \mathbb{E}[x] \\ \mathbb{E}[x] & \mathbb{E}[x^2] \end{bmatrix}$$

Then $\mathbb{E}[\mathbf{x}y]$:

$$\mathbb{E} \left[\begin{bmatrix} 1 \\ x \end{bmatrix} \cdot y \right] = \begin{bmatrix} \mathbb{E}[y] \\ \mathbb{E}[xy] \end{bmatrix}$$

To calculate these two parts, let's first break the joint distribution into the two marginal distributions:

$$\begin{aligned}
 f(x) &= \int_{-\infty}^{\infty} f(x, y) dy \\
 &= \int_0^1 \frac{3}{2}(x^2 + y^2) dy \\
 &= \left. \frac{3}{2}x^2y + \frac{3}{2} \frac{1}{3}y^3 \right|_0^1 \\
 &= \frac{3}{2}x^2 + \frac{1}{2}
 \end{aligned}$$

$$\begin{aligned}
 f(y) &= \int_{-\infty}^{\infty} f(x, y) dx \\
 &= \int_0^1 \frac{3}{2}(x^2 + y^2) dx \\
 &= \left. \frac{3}{2} \frac{1}{3}x^3 + \frac{3}{2}y^2x \right|_0^1 \\
 &= \frac{3}{2}y^2 + \frac{1}{2}
 \end{aligned}$$

Now we can calculate the components of the estimator that we need:

$$\begin{aligned}
 \mathbb{E}[x] &= \int_{-\infty}^{\infty} xf(x) dx \\
 &= \int_0^1 \frac{3}{2}x^3 + \frac{1}{2}x dx \\
 &= \left. \frac{3}{2} \frac{1}{4}x^4 + \frac{1}{4}x^2 \right|_0^1 \\
 &= \frac{3}{8} + \frac{1}{4} \\
 &= \frac{5}{8}
 \end{aligned}$$

$$\begin{aligned}
 \mathbb{E}[y] &= \int_{-\infty}^{\infty} yf(y) dy \\
 &= \int_0^1 \frac{3}{2}y^3 + \frac{1}{2}y dy \\
 &= \left. \frac{3}{2} \frac{1}{4}y^4 + \frac{1}{4}y^2 \right|_0^1 \\
 &= \frac{3}{8} + \frac{1}{4} \\
 &= \frac{5}{8}
 \end{aligned}$$

$$\begin{aligned}
\mathbb{E}[x^2] &= \int_{-\infty}^{\infty} x^2 f(x) dx \\
&= \int_0^1 \frac{3}{2} x^4 + \frac{1}{2} x^2 dx \\
&= \left. \frac{3}{2} \frac{1}{5} x^5 + \frac{1}{2} \frac{1}{3} x^3 \right|_0^1 \\
&= \frac{3}{10} + \frac{1}{6} \\
&= \frac{7}{15}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[xy] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x, y) dx dy \\
&= \int_0^1 \int_0^1 \frac{3}{2} xy(x^2 + y^2) dx dy \\
&= \int_0^1 \int_0^1 \frac{3}{2} (x^3 y + xy^3) dx dy \\
&= \int_0^1 \frac{3}{2} \left[\frac{1}{4} x^4 y + \frac{1}{2} x^2 y^3 \right]_0^1 dy \\
&= \frac{3}{2} \int_0^1 \frac{1}{4} y + \frac{1}{2} y^3 dy \\
&= \frac{3}{2} \left[\frac{1}{8} y^2 + \frac{1}{8} y^4 \right]_0^1 \\
&= \frac{3}{2} \times \frac{1}{4} \\
&= \frac{3}{8}
\end{aligned}$$

We now have all the moments that we need. Let's plug them into the β^* equation:

$$\begin{aligned}
\beta^* &= \begin{bmatrix} 1 & \frac{5}{8} \\ \frac{5}{8} & \frac{7}{15} \end{bmatrix}^{-1} \begin{bmatrix} \frac{5}{8} \\ \frac{3}{8} \end{bmatrix} \\
&= \frac{1}{\frac{7}{15} - \left(\frac{5}{8}\right)^2} \begin{bmatrix} \frac{7}{15} & -\frac{5}{8} \\ -\frac{5}{8} & 1 \end{bmatrix} \begin{bmatrix} \frac{5}{8} \\ \frac{3}{8} \end{bmatrix} \\
&= \frac{960}{73} \begin{bmatrix} \frac{11}{192} \\ -\frac{1}{64} \end{bmatrix} \\
&= \begin{bmatrix} \frac{55}{73} \\ -\frac{15}{73} \end{bmatrix}
\end{aligned}$$

Therefore, the BLP is:

$$P(y|x) = \frac{55}{73} - \frac{15}{73}x$$

Part b

To find the conditional expectation, we need to find the conditional distribution:

$$\begin{aligned} f(y|x) &= \frac{f(x, y)}{f(x)} \\ &= \frac{\frac{3}{2}(x^2 + y^2)}{\frac{3}{2}x^2 + \frac{1}{2}} \\ &= \frac{3x^2 + 3y^2}{3x^2 + 1} \end{aligned}$$

Now let's take the conditional expectation:

$$\begin{aligned} \mathbb{E}[y|x] &= \int_{-\infty}^{\infty} yf(y|x)dy \\ &= \int_0^1 y \frac{3x^2 + 3y^2}{3x^2 + 1} dy \\ &= \frac{1}{3x^2 + 1} \int_0^1 3x^2y + 3y^3 dy \\ &= \frac{1}{3x^2 + 1} \left[\frac{3}{2}x^2y^2 + \frac{3}{4}y^4 \right]_0^1 \\ &= \frac{3x^2}{6x^2 + 2} + \frac{3}{12x^2 + 4} \\ &= \frac{6x^2 + 3}{12x^2 + 4} \end{aligned}$$

This is not the same as the BLP.

3.2 Asymptotic Theorems

There are three main consistency theorems you will need to know well.

3.2.1 Weak Law of Large Numbers

Let $\{x_i\}$ be a sequence of *i.i.d.* random variables with $\mathbb{E}[x_i] = \mu$ and $\text{Var}(x_i) = \sigma^2 < \infty$. Then:

$$\frac{1}{n} \sum_{i=1}^n x_i \xrightarrow{P} \mu$$

Theorem 2. *If $\{x_i\}$ is an *i.i.d.* sequence of random variables with $\mathbb{E}[|g(x_i)|] < \infty$, then:*

$$\frac{1}{n} \sum_{i=1}^n g(x_i) \xrightarrow{P} \mathbb{E}[g(x)]$$

To prove this theorem, simply use the weak law of large numbers and the properties of independent

variables.

3.2.2 Continuous Mapping Theorem

Let x_n be a sequence of real-valued random vectors and let $h : \mathcal{R}^k \rightarrow \mathcal{R}^m$. Define the set of discontinuous points as:

$$D_h = \{x \in \mathcal{X} : h(\cdot) \text{ is discontinuous at } x\}$$

Now, if $P(x \in D_h) = 0$ and:

- (i) if $x_n \xrightarrow{P} x$, then $h(x_n) \xrightarrow{P} h(x)$
- (ii) if $x_n \xrightarrow{d} x$, then $h(x_n) \xrightarrow{d} h(x)$
- (iii) if $x_n \xrightarrow{a.s.} x$, then $h(x_n) \xrightarrow{a.s.} h(x)$

3.2.3 Slutsky's Theorem

If $x_n \xrightarrow{d} x$ and $y_n \xrightarrow{P} y$, then:

- (i) $x_n + y_n \xrightarrow{d} x + y$
- (ii) $x_n y_n \xrightarrow{d} xy$
- (iii) $\frac{x_n}{y_n} \xrightarrow{d} \frac{x}{y}$ if $y \neq 0$

Slutsky's theorem is a special case of the continuous mapping theorem. Keep in mind that it still holds if x_n converges in probability instead. In that scenario, parts (i), (ii), and (iii) converge in probability instead of in distribution.

3.3 Practice Problem 2: Hansen 4.23 and 7.2

Define the **ridge regression** estimator as:

$$\hat{\beta} = \left(\sum_{i=1}^n (X_i X_i') + \lambda I_k \right)^{-1} \left(\sum_{i=1}^n X_i Y_i \right)$$

where $\lambda > 0$ is a constant.

- a) Find $\mathbb{E}[\hat{\beta}|X]$. Is $\hat{\beta}$ a biased estimator for β ?
- b) Find the probability limit of $\hat{\beta}$ as $n \rightarrow \infty$. Is $\hat{\beta}$ consistent for β ?

3.3.1 Solution

Part a

Take the conditional expectation of the ridge estimator:

$$\begin{aligned}
 \mathbb{E}[\hat{\beta}|X] &= \mathbb{E} \left[\left(\sum_{i=1}^n (X_i X_i') + \lambda I_k \right)^{-1} \left(\sum_{i=1}^n X_i Y_i \right) \middle| X \right] \\
 &= \mathbb{E} \left[\left(\sum_{i=1}^n (X_i X_i') + \lambda I_k \right)^{-1} \left(\sum_{i=1}^n X_i (X_i' \beta_i + \varepsilon_i) \right) \middle| X \right] \\
 &= \left(\sum_{i=1}^n (X_i X_i') + \lambda I_k \right)^{-1} \left(\sum_{i=1}^n X_i X_i' \right) \beta + \left(\sum_{i=1}^n (X_i X_i') + \lambda I_k \right)^{-1} \mathbb{E} \left[\left(\sum_{i=1}^n X_i \varepsilon_i \right) \middle| X \right] \\
 &= \left(\sum_{i=1}^n (X_i X_i') + \lambda I_k \right)^{-1} \left(\sum_{i=1}^n X_i X_i' \right) \beta + \left(\sum_{i=1}^n (X_i X_i') + \lambda I_k \right)^{-1} \sum_{i=1}^n X_i \mathbb{E} [\varepsilon_i | X] \\
 &= \left(\sum_{i=1}^n (X_i X_i') + \lambda I_k \right)^{-1} \left(\sum_{i=1}^n X_i X_i' \right) \beta
 \end{aligned}$$

This is not β , so the ridge regression estimator is biased.

Part b

Start from the estimator (I have simply multiplied and divided by n):

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^n (X_i X_i') + \frac{\lambda}{n} I_k \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i Y_i \right)$$

Take each piece individually first. We know that by the Weak Law of Large Numbers:

$$\begin{aligned}
 \frac{1}{n} \sum_{i=1}^n x_i x_i' &\xrightarrow{P} \mathbb{E}[x_i x_i'] \\
 \frac{1}{n} \lambda I_k &\longrightarrow 0 \\
 \frac{1}{n} \sum_{i=1}^n x_i y_i &\xrightarrow{P} \mathbb{E}[x_i y_i]
 \end{aligned}$$

In addition, from Slutsky's theorem:

$$\frac{1}{n} \sum_{i=1}^n (X_i X_i') + \frac{\lambda}{n} I_k \xrightarrow{P} \mathbb{E}[X_i X_i']$$

and by the continuous mapping theorem:

$$\left(\frac{1}{n} \sum_{i=1}^n (X_i X_i') + \frac{\lambda}{n} I_k \right)^{-1} \xrightarrow{P} \mathbb{E}[X_i X_i']^{-1}$$

So by using Slutsky's theorem again:

$$\begin{aligned} \left(\frac{1}{n} \sum_{i=1}^n (X_i X_i') + \frac{\lambda}{n} I_k \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i Y_i \right) &\xrightarrow{P} \mathbb{E}[X_i X_i]^{-1} \mathbb{E}[X_i Y_i] \\ \hat{\beta} &\xrightarrow{P} \beta \end{aligned}$$

So the ridge regression estimator is consistent.

Chapter 4

Applying OLS Fundamentals

4.1 Practice Problem: Hansen 7.7

Of the variables (Y^*, Y, X) , only (Y, X) are observed. In this case, we say that Y^* is a latent variable. Suppose

$$\begin{aligned} Y^* &= X'\beta + \varepsilon \\ \mathbb{E}[X\varepsilon] &= 0 \\ Y &= Y^* + u \end{aligned}$$

u is a measurement error and satisfies:

$$\begin{aligned} \mathbb{E}[Xu] &= 0 \\ \mathbb{E}[Y^*u] &= 0 \end{aligned}$$

Denote the OLS coefficient from the regression of Y on X as $\hat{\beta}$.

- (a) Is β the coefficient from the linear projection of Y on X ?
- (b) Is $\hat{\beta}$ consistent for β as $n \rightarrow \infty$?
- (c) Find the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta)$ as $n \rightarrow \infty$.

4.1.1 Part a

Starting with β :

$$\begin{aligned} \beta &= \mathbb{E}[x_i x_i']^{-1} \mathbb{E}[x_i y_i^*] \\ &= \mathbb{E}[x_i x_i']^{-1} \mathbb{E}[x_i (y_i - u_i)] \\ &= \mathbb{E}[x_i x_i']^{-1} \mathbb{E}[x_i y_i] - \mathbb{E}[x_i x_i']^{-1} \mathbb{E}[x_i u_i] \\ &= \mathbb{E}[x_i x_i']^{-1} \mathbb{E}[x_i y_i] \end{aligned}$$

So the true β is the same as the linear projection.

4.1.2 Part b

Use the analogy principle and then go from there:

$$\begin{aligned}
\hat{\beta} &= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) \\
&= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i (y_i^* + u_i) \right) \\
&= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i y_i^* \right) + \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i u_i \right) \\
&= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i (x_i' \beta + \varepsilon_i) \right) + \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i u_i \right) \\
&= \beta + \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i \varepsilon_i \right) + \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i u_i \right) \quad (*) \\
&\xrightarrow{P} \beta + \mathbb{E}[x_i x_i']^{-1} \mathbb{E}[x_i \varepsilon_i] + \mathbb{E}[x_i x_i']^{-1} \mathbb{E}[x_i u_i] \\
&= \beta
\end{aligned}$$

So $\hat{\beta} \xrightarrow{P} \beta$.

4.1.3 Part c

Starting from (*), subtract β from both sides:

$$\begin{aligned}
\hat{\beta} - \beta &= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i \varepsilon_i \right) + \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i u_i \right) \\
\sqrt{n} (\hat{\beta} - \beta) &= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \varepsilon_i \right) + \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i u_i \right) \\
&= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i (\varepsilon_i + u_i) \right) \quad (1)
\end{aligned}$$

First, we must show that the mean of our \sqrt{n} term is zero:

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n x_i (\varepsilon_i + u_i) \right] &= \frac{1}{n} \sum_{i=1}^n (\mathbb{E}[x_i \varepsilon_i] + \mathbb{E}[x_i u_i]) \\
&= 0
\end{aligned}$$

We know that the multivariate central limit theorem will therefore hold. Now we must find the variance:

$$\begin{aligned}
 \text{Var} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i (\varepsilon_i + u_i) \right) &= \frac{1}{n} \text{Var} \left(\sum_{i=1}^n (x_i \varepsilon_i + x_i u_i) \right) \\
 &= \frac{1}{n} \left\{ \mathbb{E} \left[\text{Var} \left(\sum_{i=1}^n (x_i \varepsilon_i + x_i u_i) \middle| x \right) \right] \right. \\
 &\quad \left. + \text{Var} \left(\mathbb{E} \left[\sum_{i=1}^n (x_i \varepsilon_i + x_i u_i) \middle| x \right] \right) \right\} \\
 &= \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n x_i \text{Var} (\varepsilon_i + u_i | x) x_i' \right] \tag{2}
 \end{aligned}$$

Before we go forward, we must additionally assume that u_i and ε_i are independent from each other so that the conditional covariance term is zero.

Continuing on from equation (2), we now make the standard assumption that all of our variables are *i.i.d.* and the errors are homoskedastic:

$$\frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n x_i \text{Var} (\varepsilon_i + u_i | x) x_i' \right] = \mathbb{E}[x_i (\sigma_\varepsilon^2 + \sigma_u^2) x_i']$$

And now we go back to equation (1):

$$\begin{aligned}
 \sqrt{n} (\hat{\beta} - \beta) &\xrightarrow{d} \mathbb{E}[x_i x_i']^{-1} N(0, \mathbb{E}[x_i (\sigma_\varepsilon^2 + \sigma_u^2) x_i']) \\
 &= N(0, (\sigma_\varepsilon^2 + \sigma_u^2) \mathbb{E}[x_i x_i']^{-1})
 \end{aligned}$$

4.2 Delta Method

If $\sqrt{n}(x_n - x) \xrightarrow{d} \xi$ and $h(u)$ is a continuously differentiable function, then:

$$\sqrt{n}(h(x_n) - h(x)) \xrightarrow{d} H' \xi$$

where $H = \frac{\partial}{\partial u} h(u)^T$. In particular, if $\xi \sim N(0, V)$, then:

$$\sqrt{n}(h(x_n) - h(x)) \xrightarrow{d} N(0, H' V H)$$

4.3 Practice Problem: Intro Hansen 8.8

Assume that:

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 - \theta_2 \end{pmatrix} \xrightarrow{d} N(0, \Sigma)$$

Use the Delta Method to find the asymptotic distribution of the following statistics:

- (a) $\hat{\theta}_1 \hat{\theta}_2$
- (b) $e^{\hat{\theta}_1 + \hat{\theta}_2}$
- (c) If $\theta_2 \neq 0$, $\frac{\hat{\theta}_1}{\hat{\theta}_2^2}$
- (d) $\hat{\theta}_1^3 + \hat{\theta}_1 \hat{\theta}_2^2$

4.3.1 Part a

Using Slutsky's Theorem, we know that $\hat{\theta}_1 \hat{\theta}_2 \xrightarrow{P} \theta_1 \theta_2$ since $\hat{\theta}_1 \xrightarrow{P} \theta_1$ and $\hat{\theta}_2 \xrightarrow{P} \theta_2$ from the set-up of the problem. So we only need to find the asymptotic variance. Using the Delta Method, we start with H :

$$\begin{aligned} H &= \frac{\partial}{\partial \boldsymbol{\theta}} \theta_1 \theta_2 \\ &= \begin{bmatrix} \theta_2 \\ \theta_1 \end{bmatrix} \end{aligned}$$

We are given that $V = \Sigma$, so the asymptotic distribution is:

$$\sqrt{n}(\hat{\theta}_1 \hat{\theta}_2 - \theta_1 \theta_2) \xrightarrow{d} N \left(0, \begin{bmatrix} \theta_2 \\ \theta_1 \end{bmatrix}^T \Sigma \begin{bmatrix} \theta_2 \\ \theta_1 \end{bmatrix} \right)$$

4.3.2 Part b

We know that $e^{\hat{\theta}_1 + \hat{\theta}_2} \xrightarrow{P} e^{\theta_1 + \theta_2}$ by the Continuous Mapping Theorem since $\hat{\theta}_1 + \hat{\theta}_2 \xrightarrow{P} \theta_1 + \theta_2$ by Slutsky's Theorem and the exponential transformation is continuous. We now find H :

$$\begin{aligned} H &= \frac{\partial}{\partial \boldsymbol{\theta}} e^{\theta_1 + \theta_2} \\ &= \begin{bmatrix} e^{\theta_1 + \theta_2} \\ e^{\theta_1 + \theta_2} \end{bmatrix} \end{aligned}$$

Putting this altogether:

$$\sqrt{n} \left(e^{\hat{\theta}_1 + \hat{\theta}_2} - e^{\theta_1 + \theta_2} \right) \xrightarrow{d} N \left(0, \begin{bmatrix} e^{\theta_1 + \theta_2} \\ e^{\theta_1 + \theta_2} \end{bmatrix}^T \Sigma \begin{bmatrix} e^{\theta_1 + \theta_2} \\ e^{\theta_1 + \theta_2} \end{bmatrix} \right)$$

4.3.3 Part c

We know that $\frac{\hat{\theta}_1}{\hat{\theta}_2^2} \xrightarrow{P} \frac{\theta_1}{\theta_2^2}$ by the CMT, as division is continuous as long as the denominator is not zero (which we are given), since by the CMT $\hat{\theta}_2^2 \xrightarrow{P} \theta_2^2$. We now find H :

$$\begin{aligned}
 H &= \frac{\partial \theta_1}{\partial \theta \theta_2^2} \\
 &= \begin{bmatrix} \frac{1}{\theta_2^2} \\ -\frac{2\theta_1}{\theta_2^3} \end{bmatrix}
 \end{aligned}$$

Altogether, we now have:

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2^2 - \theta_2^2 \end{pmatrix} \xrightarrow{d} N \left(0, \begin{bmatrix} \frac{1}{\theta_2^2} \\ -\frac{2\theta_1}{\theta_2^3} \end{bmatrix}^T \Sigma \begin{bmatrix} \frac{1}{\theta_2^2} \\ -\frac{2\theta_1}{\theta_2^3} \end{bmatrix} \right)$$

4.3.4 Part d

We know that $\hat{\theta}_1^3 + \hat{\theta}_1 \hat{\theta}_2^2 \xrightarrow{P} \theta_1^3 + \theta_1 \theta_2^2$ by the CMT since $\hat{\theta}_1^3 \xrightarrow{P} \theta_1^3$ by the CMT, $\hat{\theta}_2^2 \xrightarrow{P} \theta_2^2$ by the CMT, and $\hat{\theta}_1 \hat{\theta}_2^2 \xrightarrow{P} \theta_1 \theta_2^2$ by Slutsky's theorem. We now find H :

$$\begin{aligned}
 H &= \frac{\partial}{\partial \theta} (\theta_1^3 + \theta_1 \theta_2^2) \\
 &= \begin{bmatrix} 3\theta_1^2 + \theta_2^2 \\ 2\theta_1 \theta_2 \end{bmatrix}
 \end{aligned}$$

Now pulling this all together:

$$\sqrt{n} \left(\hat{\theta}_1^3 + \hat{\theta}_1 \hat{\theta}_2^2 - (\theta_1^3 + \theta_1 \theta_2^2) \right) \xrightarrow{d} N \left(0, \begin{bmatrix} 3\theta_1^2 + \theta_2^2 \\ 2\theta_1 \theta_2 \end{bmatrix}^T \Sigma \begin{bmatrix} 3\theta_1^2 + \theta_2^2 \\ 2\theta_1 \theta_2 \end{bmatrix} \right)$$

4.4 Practice Problem: Hansen 3.13

Let D_1 and D_2 be vectors of ones and zeroes, with the i^{th} element of D_1 equaling one if that observation is male and zero if that observation is female (D_2 being the opposite). Then:

(a) In the OLS regression

$$Y = D_1 \hat{\gamma}_1 + D_2 \hat{\gamma}_2 + \hat{\mu}$$

show that $\hat{\gamma}_1$ is the sample mean of the dependent variable among men in the sample and that $\hat{\gamma}_2$ is the sample mean among women.

(b) Let $X_{n \times k}$ be an additional matrix of regressors. Describe in words the transformations

$$\begin{aligned} Y^* &= Y - D_1 \bar{Y}_1 - D_2 \bar{Y}_2 \\ X^* &= X - D_1 \bar{X}_1' - D_2 \bar{X}_2' \end{aligned}$$

Where \bar{X}_1 and \bar{X}_2 are the $k \times 1$ means of the regressors for men and women, respectively.

(c) Compare $\tilde{\beta}$ from the OLS regression

$$Y^* = X^* \tilde{\beta} + \tilde{e}$$

with the $\hat{\beta}$ from the OLS regression

$$Y = D_1 \hat{\alpha}_1 + D_2 \hat{\alpha}_2 + X \hat{\beta} + \hat{e}$$

4.4.1 Part a

We first take the general formula for the OLS estimator:

$$\beta = (X'X)^{-1}(X'Y)$$

Looking at the OLS equation we are estimating, we see that

$$X = \begin{bmatrix} D_1 & D_2 \end{bmatrix}$$

Calculating $X'X$ then:

$$\begin{aligned} X'X &= \begin{bmatrix} D_1' D_1 & D_1' D_2 \\ D_2' D_1 & D_2' D_2 \end{bmatrix} \\ &= \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix} \end{aligned}$$

We can then invert this by using properties of a diagonal block matrix:

$$(X'X)^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix}$$

Now we need to find the second part of the OLS estimator:

$$\begin{aligned} X'Y &= \begin{bmatrix} D_1'Y \\ D_2'Y \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n y_i \mathbb{1}(d_{1,i} = 1) \\ \sum_{i=1}^n y_i \mathbb{1}(d_{2,i} = 1) \end{bmatrix} \end{aligned}$$

Combining the two pieces together, we get:

$$\begin{aligned} \hat{\beta}_{OLS} &= \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \mathbb{1}(d_{1,i} = 1) \\ \sum_{i=1}^n y_i \mathbb{1}(d_{2,i} = 1) \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n_1} \sum_{i=1}^n y_i \mathbb{1}(d_{1,i} = 1) \\ \frac{1}{n_2} \sum_{i=1}^n y_i \mathbb{1}(d_{2,i} = 1) \end{bmatrix} \\ &= \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{bmatrix} \end{aligned}$$

4.4.2 Part b

The first transformation demeans y so that y^* has a sample mean of zero. The second transformation demeans X so that X^* has a sample mean of zero. When running regressions with these variables, the β coefficients are now deviations from the mean of the data.

4.4.3 Part c

Let's start from the second OLS regression. We can rewrite this equation as

$$\begin{aligned} Y &= X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \hat{e} \\ X_1 &= \begin{bmatrix} D_1 & D_2 \end{bmatrix} \\ \hat{\beta}_1 &= \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} \\ X_2 &= X \\ \hat{\beta}_2 &= \hat{\beta} \end{aligned}$$

Now, we don't care about X_1 , so we can use the elimination matrix to simplify the problem. Define $M_1 \equiv I - X_1'(X_1'X_1)^{-1}X_1$. We can now write $\hat{\beta}_2$ as:

$$\hat{\beta}_2 = ((M_1X)'M_1X)^{-1}(M_1X)'M_1Y$$

Okay, let's look at this equation piece-by-piece. First, we look at M_1Y :

$$\begin{aligned}
M_1 Y &= (I - X_1'(X_1'X_1)^{-1}X_1) Y \\
&= Y - X_1'(X_1'X_1)^{-1}X_1 Y \\
&= Y - X_1' \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{bmatrix} \\
&= Y - D_1 \bar{Y}_1 - D_2 \bar{Y}_2 \\
&= Y^*
\end{aligned}$$

Now we can look at $M_1 X$:

$$\begin{aligned}
M_1 X &= (I - X_1'(X_1'X_1)^{-1}X_1) X \\
&= X - X_1'(X_1'X_1)^{-1}X_1 X \\
&= X - X_1' \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n x_i \mathbb{1}(d_{1,i} = 1) \\ \sum_{i=1}^n x_i \mathbb{1}(d_{2,i} = 1) \end{bmatrix} \\
&= X - X_1' \begin{bmatrix} \bar{x}'_1 \\ \bar{x}'_2 \end{bmatrix} \\
&= X - D_1 \bar{x}'_1 - D_2 \bar{x}'_2 \\
&= X^*
\end{aligned}$$

Now we have all the pieces we need to find $\hat{\beta}_2$:

$$\begin{aligned}
\hat{\beta}_2 &= ((M_1 X)' M_1 X)^{-1} (M_1 X)' M_1 Y \\
&= ((X^*)' X^*)^{-1} ((X^*)' Y^*)
\end{aligned}$$

Turning to the first OLS regression, we apply our normal formula for the OLS estimator:

$$\tilde{\beta} = ((X^*)' X^*)^{-1} ((X^*)' Y^*)$$

which is the same estimator we derived from the second OLS regression. Therefore, the two regressions deliver the same results for the target β 's.

Chapter 5

More Regression

5.1 Practice Problem: Hansen 3.13

Let D_1 and D_2 be vectors of ones and zeroes, with the i^{th} element of D_1 equaling one if that observation is male and zero if that observation is female (D_2 being the opposite). Then:

- (a) In the OLS regression

$$Y = D_1\hat{\gamma}_1 + D_2\hat{\gamma}_2 + \hat{\mu}$$

show that $\hat{\gamma}_1$ is the sample mean of the dependent variable among men in the sample and that $\hat{\gamma}_2$ is the sample mean among women.

- (b) Let $X_{n \times k}$ be an additional matrix of regressors. Describe in words the transformations

$$\begin{aligned} Y^* &= Y - D_1\bar{Y}_1 - D_2\bar{Y}_2 \\ X^* &= X - D_1\bar{X}_1' - D_2\bar{X}_2' \end{aligned}$$

Where \bar{X}_1 and \bar{X}_2 are the $k \times 1$ means of the regressors for men and women, respectively.

- (c) Compare $\tilde{\beta}$ from the OLS regression

$$Y^* = X^*\tilde{\beta} + \tilde{e}$$

with the $\hat{\beta}$ from the OLS regression

$$Y = D_1\hat{\alpha}_1 + D_2\hat{\alpha}_2 + X\hat{\beta} + \hat{e}$$

5.1.1 Part a

We first take the general formula for the OLS estimator:

$$\beta = (X'X)^{-1}(X'Y)$$

Looking at the OLS equation we are estimating, we see that

$$X = \begin{bmatrix} D_1 & D_2 \end{bmatrix}$$

Calculating $X'X$ then:

$$\begin{aligned} X'X &= \begin{bmatrix} D_1'D_1 & D_1'D_2 \\ D_2'D_1 & D_2'D_2 \end{bmatrix} \\ &= \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix} \end{aligned}$$

We can then invert this by using properties of a diagonal block matrix:

$$(X'X)^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix}$$

Now we need to find the second part of the OLS estimator:

$$\begin{aligned} X'Y &= \begin{bmatrix} D_1'Y \\ D_2'Y \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n y_i \mathbb{1}(d_{1,i} = 1) \\ \sum_{i=1}^n y_i \mathbb{1}(d_{2,i} = 1) \end{bmatrix} \end{aligned}$$

Combining the two pieces together, we get:

$$\begin{aligned} \hat{\beta}_{OLS} &= \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \mathbb{1}(d_{1,i} = 1) \\ \sum_{i=1}^n y_i \mathbb{1}(d_{2,i} = 1) \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n_1} \sum_{i=1}^n y_i \mathbb{1}(d_{1,i} = 1) \\ \frac{1}{n_2} \sum_{i=1}^n y_i \mathbb{1}(d_{2,i} = 1) \end{bmatrix} \\ &= \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{bmatrix} \end{aligned}$$

5.1.2 Part b

The first transformation demeanes y so that y^* has a sample mean of zero. The second transformation demeanes X so that X^* has a sample mean of zero. When running regressions with these variables, the

β coefficients are now deviations from the mean of the data.

5.1.3 Part c

Let's start from the second OLS regression. We can rewrite this equation as

$$\begin{aligned} Y &= X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e} \\ X_1 &= \begin{bmatrix} D_1 & D_2 \end{bmatrix} \\ \hat{\beta}_1 &= \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} \\ X_2 &= X \\ \hat{\beta}_2 &= \hat{\beta} \end{aligned}$$

Now, we don't care about X_1 , so we can use the elimination matrix to simplify the problem. Define $M_1 \equiv I - X_1'(X_1'X_1)^{-1}X_1$. We can now write $\hat{\beta}_2$ as:

$$\hat{\beta}_2 = ((M_1X)'M_1X)^{-1}(M_1X)'M_1Y$$

Okay, let's look at this equation piece-by-piece. First, we look at M_1Y :

$$\begin{aligned} M_1Y &= (I - X_1'(X_1'X_1)^{-1}X_1)Y \\ &= Y - X_1'(X_1'X_1)^{-1}X_1Y \\ &= Y - X_1' \begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{bmatrix} \\ &= Y - D_1\bar{Y}_1 - D_2\bar{Y}_2 \\ &= Y^* \end{aligned}$$

Now we can look at M_1X :

$$\begin{aligned} M_1X &= (I - X_1'(X_1'X_1)^{-1}X_1)X \\ &= X - X_1'(X_1'X_1)^{-1}X_1X \\ &= X - X_1' \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n x_i \mathbb{1}(d_{1,i} = 1) \\ \sum_{i=1}^n x_i \mathbb{1}(d_{2,i} = 1) \end{bmatrix} \\ &= X - X_1' \begin{bmatrix} \bar{x}'_1 \\ \bar{x}'_2 \end{bmatrix} \\ &= X - D_1\bar{x}'_1 - D_2\bar{x}'_2 \\ &= X^* \end{aligned}$$

Now we have all the pieces we need to find $\hat{\beta}_2$:

$$\begin{aligned}\hat{\beta}_2 &= ((M_1 X)' M_1 X)^{-1} (M_1 X)' M_1 Y \\ &= ((X^*)' X^*)^{-1} ((X^*)' Y^*)\end{aligned}$$

Turning to the first OLS regression, we apply our normal formula for the OLS estimator:

$$\tilde{\beta} = ((X^*)' X^*)^{-1} ((X^*)' Y^*)$$

which is the same estimator we derived from the second OLS regression. Therefore, the two regressions deliver the same results for the target β 's.

5.2 Previous Problem: Question 2

Consider the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. If the true value of $\beta_0 = 0$, compare the variance of $\hat{\beta}_1$ with the variance of $\tilde{\beta}_1$, where $(\hat{\beta}_1, \tilde{\beta}_1)$ are the slope estimates of y on x_1 in models with and without an intercept, respectively.

5.2.1 Solution

We first derive the slope estimate $\hat{\beta}_1$:

$$\begin{aligned}\hat{\beta} &= (x'x)^{-1} (x'y) \\ &= \frac{1}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \\ &= \frac{1}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ -\sum_{i=1}^n x_i \sum_{i=1}^n y_i + n \sum_{i=1}^n x_i y_i \end{bmatrix}\end{aligned}$$

Let's look at just the bottom row of the matrix:

$$\begin{aligned}-\sum_{i=1}^n x_i \sum_{i=1}^n y_i + n \sum_{i=1}^n x_i y_i &= n \sum_{i=1}^n x_i y_i - n^2 \bar{x} \bar{y} \\ &= n \sum_{i=1}^n x_i y_i - 2n^2 \bar{x} \bar{y} + n^2 \bar{x} \bar{y} \\ &= n \sum_{i=1}^n x_i y_i - n \sum_{i=1}^n x_i \bar{y} - n \sum_{i=1}^n y_i \bar{x} + n \sum_{i=1}^n \bar{x} \bar{y} \\ &= n \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\end{aligned}$$

The denominator out front can be equated to the sample variance:

$$\begin{aligned}
 n \sum_{i=1}^n (x_i - \bar{x})^2 &= n \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\
 &= n \left(\sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 \right) \\
 &= n \left(\sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \right) \\
 &= n \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \\
 &= n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2
 \end{aligned}$$

Therefore:

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Now, let's find $\tilde{\beta}_1$:

$$\begin{aligned}
 \tilde{\beta}_1 &= (x'x)^{-1}(x'y) \\
 &= \left(\sum_{i=1}^n x_i^2 \right)^{-1} \left(\sum_{i=1}^n x_i y_i \right) \\
 &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}
 \end{aligned}$$

Now that we have our slope coefficients, we can compare the variances. First, let's take the conditional variance of $\hat{\beta}_1$:

$$\begin{aligned}
 Var(\hat{\beta}_1 | x) &= Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \middle| x \right) \\
 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 Var(y_i | x)}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} \\
 &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}
 \end{aligned}$$

Now we can move to the variance of $\tilde{\beta}_1$:

$$\begin{aligned} \text{Var} \left(\tilde{\beta}_1 \mid x \right) &= \text{Var} \left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \mid x \right) \\ &= \frac{\sum_{i=1}^n x_i^2 \text{Var}(y_i \mid x)}{\left(\sum_{i=1}^n x_i^2 \right)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n x_i^2} \end{aligned}$$

If $\bar{x} \neq 0$, then the denominator of the variance for $\hat{\beta}_1$ will be smaller. Therefore:

$$\text{Var} \left(\hat{\beta}_1 \mid x \right) \geq \text{Var} \left(\tilde{\beta}_1 \mid x \right)$$

5.3 Previous Problem: Question 5

Let the random variable X_n have a binomial distribution, $\text{Bin}(n, p)$. Then $\mathbb{E}[X_n] = np$ and $\text{Var}(X_n) = np(1-p)$.

- (a) Prove that X_n/n converges to p in probability using the Chebyshev Inequality.
- (b) Prove that $1 - X_n/n$ converges to $1 - p$ in probability.
- (c) Prove that $X_n/n(1 - X_n/n)$ converges to $p(1 - p)$ in probability. Use the theorem that if $X_n \xrightarrow{P} a$, then $g(X_n) \xrightarrow{P} g(a)$.

5.3.1 Part a

First, note that $\mathbb{E}[X_n/n] = p$ and $\text{Var}(X_n/n) = np(1-p)/n^2$. Now we can use the Chebyshev Inequality:

$$\begin{aligned} P \left(\left| \frac{X_n}{n} - p \right| \geq \varepsilon \right) &\leq \frac{\mathbb{E}[X_n - p]^2}{\varepsilon^2} \\ &= \frac{\text{Var}(X_n/n)}{\varepsilon^2} \\ &= \frac{p(1-p)}{n\varepsilon^2} \\ &\xrightarrow{P} 0 \end{aligned}$$

Therefore, since $\lim_{n \rightarrow \infty} P \left(\left| \frac{X_n}{n} - p \right| \geq \varepsilon \right) = 0$, we know that $X_n/n \xrightarrow{P} p$.

5.3.2 Part b

Again, note that $\mathbb{E}[1 - X_n/n] = 1 - p$ and $Var(1 - X_n/n) = np(1 - p)/n^2$. Apply the Chebyshev Inequality:

$$\begin{aligned} P\left(\left|1 - \frac{X_n}{n} - (1 - p)\right| \geq \varepsilon\right) &\leq \frac{\mathbb{E}[1 - X_n/n - (1 - p)]^2}{\varepsilon^2} \\ &= \frac{Var(1 - X_n/n)}{\varepsilon^2} \\ &= \frac{p(1 - p)}{n\varepsilon^2} \\ &\xrightarrow{P} 0 \end{aligned}$$

Therefore, since $\lim_{n \rightarrow \infty} P(|1 - \frac{X_n}{n} - (1 - p)| \geq \varepsilon) = 0$, we know that $1 - X_n/n \xrightarrow{P} 1 - p$.

5.3.3 Part c

Note that $\mathbb{E}[(X_n/n)^2] = p^2$. Then by the weak law of large numbers:

$$\left(\frac{X_n}{n}\right)^2 \xrightarrow{P} \mathbb{E}\left[\left(\frac{X_n}{n}\right)^2\right] = p^2$$

Now define $g(X_n) = X_n/n - (X_n/n)^2$. Then by the provided theorem:

$$\begin{aligned} g(X_n) &\xrightarrow{P} g(a) \\ &= p - p^2 \end{aligned}$$

5.4 Matlab Functions

In class, you were told that building a function in Matlab to calculate OLS estimates would be a good idea for your midterm. I'm going to go through constructing a function in Matlab to make sure that we are all on the same page.

There are two ways to create a function. First, you can include a functions section at the end of your file. This method works fine if you are only using that function in the current file.

Secondly, you can create a new Matlab file starting with the "function" command. As long as this file is in your directory, you can call this function in any other file. This method tends to be more useful.

First, let's look at the function syntax:

```
function [ols, se, tstat, ci95] = olsreg(x,y,variables,newey)
```

I declare that I am writing a function. Then, in brackets, I list the output variables I want from the function. Note that these output variables must be named inside the function. Next, I say that

these output function come from the function, in this case “olsreg”, consisting of the input variables “x”, “y”, “variables”, and “newey.”

To call this function, let’s look at my Matlab file:

```
clear, close all
newey = 0;

% Load data
load taylor.mat;

start = 25; % 25 for FF, 49 for SR
stop = 160; % 172 is the end, 160 for pre-covid (2018Q4)

y = taylor.ff(start:stop); % Lhs variable
T = size(y,1); % Number of time periods

% Include a lag of the federal funds rate
lagff1 = lagmatrix(taylor.ff,1);

x = [taylor.election1(start:stop), taylor.pceinf(start:stop),...
     taylor.Unemployment(start:stop), log(taylor.uncertainty(start:stop)),...
     lagff1(start:stop), ones(T,1)];

variables = ["Election Cycle"; "Inflation"; "Unemployment"; "Uncertainty"; "Lagged Rate"; "Constant"];

[beta, rse, tstat, ci95] = olsreg(x,y,variables,newey);
```

I first define the input variables. “newey” is defined in the second line. “y” is defined as the federal funds rate. “x” is the matrix of right-hand side variables. “variables” is a string vector of my right-hand side variable names.

At the end of the file, I say that I want to save the OLS coefficients as “beta”, the standard errors as “rse”, the t-stats as “t-stat”, and the 95% confidence interval as “ci95.” I then call the function name and put in my input variables. Now let’s look at the function file:

```
function [ols, se, tstat, ci95] = olsreg(x,y,variables,newey)

k = size(x,2);
T = size(y,1);
% OLS estimator
ols = inv(x'*x)*(x'*y);

% Residuals
resid = y - x*ols;

if newey == 1

    % NW estimator
    p=1;
    om0 = zeros(k,k);
    for t = 1:T
        om0 = om0 + resid(t)*resid(t)*x(t,:)'*x(t,:);
    end
    om0 = om0;
    om1 = zeros(k,k);
    for lags = 1:p
        for t = p+1:T
            om1 = om1 + (1-lags/(p+1))*resid(t)*resid(t-lags)*x(t,:)'*x(t-lags,:) + x(t-lags,:)'*x(t,:);
        end
    end
    om1 = om1;

end
```

Right away, I use the input variables “x” and “y” to define “T”, the length of the time series, and “k”, the number of right-hand side variables. I then calculate the ols estimator vector. I then need my measure of precision, the standard error. Because I am dealing with time series, I include the option to use the Newey-West variance calculation. My indicator variable, “newey”, is turned off though, so I use robust standard errors, as seen below:

```

var = om0 + om1;

else
|
% Calculate White standard errors
var = zeros(k,k);
for t = 1:T
    var = var + x(t,:)'*resid(t)*resid(t)*x(t,:);
end

end

avar = T/(T-k)*inv(x'*x)*var*inv(x'*x);

se = sqrt(diag(avar));

% t-stats
for i = 1:k
    tstat(i) = ols(i)/se(i);
    p(i) = 2*(1-tcdf(abs(tstat(i)),T-k));

% CIs
lb = ols(i) - tinv(1-(1-.95)/2,T-k)*se(i);
ub = ols(i) + tinv(1-(1-.95)/2,T-k)*se(i);
ci95(i,:) = [lb, ub];

```

I can then calculate the standard errors. From there, I can calculate the t-statistics, p-values, and confidence intervals.

```

lb = ols(i) - tinv(1-(1-.90)/2,T-k)*se(i);
ub = ols(i) + tinv(1-(1-.90)/2,T-k)*se(i);
ci90(i,:) = [lb, ub];
end
names = ["Variable", "Coef.", "Robust Std. Err.",...
        "t", "P value", "95 Conf. Int. (lower)", "95 Conf. Int. (upper)"];
Regression_table = table(variables, ols, se, tstat, p, ci95(:,1),...
        ci95(:,2), 'VariableNames',names);
display(Regression_table)

```

Lastly, I build the regression table and display it, so that when the function is finished, the table will appear in the console. You should make your own function for the exam, not copy mine. But you can use my function as inspiration for your own function.

Chapter 6

Algebra Review

6.1 Previous Problem: Homework 3 Question 1

Consider a simple regression model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where $\varepsilon_i \sim \text{Exp}(\lambda)$ with pdf:

$$f(\varepsilon_i|x_i) = \frac{1}{\lambda}e^{-\varepsilon_i/\lambda}, \quad \varepsilon_i \geq 0$$

Show that the OLS estimate for β is still unbiased but that $\hat{\alpha}$ is now biased.

6.1.1 Solution - Finding the Expectation

First we must find the expected value of an exponential distribution:

$$\mathbb{E}[\varepsilon_i|x_i] = \int_0^{\infty} \frac{\varepsilon_i}{\lambda} e^{-\varepsilon_i/\lambda} d\varepsilon_i$$

Using integration by parts:

$$\begin{aligned} u &= \varepsilon_i & dv &= e^{-\varepsilon_i/\lambda} d\varepsilon_i \\ du &= d\varepsilon_i & v &= -\lambda e^{-\varepsilon_i/\lambda} \end{aligned}$$

Then the integral becomes:

$$= \frac{1}{\lambda} \left(\left[-\varepsilon_i \lambda e^{-\varepsilon_i/\lambda} \right]_0^{\infty} + \lambda \int_0^{\infty} e^{-\varepsilon_i/\lambda} d\varepsilon_i \right)$$

To evaluate the first part, we need rearrange it into an indeterminate form:

$$\lim_{\varepsilon_i \rightarrow \infty} -\varepsilon_i \lambda e^{-\varepsilon_i/\lambda} = \frac{-\varepsilon_i \lambda}{e^{\varepsilon_i/\lambda}}$$

This form is ∞/∞ . We can now use L'Hopital's rule:

$$\begin{aligned} \lim_{\varepsilon_i \rightarrow \infty} \frac{-\varepsilon_i \lambda}{e^{\varepsilon_i/\lambda}} &= \lim_{\varepsilon_i \rightarrow \infty} \frac{-\lambda}{\frac{1}{\lambda} e^{\varepsilon_i/\lambda}} \\ &= 0 \end{aligned}$$

Going back to the expectation:

$$\begin{aligned} &= \frac{\lambda}{\lambda} \int_0^{\infty} e^{-\varepsilon_i/\lambda} d\varepsilon_i \\ &= -\lambda e^{-\varepsilon_i/\lambda} \Big|_0^{\infty} \\ &= \lambda \end{aligned}$$

6.1.2 Solution: Evaluating the Bias

We start with the slope coefficient that we derived last week in recitation:

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}) (\alpha + \beta x_i + \varepsilon_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}) \alpha + \sum_{i=1}^n (x_i - \bar{x}) \beta x_i + \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{(n\bar{x} - n\bar{x})\alpha + \sum_{i=1}^n (x_i^2 - x_i \bar{x})\beta + \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i^2 - n\bar{x}^2)\beta + \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \beta + \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \beta + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Now take the expectation:

$$\mathbb{E} [\hat{\beta}] = \beta + \mathbb{E} \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\begin{aligned}
&= \beta + \mathbb{E}_x \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}[\varepsilon_i | x]}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \\
&= \beta + \mathbb{E}_x \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) \lambda}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \\
&= \beta
\end{aligned}$$

So the slope coefficient is still unbiased. Now we look at the intercept term. As a refresher, let's derive it from first principles. We want to minimize the sum of squared residuals:

$$\begin{aligned}
\min_{\{\hat{\alpha}\}} \sum_{i=1}^n \hat{\varepsilon}_i &= \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \\
\frac{\partial SSR}{\partial \hat{\alpha}} &= -2 \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \\
n\bar{y} - n\hat{\alpha} - \hat{\beta}n\bar{x} &= 0 \\
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}
\end{aligned}$$

Now take the expectation:

$$\begin{aligned}
\mathbb{E}[\hat{\alpha}] &= \mathbb{E}[\bar{y}] - \mathbb{E}[\hat{\beta}\bar{x}] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\alpha + \beta x_i + \varepsilon_i] - \frac{1}{n} \mathbb{E} \left[\hat{\beta} \sum_{i=1}^n x_i \right] \\
&= \alpha + \beta \mathbb{E}[x_i] + \mathbb{E}_x[\varepsilon_i | x] - \frac{1}{n} \mathbb{E} \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n x_i \right] \\
&= \alpha + \beta \mathbb{E}[x_i] + \lambda - \frac{1}{n} \mathbb{E} \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) (\alpha + \beta x_i + \varepsilon_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n x_i \right] \\
&= \alpha + \beta \mathbb{E}[x_i] + \lambda - \frac{1}{n} \mathbb{E} \left[\beta \sum_{i=1}^n x_i + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n x_i \right] \\
&= \alpha + \beta \mathbb{E}[x_i] + \lambda - \frac{1}{n} \mathbb{E}_x \left[\beta \sum_{i=1}^n x_i + \frac{\sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}[\varepsilon_i | x]}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n x_i \right] \\
&= \alpha + \beta \mathbb{E}[x_i] + \lambda - \frac{1}{n} \mathbb{E}_x \left[\beta \sum_{i=1}^n x_i + \frac{\sum_{i=1}^n (x_i - \bar{x}) \lambda}{\sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n x_i \right] \\
&= \alpha + \beta \mathbb{E}[x_i] + \lambda - \beta \mathbb{E}[x_i] \\
&= \alpha + \lambda
\end{aligned}$$

6.2 Previous Question: Homework 3 Question 2

Consider the simple regression model:

$$y_i = \alpha + \varepsilon_i$$

We estimate α via OLS.

(a) Show that $\hat{\alpha} = \bar{y}$. Also show that $\hat{\alpha}$ is consistent and asymptotically normal.

(b) Consider an alternative estimate $\tilde{\alpha} = \sum_{i=1}^n w_i y_i$, where:

$$w_i = \frac{i}{n(n+1)/2} = \frac{i}{\sum_{i=1}^n i}$$

This is a weighted sample mean of y . Prove that $\tilde{\alpha}$ is consistent and obtain its asymptotic variance. Note that $\sum_{i=1}^n w_i = 1$ and $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$.

6.2.1 Part a

Using the OLS estimator:

$$\begin{aligned} \hat{\alpha} &= (x'x)^{-1}(x'y) \\ &= \left(\sum_{i=1}^n (1)^2 \right)^{-1} \left(\sum_{i=1}^n y_i \right) \\ &= \frac{1}{n} \sum_{i=1}^n y_i \\ &= \bar{y} \end{aligned}$$

Now we can appeal to asymptotics:

$$\begin{aligned} \hat{\alpha} &= \frac{1}{n} \sum_{i=1}^n y_i \\ &= \frac{1}{n} \sum_{i=1}^n (\alpha + \varepsilon_i) \\ &= \alpha + \frac{1}{n} \sum_{i=1}^n \varepsilon_i \\ &\xrightarrow{P} \alpha + \mathbb{E}[\varepsilon_i] \\ &= \alpha + \mathbb{E}_x[\mathbb{E}[\varepsilon_i|x]] \\ \hat{\alpha} &\xrightarrow{P} \alpha \end{aligned} \tag{1}$$

So α is consistent. Going back to equation (1):

$$\begin{aligned}\hat{\alpha} &= \alpha + \frac{1}{n} \sum_{i=1}^n \varepsilon_i \\ \sqrt{n}(\hat{\alpha} - \alpha) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \\ \sqrt{n}(\hat{\alpha} - \alpha) &\xrightarrow{d} N(0, \sigma^2)\end{aligned}$$

6.2.2 Part b

Let's simplify $\tilde{\alpha}$:

$$\begin{aligned}\tilde{\alpha} &= \sum_{i=1}^n w_i y_i \\ &= \sum_{i=1}^n w_i (\alpha + \varepsilon_i) \\ &= \alpha + \sum_{i=1}^n w_i \varepsilon_i\end{aligned}$$

Now we go the Chebyshev inequality:

$$\begin{aligned}P(|\tilde{\alpha} - \alpha| \geq \chi) &\leq \frac{\mathbb{E}[(\tilde{\alpha} - \alpha)^2]}{\chi^2} \\ &= \frac{\mathbb{E}[\tilde{\alpha}^2 - 2\alpha\tilde{\alpha} + \alpha^2]}{\chi^2} \\ &= \frac{\mathbb{E}\left[\alpha^2 + 2\alpha \sum_{i=1}^n w_i \varepsilon_i + \left(\sum_{i=1}^n w_i \varepsilon_i\right)^2 - 2\alpha^2 - 2\alpha \sum_{i=1}^n w_i \varepsilon_i + \alpha^2\right]}{\chi^2} \\ &= \frac{\mathbb{E}\left[\left(\sum_{i=1}^n w_i \varepsilon_i\right)^2\right]}{\chi^2}\end{aligned}$$

Here, we note that:

$$\mathbb{E}\left[\sum_{i=1}^n w_i \varepsilon_i\right] = \mathbb{E}_x\left[\sum_{i=1}^n w_i \mathbb{E}[\varepsilon_i | x]\right] = 0$$

So $\mathbb{E}\left[\left(\sum_{i=1}^n w_i \varepsilon_i\right)^2\right] = \text{Var}\left(\sum_{i=1}^n w_i \varepsilon_i\right)$. Continuing on:

$$\frac{\mathbb{E}\left[\left(\sum_{i=1}^n w_i \varepsilon_i\right)^2\right]}{\chi^2} = \frac{\text{Var}\left(\sum_{i=1}^n w_i \varepsilon_i\right)}{\chi^2}$$

$$\begin{aligned}
&= \frac{\sum_{i=1}^n w_i^2 \text{Var}(\varepsilon_i)}{\chi^2} \\
&= \frac{\sigma^2}{\chi^2} \sum_{i=1}^n \frac{i^2}{(n(n+1)/2)^2} \\
&= \frac{\sigma^2}{\chi^2} \frac{n(n+1)(2n+1)/6}{n^2(n+1)^2/4} \\
&= \frac{\sigma^2}{\chi^2} \frac{2n^2 + 3n + 1}{n^3 + 2n^2 + n} \cdot \frac{4}{6}
\end{aligned} \tag{2}$$

Taking n to infinity will generate an ∞/∞ form. Using L'Hopital's rule:

$$\begin{aligned}
\lim_{n \rightarrow \infty} \frac{\sigma^2}{\chi^2} \frac{2n^2 + 3n + 1}{n^3 + 2n^2 + n} \cdot \frac{4}{6} &= \lim_{n \rightarrow \infty} \frac{\sigma^2}{\chi^2} \frac{4n + 3}{3n^2 + 4n + 1} \cdot \frac{4}{6} \\
&= \lim_{n \rightarrow \infty} \frac{\sigma^2}{\chi^2} \frac{4n + 3}{3n^2 + 4n + 1} \cdot \frac{4}{6} \\
&= \lim_{n \rightarrow \infty} \frac{\sigma^2}{\chi^2} \frac{4}{6n + 4} \cdot \frac{4}{6} \\
&= 0
\end{aligned}$$

Therefore, as $n \rightarrow \infty$, $P(|\tilde{\alpha} - \alpha| \geq \chi) \leq 0$. By definition, then, $\tilde{\alpha} \xrightarrow{P} \alpha$. To find the asymptotic variance, we simply take $\text{Var}(\sqrt{n}(\tilde{\alpha} - \alpha))$:

$$\begin{aligned}
\text{Var}(\sqrt{n}(\tilde{\alpha} - \alpha)) &= n \text{Var}(\tilde{\alpha}) \\
&= n \text{Var}\left(\alpha + \sum_{i=1}^n w_i \varepsilon_i\right) \\
&= n \text{Var}\left(\sum_{i=1}^n w_i \varepsilon_i\right) && \text{(Use (2))} \\
&= \sigma^2 \frac{2n^3 + 3n^2 + n}{n^3 + 2n^2 + n} \cdot \frac{4}{6}
\end{aligned}$$

Once again, this is in an ∞/∞ form. Applying L'Hopital's rule:

$$\begin{aligned}
\lim_{n \rightarrow \infty} \sigma^2 \frac{2n^3 + 3n^2 + n}{n^3 + 2n^2 + n} \cdot \frac{4}{6} &= \lim_{n \rightarrow \infty} \sigma^2 \frac{6n^2 + 6n + 1}{3n^2 + 4n + 1} \cdot \frac{4}{6} \\
&= \lim_{n \rightarrow \infty} \sigma^2 \frac{12n + 6}{6n + 4} \cdot \frac{4}{6} \\
&= \lim_{n \rightarrow \infty} \sigma^2 \frac{12}{6} \cdot \frac{4}{6} \\
&= \frac{4\sigma^2}{3}
\end{aligned}$$

Therefore, because $\tilde{\alpha}$ is consistent, the asymptotic distribution is:

$$\sqrt{n}(\tilde{\alpha} - \alpha) \xrightarrow{d} N\left(0, \frac{4\sigma^2}{3}\right)$$

6.3 Violating the Exogeneity Assumption

Take the following regression specification:

$$y_i = x_i' \beta + \varepsilon_i$$

Define $x_i = [1, x_0, \dots, x_k]'$. Suppose that $\mathbb{E}[x_k \varepsilon] \neq 0$. The proof for consistency of our β estimator would therefore fail:

$$\begin{aligned} \hat{\beta} &= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) \\ &= \beta + \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i \varepsilon_i \right) \\ &\xrightarrow{P} \beta + \mathbb{E}[x_i x_i']^{-1} \mathbb{E}[x_i \varepsilon_i] \end{aligned}$$

This is as far as we can go without assuming orthogonality. Now suppose that we have a variable z_k that is uncorrelated with ε_i such that $\mathbb{E}[z_k \varepsilon_i] = 0$. Define z_i as $[1, x_1, \dots, x_{k-1}, z_k]'$. Go back to the standard regression equation:

$$y_i = x_i' \beta + \varepsilon_i$$

Instead of pre-multiplying by x_i , pre-multiply by z_i :

$$\begin{aligned} z_i y_i &= z_i x_i' \beta + z_i \varepsilon_i \\ \mathbb{E}[z_i y_i] &= \mathbb{E}[z_i x_i'] \beta + \mathbb{E}[z_i \varepsilon_i] \\ \mathbb{E}[z_i x_i']^{-1} \mathbb{E}[z_i y_i] &= \beta \end{aligned}$$

Apply the analogy principle:

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^n z_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n z_i y_i \right)$$

Let's see if this estimator is consistent:

$$\begin{aligned}
 \hat{\beta} &= \left(\frac{1}{n} \sum_{i=1}^n z_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n z_i (x_i' \beta + \varepsilon_i) \right) \\
 &= \beta + \left(\frac{1}{n} \sum_{i=1}^n z_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n z_i \varepsilon_i \right) \\
 &\xrightarrow{P} \beta + \mathbb{E}[z_i x_i']^{-1} \mathbb{E}[z_i \varepsilon_i] \\
 &= \beta
 \end{aligned}$$

So the estimator is consistent. Is it biased?

$$\begin{aligned}
 \mathbb{E}[\hat{\beta}] &= \beta + \mathbb{E} \left[\left(\frac{1}{n} \sum_{i=1}^n z_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n z_i \varepsilon_i \right) \right] \\
 &= \beta + \mathbb{E}_{z,x} \left[\left(\frac{1}{n} \sum_{i=1}^n z_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n z_i \mathbb{E}[\varepsilon_i | z, x] \right) \right]
 \end{aligned}$$

But we do not know if the inner expectation is zero. So it is safest to assume that this estimator is biased.

Chapter 7

Non-Linear Least Squares

7.1 Theory

Non-linear least squares is a topic I did not cover my first year in the program, but the idea behind it is cool. We start with the following general function:

$$y_i = h(x_i, \beta) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | x] = 0$$

As in linear least squares, we want to minimize the sum of squared residuals:

$$\min_{\{\beta\}} S(\beta) = \min_{\{\beta\}} \sum_{i=1}^n (y_i - h(x_i, \beta))^2$$

Taking the first order condition with respect to β yields:

$$-2 \sum_{i=1}^n (y_i - h(x_i, \beta)) \frac{\partial h}{\partial \beta} = 0$$

In ordinary least squares, we can solve this first order condition analytically. With non-linear least squares, we usually cannot derive an analytical solution. As such, the best we can do is:

$$\hat{\beta}_{NL} = \underset{\{\beta\}}{\operatorname{argmin}} S(\beta)$$

By definition of the least squares estimator, $\hat{\beta}_{NL}$ solves the first order condition. That is:

$$\frac{\partial S(\hat{\beta}_{NL})}{\partial \beta} = 0$$

7.1.1 Consistency

Despite not having an analytical solution, we can show that the non-linear least squares estimator is consistent. We need to make one more assumption: that β is identifiable. In other words, there exists

only one β , denoted by β_0 , that minimizes the sum of squared residuals.

Now, recall the continuous mapping theorem:

Theorem 3 (Continuous Mapping Theorem). *Let x_n be a sequence of real-valued random vectors and let $h : \mathcal{R}^k \rightarrow \mathcal{R}^m$. Define the set of discontinuous points as:*

$$D_h = \{x \in \mathcal{X} : h(\cdot) \text{ is discontinuous at } x\}$$

Now, if $P(x \in D_h) = 0$ and:

$$(i) \text{ if } x_n \xrightarrow{P} x, \text{ then } h(x_n) \xrightarrow{P} h(x)$$

$$(ii) \text{ if } x_n \xrightarrow{d} x, \text{ then } h(x_n) \xrightarrow{d} h(x)$$

$$(iii) \text{ if } x_n \xrightarrow{a.s.} x, \text{ then } h(x_n) \xrightarrow{a.s.} h(x)$$

The converse of this theorem holds true when $h(\cdot)$ is an injective (one-to-one) function. Stated another way:

Theorem 4 (Converse). *Let $h(\cdot)$ be a continuous, injective function such that $h(x_n) \xrightarrow{P} h(x)$. Then:*

$$x_n \xrightarrow{P} x$$

How do we plan on applying this converse? x_n in our scenario is $\hat{\beta}_{NL}$. We want to show that $\hat{\beta}_{NL} \xrightarrow{P} \beta$. We also have a continuous, one-to-one function in $\frac{\partial S(\hat{\beta})}{\partial \beta}$. We already know that:

$$\frac{\partial S(\hat{\beta})}{\partial \beta} = 0$$

If we take the probability limit:

$$\frac{1}{n} \frac{\partial S(\hat{\beta})}{\partial \beta} \xrightarrow{P} 0$$

Now we need to show that $\frac{\partial S(\beta_0)}{\partial \beta} = 0$ too. We begin with the sample mean of the derivative:

$$\begin{aligned} \frac{1}{n} \frac{\partial S(\beta_0)}{\partial \beta} &= \frac{1}{n} \sum_{i=1}^n (y_i - h(x_i, \beta)) \frac{\partial h(x_i, \beta)}{\partial \beta} \\ &\xrightarrow{P} \mathbb{E} \left[\varepsilon_i \frac{\partial h(x_i, \beta)}{\partial \beta} \right] \\ &= \mathbb{E}_x \left[\mathbb{E}[\varepsilon_i | x] \frac{\partial h(x_i, \beta)}{\partial \beta} \right] \\ &= 0 \end{aligned}$$

Therefore:

$$\frac{1}{n} \frac{\partial S(\hat{\beta}_{NL})}{\partial \beta} \xrightarrow{P} \frac{\partial S(\beta_0)}{\partial \beta}$$

By the converse of the continuous mapping theorem then:

$$\hat{\beta}_{NL} \xrightarrow{P} \beta_0$$

We have proven that the non-linear least squares estimator is consistent.

7.1.2 Asymptotic Normality

Start with the first order condition:

$$\frac{\partial S(\hat{\beta})}{\partial \beta} = -2 \sum_{i=1}^n (y_i - h(x_i, \beta)) \frac{\partial h(x_i, \beta)}{\partial \beta}$$

Denote the right-hand side as $g(\hat{\beta})$. Doing a first-order Taylor approximation around the true β delivers:

$$g(\hat{\beta}) = g(\beta) + \underline{H(\tilde{\beta})}(\hat{\beta} - \beta) \quad \text{for } \tilde{\beta} \in (\hat{\beta}, \beta) \quad (7.1)$$

Looking just at the Hessian (in red):

$$\begin{aligned} H(\tilde{\beta}) &= \frac{\partial^2 S(\tilde{\beta})}{\partial \beta \partial \beta'} \\ &= 2 \sum_{i=1}^n \frac{\partial h(x_i, \tilde{\beta})}{\partial \beta} \frac{\partial h(x_i, \tilde{\beta})}{\partial \beta'} - 2 \sum_{i=1}^n (y_i - h(x_i, \tilde{\beta})) \frac{\partial^2 h(x_i, \tilde{\beta})}{\partial \beta \partial \beta'} \\ \frac{1}{n} H(\tilde{\beta}) &\xrightarrow{P} 2\mathbb{E} \left[\frac{\partial h(x_i, \tilde{\beta})}{\partial \beta} \frac{\partial h(x_i, \tilde{\beta})}{\partial \beta'} \right] \\ &= 2Q_0 \end{aligned}$$

Going back to equation (1) and dividing by \sqrt{n} :

$$\underbrace{\frac{1}{\sqrt{n}} g(\hat{\beta})}_{=0} = \frac{1}{\sqrt{n}} g(\beta) + \frac{1}{n} H(\tilde{\beta}) \sqrt{n} (\hat{\beta} - \beta) \quad (7.2)$$

Now let's look at the gradient (in blue):

$$\begin{aligned}
\frac{1}{\sqrt{n}}g(\beta) &= \frac{-2}{\sqrt{n}} \sum_{i=1}^n (y_i - h(x_i, \beta)) \frac{\partial h(x_i, \beta)}{\partial \beta} \\
&= \frac{-2}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \frac{\partial h(x_i, \beta)}{\partial \beta} \\
&\xrightarrow{d} N \left(0, 4\mathbb{E} \left[\varepsilon_i \frac{\partial h(x_i, \beta)}{\partial \beta} \frac{\partial h(x_i, \beta)'}{\partial \beta} \varepsilon_i' \right] \right) \\
&= N \left(0, 4\mathbb{E}_x \left[\frac{\partial h(x_i, \beta)'}{\partial \beta} \mathbb{E}[\varepsilon_i \varepsilon_i' | x] \frac{\partial h(x_i, \beta)'}{\partial \beta} \right] \right) \\
&= N(0, 4\sigma^2 Q_0)
\end{aligned}$$

Rearranging equation (2) gives:

$$\begin{aligned}
\sqrt{n}(\hat{\beta} - \beta) &= -\frac{1}{n} H(\tilde{\beta})^{-1} \frac{1}{\sqrt{n}} g(\beta) \\
&\xrightarrow{d} \frac{1}{2} Q_0^{-1} N(0, 4\sigma^2 Q_0) \\
&= N(0, \sigma^2 Q_0^{-1})
\end{aligned}$$

This completes the asymptotic normality proof.

7.2 Example

Suppose we have the following regression specification:

$$y = \frac{\beta_1 x}{\beta_2 + x} + \varepsilon$$

We want to estimate this via non-linear least squares. To do so, we use the Newton-Raphson algorithm:

$$S(\theta) \approx S(\theta_0) + g(\theta_0)'(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)'H(\theta_0)(\theta - \theta_0)$$

where $S(\theta)$ is the function we want to approximate and θ is the estimated parameter. θ_0 is the guess made for the value of θ at the beginning of that iteration. To update our guesses, we use the following formula:

$$\theta_{i+1} = \theta_i - H(\theta_i)^{-1}g(\theta_i)$$

Lastly, we stop iterating when the algorithm converges:

$$\|\theta_{i+1} - \theta_i\| < tol$$

where tol is a tolerance parameter, usually set to 10^{-8} .

In our problem, θ is $\beta = [\beta_1, \beta_2]'$. $S(\theta)$ is the sum of squared residuals. Let's define the constructs we need:

$$e_i = y_i - \frac{\beta_1 x_i}{\beta_2 + x_i} \quad (\text{Residuals})$$

$$S(\beta) = \sum_{i=1}^n e_i^2 \quad (\text{SSR})$$

$$g(\beta) = \begin{bmatrix} \frac{-x_i}{\beta_2 + x_i} \\ \frac{\beta_1 x_i}{(\beta_2 + x_i)^2} \end{bmatrix} e_i \quad (\text{Gradient})$$

$$H(\beta) \approx \begin{bmatrix} \frac{-x_i}{\beta_2 + x_i} \\ \frac{\beta_1 x_i}{(\beta_2 + x_i)^2} \end{bmatrix}' \begin{bmatrix} \frac{-x_i}{\beta_2 + x_i} \\ \frac{\beta_1 x_i}{(\beta_2 + x_i)^2} \end{bmatrix} \quad (\text{Hessian})$$

To run this example, I use the following data set:¹

$$y = [0.05, 0.127, 0.094, 0.2122, 0.2729, 0.2665, 0.3317]'$$

$$x = [0.038, 0.194, 0.425, 0.626, 1.253, 2.5, 3.74]'$$

Below is a screenshot of the algorithm I wrote and a graph demonstrating the non-linear regression model. Note that before running the algorithm, I needed to set initial values for β_1 , iter, and diff.

```

while abs(diff) > 1e-15

    % Set last iteration's beta
    beta0 = beta1;

    % Calculate residuals
    e = y - beta0(1)*x./(beta0(2)+x);

    % Calculate the gradient
    grad1 = [-x./(beta0(2)+x), beta0(1)*x./(beta0(2) + x).^2]'*e;

    % Calculate the Hessian using the gradient squared
    hess1 = [-x./(beta0(2)+x), beta0(1)*x./(beta0(2) + x).^2]'...'
            *[-x./(beta0(2)+x), beta0(1)*x./(beta0(2) + x).^2];

    % Update beta guess
    beta1 = beta0 - (hess1)^(-1)*grad1;

    % Second-order Taylor approximation of FOC
    s = e'*e + grad1*(beta1 - beta0) + 1/2*(beta1-beta0)'*hess1*(beta1-beta0);

    % Difference
    diff = beta1 - beta0;

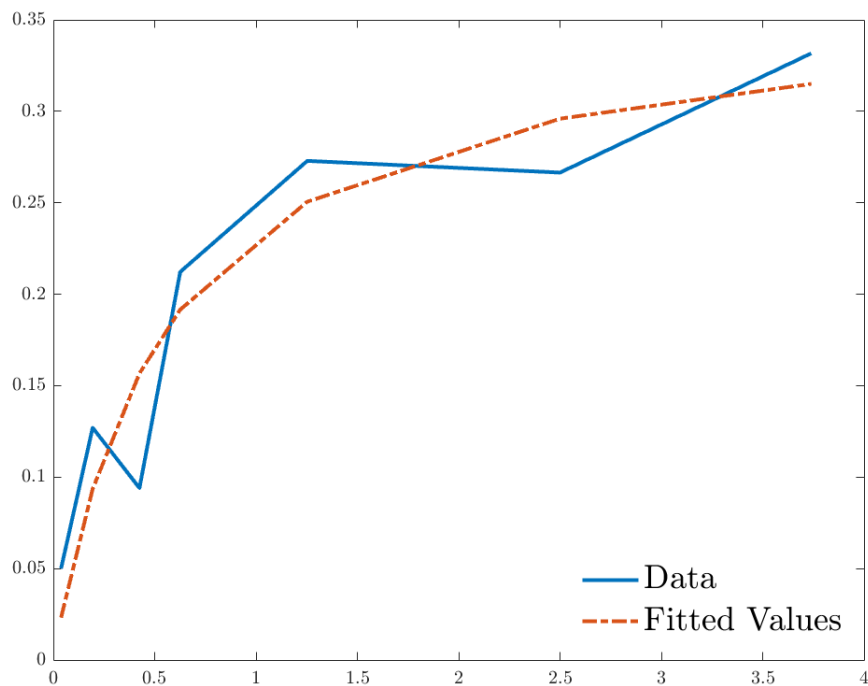
    % Count the iteration
    iter = iter + 1;

    % Display the iteration we are on
    display(iter)

end

```

¹These are taken from an example found [here](#).



7.3 Homework 4 F-Statistics

The F-statistic calculations in homework 4 were rough, so we're going to go through question 1, focusing specifically on the F-tests:

Test the null hypothesis of constant returns of scale of the Cobb-Douglas production function, $H_0 : \beta_1 + \beta_2 = 1$, in three different ways: 1) t-test 2) F-test using a set of linear restriction, 3) F-test by comparing the sum of squares of restricted and unrestricted regression. Compare these test statistics, and comment on your findings.

7.3.1 Part 1

We first need to run an OLS regression. Once we have the β coefficients and the variance-covariance matrix, we can then run a t-test on the null hypothesis.

```

===== OLS Regression Results =====
R2 = 0.943463 Rbar2 = 0.938751
Sum of Squared Residual = 0.851634
F-statistic( 2, 24) = 200.248945 ( 0.00000)
Durbin-Watson Statistic = 1.885989
Number of Observations = 27
=====
Variables      Estimates      Std. Err      t-stat      p-value
-----
Var 0          1.170644      0.326782      3.582339    0.001502
Var 1          0.602999      0.125954      4.787457    0.000071
Var 2          0.375710      0.085346      4.402204    0.000190
=====

```

Why can we run a t-test? We have one linear restriction, so the t-test squared is the F-test. Our t-test will look like:

$$\begin{aligned}
 t &= \frac{\beta_1 + \beta_2 - 1}{\sqrt{\text{Var}(\beta_1) + \text{Var}(\beta_2) + 2\text{Cov}(\beta_1, \beta_2)}} \\
 &= -0.3402 \\
 F &= 0.1158
 \end{aligned}$$

The other two ways of calculating the F-stat should deliver the same value.

7.3.2 Part 2

The next F-test is done via linear restriction. We are setting $\beta_1 + \beta_2 = 1$, so the restriction matrix will look like:

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

with a value, q , of 1. We now follow the formula:

$$\begin{aligned}
 F &= \frac{(R\hat{\beta} - q)' (R(x'x)^{-1}R')^{-1} (R\hat{\beta} - q)}{s^2r} \\
 &= (\hat{\beta}_1 + \hat{\beta}_2 - 1)' (Rs^2(x'x)^{-1}R')^{-1} (\hat{\beta}_1 + \hat{\beta}_2 - 1) \\
 &= (\hat{\beta}_1 + \hat{\beta}_2 - 1)' (R \text{Var}(\hat{\beta}) R')^{-1} (\hat{\beta}_1 + \hat{\beta}_2 - 1) \\
 &= 0.1158
 \end{aligned}$$

This statistic matches the one from Part 1.

7.3.3 Part 3

Now we want to run a restricted regression and obtain the sum of squared residuals. We can rewrite the restriction as:

$$\beta_2 = 1 - \beta_1$$

Sub this into the regression specification:

$$\begin{aligned} \ln(Y) &= \beta_0 + \beta_1 \ln(K) + \beta_2 \ln(L) + \varepsilon \\ \ln(Y) &= \beta_0 + \beta_1 \ln(K) + (1 - \beta_1) \ln(L) + \varepsilon \\ \ln(Y) - \ln(L) &= \beta_0 + \beta_1 (\ln(K) - \ln(L)) + \varepsilon \\ y^* &= \beta_0 + \beta_1 \ln(K^*) + \varepsilon \end{aligned}$$

Reformat the data to match this regression specification. Then calculate the SSR_r :

```

===== OLS Regression Results =====
R2          = 0.481076          Rbar2       = 0.460319
Sum of Squared Residual = 0.855741
F-statistic( 1, 25) = 23.176629 ( 0.00006)
Durbin-Watson Statistic = 1.903585
Number of Observations = 27
=====
Variables    Estimates    Std. Err    t-stat    p-value
-----
Var 0        1.069265    0.131759    8.115322  0.000000
Var 1        0.363030    0.075408    4.814211  0.000060
=====

```

Compare this to the SSR_{ur} from part 1:

$$SSR_{ur} = 0.8516$$

$$SSR_r = 0.8557$$

Use the F-test formula for sum of squares:

$$\begin{aligned} F &= \frac{(SSR_r - SSR_{ur})(n - k)}{SSR_{ur}r} \\ &= \frac{(0.8557 - 0.8516)(27 - 3)}{0.8516} \\ &= 0.1158 \end{aligned}$$

Note that k is the total number of regressors (3, as we have the constant, $\ln(K)$, and $\ln(L)$).

Chapter 8

Midterm Review

8.1 Frisch-Waugh-Lovell Theorem

Theorem 5 (Frisch-Waugh-Lovell). *Let the regression model be as follows:*

$$y = x_1\beta_1 + x_2\beta_2 + \varepsilon$$

Then the estimate for β_2 will be the same as the estimate from the following model:

$$M_1y = M_1x_2\beta_2 + M_1u \tag{1}$$

This yields a coefficient estimate of:

$$\hat{\beta}_2 = (x_2'M_1x_2)^{-1}(x_2'M_1y)$$

We can write the theorem in another, equivalent way. First, start from equation (1). Note that M_1y is the part of y not explained by x_1 . Therefore, M_1y represents the residuals from the following regression:

$$y = \beta x_1 + \varepsilon_1$$

Call those residuals $\hat{\varepsilon}_1$. Then look at M_1x_2 . This represents the residuals of the following regression:

$$x_2 = \pi x_1 + u_1$$

Call those residuals \tilde{x}_2 . Then the estimate for $\hat{\beta}_2$ becomes:

$$\hat{\beta}_2 = (\tilde{x}_2'\tilde{x}_2)^{-1}(\tilde{x}_2'\hat{\varepsilon}_1)$$

This is just the slope coefficient from the following regression:

$$\hat{\varepsilon}_1 = \beta_2 \tilde{x}_2 + \nu$$

8.1.1 When to use this?

We use the FWL theorem when we can partition the regression. For example, take the following specification:

$$y = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \alpha_4 z_4 + \varepsilon$$

Suppose we are only interested in α_1 . Define $x_{1i} = [1 \quad z_{2i} \quad z_{3i} \quad z_{4i}]'$ and $\beta_1 = [\alpha_0 \quad \alpha_2 \quad \alpha_3 \quad \alpha_4]'$. Then define $x_{2i} = z_{1i}'$ and $\beta_2 = \alpha_1$. This rewrites the regression as:

$$y = x_1 \beta_1 + x_2 \beta_2 + \varepsilon$$

Back in the early days of computing, computational power was limited. To save time, researchers would use the FWL theorem to find only the coefficient in which they were interested. Today, we use it as a regression exercise mostly. The theorem is nice for some proofs.

8.1.2 Homework 1, Question 4

Consider the following demand equation system:

$$\begin{aligned} E_d &= \alpha_d + \beta_d Y + \gamma_{dd} P_d + \gamma_{dn} P_n + \gamma_{ds} P_s + \varepsilon_d \\ E_n &= \alpha_n + \beta_n Y + \gamma_{nd} P_d + \gamma_{nn} P_n + \gamma_{ns} P_s + \varepsilon_n \\ E_s &= \alpha_s + \beta_s Y + \gamma_{sd} P_d + \gamma_{sn} P_n + \gamma_{ss} P_s + \varepsilon_s \end{aligned}$$

As defined, $Y = E_d + E_n + E_s$. Prove that $\beta_d + \beta_n + \beta_s = 1$.

8.1.3 Solution

Define x as $x = [1 \quad P_d \quad P_n \quad P_s \quad Y]$. Then the OLS estimate for each regression is:

$$\begin{aligned} \hat{\beta}_d &= (x'x)^{-1}(x'E_d) \\ \hat{\beta}_n &= (x'x)^{-1}(x'E_n) \\ \hat{\beta}_s &= (x'x)^{-1}(x'E_s) \end{aligned}$$

Sum these three together:

$$\hat{\beta}_d + \hat{\beta}_n + \hat{\beta}_s = (x'x)^{-1}(x'Y) \tag{2}$$

Now partition x into $x_1 = \begin{bmatrix} 1 & P_d & P_n & P_s \end{bmatrix}$ and $x_2 = Y$. Then we obtain equation (2) from the following regression:

$$Y = x_1\beta_1 + x_2\beta_2 + \varepsilon$$

Using the FWL theorem:

$$\begin{aligned} \hat{\beta}_2 &= (x_2' M_1 x_2)^{-1} (x_2' M_1 Y) \\ &= (x_2' M_1 x_2)^{-1} (x_2' M_1 x_2) \\ &= 1 \end{aligned}$$

This proves our claim.

8.2 Wald Test

The Wald test allows us to test restrictions on parameters. In class, you saw the linear restriction Wald test. It can be generalized to non-linear restrictions, but let's stick with the basic test for today. We have the following regression:

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + \varepsilon$$

with the following hypotheses:

$$H_0 : \beta_1 = \beta_2 \quad \text{and} \quad \beta_3 = \beta_4$$

The Wald test formula is as follows:

$$W = (R\beta - q)' \left[R \text{Var}(\hat{\beta}) R' \right]^{-1} (R\beta - q)$$

For statistical testing, we need to take the asymptotic distribution. Therefore, the Wald test becomes:

$$\begin{aligned} W &\xrightarrow{d} (R\beta - q)' \left[R \text{Avar}(\hat{\beta}) R' \right]^{-1} (R\beta - q) \\ &= \chi_r^2 \end{aligned}$$

What's the intuition behind the Wald test? We are measuring the distance from our hypothesized values given the coefficients we estimated from the data. Now, let's construct the R matrix:

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

When we multiply R against the β vector:

$$\begin{aligned} R\beta &= \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \\ &= \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_3 - \beta_4 \end{bmatrix} \end{aligned}$$

Now, what is q ? q will be the value we hypothesize for each restriction. Then:

$$q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Now, what about the asymptotic variance term? Assuming homoskedasticity in the error term, we know the asymptotic variance of $\hat{\beta}$:

$$Avar(\hat{\beta}) = \sigma^2 \mathbb{E}[x'x]^{-1}$$

Estimate this using the analogy principle:

$$\widehat{Avar}(\hat{\beta}) = \hat{s}^2 \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1}$$

We have all the pieces for the Wald test now. Just plug them into the formula:

$$W = (R\hat{\beta} - q)' \left[R \widehat{Avar}(\hat{\beta}) R' \right]^{-1} (R\hat{\beta} - q)$$

In our example, we compare the resulting statistic to the critical values of a χ^2 distribution with $r = 2$ degrees of freedom (as we have two restrictions).

8.3 Lagrange Multiplier Test

To solidify intuition, let's look at the set-up of the Lagrangian:

$$\mathcal{L} = S(\beta) + (R(\beta) - q)' \lambda$$

where λ is the Lagrange multiplier and $S(\beta)$ is the sum of squared residuals (for least squares regressions). By solving the first-order conditions for λ and finding its asymptotic distribution, we get a

finite estimate of:

$$LM = \frac{\partial S(\hat{\beta}_r)}{\partial \beta'} \left(\frac{1}{n} \frac{\partial S(\hat{\beta}_r)}{\partial \beta} \frac{\partial S(\hat{\beta}_r)}{\partial \beta'} \right)^{-1} \frac{\partial S(\hat{\beta}_r)}{\partial \beta}$$

$$\xrightarrow{d} \chi_r^2$$

In our case, with the generalized version of least squares ($y = h(x, \beta)$):

$$\frac{\partial S(\hat{\beta}_r)}{\partial \beta} = -2 \sum_{i=1}^n \hat{\varepsilon}_{ri} \frac{\partial h(x_i, \beta)}{\partial \beta}$$

Define x_0 as $\frac{\partial h(x_i, \beta)}{\partial \beta}$. Then we can rewrite the statistic as:

$$LM = \frac{n \hat{\varepsilon}_r' x_r^0 \left((x_r^0)' x_r^0 \right)^{-1} (x_r^0)' \hat{\varepsilon}_r}{\hat{\varepsilon}_r' \hat{\varepsilon}_r}$$

Note the r subscripts and superscripts. All of these objects are evaluated using the restricted regression. In addition, an important assumption working behind the scenes here: that $\hat{\beta}_r$ must be consistent. If it is not consistent, the derivation of the asymptotic distribution for the Lagrange multiplier fails.

8.4 Chow Test

The Chow test test whether coefficients in two regressions, notably on different data sets, are equal. Most often, we use this in time series and on the constant or time trend variable to test for structural breaks.

8.4.1 Homework 5, Question 3

Consider a regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_9 x_{i9} + \varepsilon_i$$

for each year 2004-2010. Regressions for each year and the pooled model produce the following statistics:

	2004	2005	2006	2007	2008	2009	2010	pooled model
Observations	65	55	87	95	103	87	78	570
$e'e$	104	88	206	144	199	308	211	1425

(a) Write the regression model to produce the sum of squared residuals for a pooled regression for 2004-2010.

(b) Test the hypothesis that the slope parameters are equal for all years.

8.4.2 Solution

For part (a):

$$\begin{bmatrix} y_{2004} \\ y_{2005} \\ y_{2006} \\ y_{2007} \\ y_{2008} \\ y_{2009} \\ y_{2010} \end{bmatrix}_{570 \times 1} = \begin{bmatrix} i_{2004} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & x_{2004} \\ 0 & i_{2005} & 0 & 0 & 0 & 0 & 0 & 0 & x_{2005} \\ 0 & 0 & i_{2006} & 0 & 0 & 0 & 0 & 0 & x_{2006} \\ 0 & 0 & 0 & i_{2007} & 0 & 0 & 0 & 0 & x_{2007} \\ 0 & 0 & 0 & 0 & i_{2008} & 0 & 0 & 0 & x_{2008} \\ 0 & 0 & 0 & 0 & 0 & i_{2009} & 0 & 0 & x_{2009} \\ 0 & 0 & 0 & 0 & 0 & 0 & i_{2010} & 0 & x_{2010} \end{bmatrix}_{570 \times 16} \begin{bmatrix} \beta_0^{2004} \\ \beta_0^{2005} \\ \beta_0^{2006} \\ \beta_0^{2007} \\ \beta_0^{2008} \\ \beta_0^{2009} \\ \beta_0^{2010} \\ \beta_{\sim 0} \end{bmatrix}_{16 \times 1}$$

with the error term at the end. For part (b), redefine x_t to include the constant. Then:

$$\begin{bmatrix} y_{2004} \\ y_{2005} \\ y_{2006} \\ y_{2007} \\ y_{2008} \\ y_{2009} \\ y_{2010} \end{bmatrix}_{570 \times 1} = \begin{bmatrix} x_{2004} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & x_{2005} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & x_{2006} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_{2007} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & x_{2008} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & x_{2009} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_{2010} & 0 \end{bmatrix}_{570 \times 70} \begin{bmatrix} \beta_{2004} \\ \beta_{2005} \\ \beta_{2006} \\ \beta_{2007} \\ \beta_{2008} \\ \beta_{2009} \\ \beta_{2010} \end{bmatrix}_{70 \times 1}$$

The Chow test is therefore:

$$F = \frac{(570 - 7 \cdot 10) \left(\hat{\varepsilon}'_p \hat{\varepsilon}_p - \sum_{t=2004}^{2010} \hat{\varepsilon}'_t \hat{\varepsilon}_t \right)}{(9 \cdot 6) \left(\sum_{t=2004}^{2010} \hat{\varepsilon}'_t \hat{\varepsilon}_t \right)}$$

We compare this statistic to the critical value from the F distribution with degrees of freedom equal to the number of restrictions (54) and free observations (500).

8.5 Strict vs. Weak Exogeneity

Let's remind ourselves what both strict and weak exogeneity are:

$$\mathbb{E}[\varepsilon_i | x_i] = 0 \quad (\text{Weak Exogeneity})$$

$$\mathbb{E}[\varepsilon_i | x] = 0 \quad (\text{Strong Exogeneity})$$

We have used strong exogeneity to prove properties of OLS. Could we use weak exogeneity?

The answer is yes, at least for our purposes. Because of our assumption that the observations are independently distributed, ε_i is independent of ε_j and x_i is independent of x_j , for $j \neq i$. Therefore, $\mathbb{E}[\varepsilon_i | x] = \mathbb{E}[\varepsilon_i | x_i]$.

This distinction between exogeneities becomes significant when the independence assumption might fail. Some examples include:

- Geographic correlation
- Time series autocorrelation
- Intra-household data

Chapter 9

Maximum Likelihood Estimation

9.1 Maximum Likelihood Theory

9.1.1 Basics of MLE

The Maximum Likelihood Estimator seeks to estimate a parameter that maximizes the likelihood function. The likelihood function describes the probability of observing the data that we've collected with parameters as the arguments. Of course, the underlying functions and therefore parameters in use are chosen by the modeller. The likelihood function is given as:

$$L_n(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^n f(\mathbf{y}_i|\boldsymbol{\theta})$$

where $f(\mathbf{y}_i|\boldsymbol{\theta})$ is the underlying DGP of the data.

To make analysis easier, we often take the natural log of the likelihood function. Note that because natural log is a monotonically increasing function, the argmax of the log transformation is the same as the argmax of the original likelihood function.

The log-likelihood function is:

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \ln(f(\mathbf{y}_i|\boldsymbol{\theta}))$$

Remember that these functions are functions of $\boldsymbol{\theta}$ and we keep \mathbf{y} fixed.

To estimate $\boldsymbol{\theta}$, we find the score vector. The score vector is defined as:

$$\mathbf{s}_n(\boldsymbol{\theta}) = \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Note that if $\boldsymbol{\theta}$ is $k \times 1$, then $\mathbf{s}_n(\boldsymbol{\theta})$ is also $k \times 1$.

If the log-likelihood function is differentiable over $\boldsymbol{\theta}$, then we can set the score vector equal to zero and solve for $\hat{\boldsymbol{\theta}}$. Recall from your microeconomics class that this FOC is necessary and not sufficient for determining whether we are at a maximum. As such, we should check SOSCs ($\frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} < 0$) just

to be sure.

9.1.2 Kullback-Leibler Information Criterion

The Kullback-Leibler Information Criterion is defined as follows:

$$\begin{aligned} KL(p||q) &= \mathbb{E}_p \left[\ln \left(\frac{p(y)}{q(y)} \right) \right] \\ &= \int_{-\infty}^{\infty} \ln \left(\frac{p(y)}{q(y)} \right) dP(y) \end{aligned}$$

The KLIC measures the ability of the likelihood ratio to distinguish between two distributions.

If you know anything about information theory, then the KLIC probably looks familiar. Entropy is defined as:

$$H(p) = -\mathbb{E}[\ln(p)]$$

and defines the degree of uncertainty in a distribution.¹

An important property of the KLIC, stated in the Gibbs Inequality theorem and derived from Jensen's Inequality, is that for any two distributions p and q , $KL(p||q) \geq 0$, with equality holding only if $p = q$ almost everywhere.²

Proof. Note that \ln is a concave function. Then:

$$\begin{aligned} -\mathbb{E}_p \left[\ln \left(\frac{p}{q} \right) \right] &= \sum_{i=1}^n p_i \ln \left(\frac{q_i}{p_i} \right) \\ &\leq \ln \sum_{i=1}^n p_i \frac{q_i}{p_i} \\ &= \ln \sum_{i=1}^n q_i \\ &\leq 0 \\ \mathbb{E}_p \left[\ln \left(\frac{p}{q} \right) \right] &\geq 0 \\ KL(p||q) &\geq 0 \end{aligned}$$

This proves that the KLIC must be weakly positive. Equality trivially holds when $p = q$. □

¹The two are related as follows: $H(p) = KL(p||u) + C$, where u is the uniform distribution and C is some constant.

²"Almost everywhere" allows $p \neq q$ at a nullity point.

9.1.3 Consistency

Take the estimated log of the likelihood function of some value θ over the likelihood function of the true, maximizing value θ_0 :

$$\frac{1}{n} \ln \left(\frac{L(\theta)}{L(\theta_0)} \right) = \frac{1}{n} \sum_{i=1}^n \ln \left(\frac{L_i(\theta)}{L_i(\theta_0)} \right)$$

By the weak law of large numbers, we know that:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \ln \left(\frac{L_i(\theta)}{L_i(\theta_0)} \right) &\xrightarrow{P} \mathbb{E}_{\theta_0} \left[\ln \left(\frac{L_i(\theta)}{L_i(\theta_0)} \right) \right] \\ &= -KL(L(\theta_0) || L(\theta)) \\ &\leq 0 \end{aligned}$$

By identifiability (that there is only one θ_0), $\theta \neq \theta_0$. Define $\hat{\theta}_n$ as the argmax of the likelihood function. Then:

$$\begin{aligned} P(\hat{\theta}_n \neq \theta_0) &= P \left(\max_{\{\theta \neq \theta_0\}} \frac{1}{n} \sum_{i=1}^n \ln \left(\frac{L(\theta_0)}{L(\theta)} \right) > 0 \right) \\ &\leq \sum_{\theta \neq \theta_0} P \left(\frac{1}{n} \sum_{i=1}^n \ln \left(\frac{L(\theta_0)}{L(\theta)} \right) \right) \\ &\rightarrow 0 \end{aligned}$$

Therefore, since there does not exist a $\hat{\theta}_n$ that maximizes the function greater than θ_0 as $n \rightarrow \infty$, $\hat{\theta}_n \xrightarrow{P} \theta_0$

9.1.4 Fisher Information Matrix

The Fisher information matrix conveys how much information the data \mathbf{Y} carries about the unknown parameter $\boldsymbol{\theta}$. It is given by:

$$\begin{aligned} \mathcal{I}_{\boldsymbol{\theta},n} &= \text{Var} \left(\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \\ &= \mathbb{E} \left[\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right] \end{aligned}$$

Note that if our model is regular (see slide 23 of Topic 7) and correctly specified (see slide 6 of Topic 7), then the Fisher information matrix equals the expected Hessian:

$$\begin{aligned} \mathcal{I}_{\boldsymbol{\theta},n} &= \mathcal{H}_{\boldsymbol{\theta},n} \\ \mathbb{E} \left[\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right] &= \mathbb{E} \left[-\frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] \end{aligned}$$

It is useful to remember that the asymptotic distribution of our MLE estimators is:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N\left(0, \mathcal{I}_{\boldsymbol{\theta},1}^{-1}\right)$$

as long as assumptions 1-11 on slide 54 of Topic 7 hold.

If the model is mis-specified, then the asymptotic distribution is:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N\left(0, \mathcal{H}_{\boldsymbol{\theta},1}^{-1} \mathcal{I}_{\boldsymbol{\theta},1} \mathcal{H}_{\boldsymbol{\theta},1}^{-1}\right)$$

9.1.5 Cramér-Rao Lower Bound

If a model is regular, parametric, and correctly specified with an interior solution for $\boldsymbol{\theta}$ that is unbiased, then the lowest possible variance is the Cramér-Rao Lower Bound. That is:

$$\text{Var}(\hat{\boldsymbol{\theta}}_n) \geq (n\mathcal{I}_{\boldsymbol{\theta},1})^{-1}$$

Note that if our data is *i.i.d.*, then $n\mathcal{I}_{\boldsymbol{\theta},1} = \mathcal{I}_{\boldsymbol{\theta},n}$.

9.2 Practice Problem 1: Hansen 10.4

Let X be distributed Cauchy with density $f(x) = \frac{1}{\pi(1+(x-\theta)^2)}$ for $x \in \mathbb{R}$.

- (a) Find the log-likelihood function of $\ell_n(\theta)$.
- (b) Find the first-order condition for the MLE $\hat{\theta}$ for θ . You will not be able to solve for $\hat{\theta}$.

9.2.1 Part a

We begin by finding the likelihood function. We will then take the natural log of that function. Using the definition of the likelihood function:

$$\begin{aligned} \mathcal{L}(\theta|x) &= \prod_{i=1}^n \frac{1}{\pi(1+(x_i-\theta)^2)} \\ \ell_n(\theta) &= \sum_{i=1}^n -\ln(\pi(1+(x_i-\theta)^2)) \\ &= \sum_{i=1}^n -\ln(\pi) - \ln(1+(x_i-\theta)^2) \\ &= -n\ln(\pi) - \sum_{i=1}^n \ln(1+(x_i-\theta)^2) \end{aligned}$$

9.2.2 Part b

Now that we have the log-likelihood function, we can find the score vector and set that equal to zero:

$$\begin{aligned}\frac{\partial \ell_n(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} \left[-n \ln(\pi) - \sum_{i=1}^n \ln(1 + (x_i - \theta)^2) \right] \\ &= \sum_{i=1}^n \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2} \\ &= 0\end{aligned}$$

Usually we'd like to solve for $\hat{\theta}$, but note that here we cannot. We need a numerical solver to find the optimal estimator. So our work, analytically at least, is done.

9.3 Practice Problem 2: Hansen 10.7

Take the Pareto model $f(x) = \alpha x^{-1-\alpha}$, $x \geq 1$. Calculate the information for α using the second derivative.

9.3.1 Solution

We start with the likelihood function. We will then take the natural log to find the log-likelihood function.

$$\begin{aligned}\mathcal{L}(\alpha|x) &= \prod_{i=1}^n (\alpha x_i^{-1-\alpha}) \\ \ell_n(\alpha) &= \sum_{i=1}^n \ln(\alpha x_i^{-1-\alpha}) \\ &= \sum_{i=1}^n \ln(\alpha) - (1 + \alpha) \ln(x_i) \\ &= n \ln(\alpha) - \sum_{i=1}^n (1 + \alpha) \ln(x_i)\end{aligned}$$

Now that we have the log-likelihood we can find the score vector:

$$\begin{aligned}\frac{\partial \ell_n(\alpha)}{\partial \alpha} &= \frac{\partial}{\partial \alpha} \left[n \ln(\alpha) - \sum_{i=1}^n (1 + \alpha) \ln(x_i) \right] \\ &= \frac{n}{\alpha} - \sum_{i=1}^n \ln(x_i)\end{aligned}$$

The problem tells us to use the second derivative, so taking the derivative with respect to α again:

$$\begin{aligned}\frac{\partial^2 \ell_n(\alpha)}{\partial \alpha^2} &= \frac{\partial}{\partial \alpha} \left[\frac{n}{\alpha} - \sum_{i=1}^n \ln(x_i) \right] \\ &= -\frac{n}{\alpha^2}\end{aligned}$$

Now that we have the second derivative, we look at the formula for the expected Hessian:

$$\begin{aligned}\mathcal{H}_{\theta,n} &= \mathbb{E} \left[-\frac{\partial^2 \ell_n(\alpha)}{\partial \alpha^2} \right] \\ &= \mathbb{E} \left[\frac{n}{\alpha^2} \right] \\ &= \frac{n}{\alpha^2}\end{aligned}$$

Assuming the information matrix equality holds, then:

$$\mathcal{I}_{\theta,n} = \frac{n}{\alpha^2}$$

9.4 Practice Problem 3

Suppose we have random variable $y_i \sim N(\alpha + \beta x_i, \sigma^2)$. The probability density function, conditional on x_i , is thus:

$$f(y_i|x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2}$$

In this problem, we want to:

- Find the log-likelihood function.
- Define the score vector and solve the first order conditions.
- Derive the observed Hessian.
- Derive the Fisher information matrix.

9.4.1 Conditional Log-Likelihood

Start first with the likelihood, then take the log:

$$\begin{aligned}L_n(\boldsymbol{\theta}) &= \prod_{i=1}^n f(y_i|x_i) \\ \ell_n(\boldsymbol{\theta}) &= \sum_{i=1}^n \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2} \right)\end{aligned}$$

$$= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

9.4.2 Score Vector

Next we take the derivative of the log-likelihood with respect to each of our three parameters:

$$\begin{aligned} \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \alpha} &= \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i) \\ \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \beta} &= \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i) x_i \\ \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \end{aligned}$$

Let's simplify one equation at a time, starting with α :

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i) &= \frac{1}{\sigma^2} \left[\sum_{i=1}^n y_i - n\alpha - \beta \sum_{i=1}^n x_i \right] \\ &= \frac{1}{\sigma^2} [n\bar{y} - n\alpha - \beta n\bar{x}] \end{aligned}$$

Recall that we set these first-order conditions equal to zero, so we can multiply by σ^2 and add $n\alpha$ to get:

$$\begin{aligned} n\alpha &= n\bar{y} - n\beta\bar{x} \\ \hat{\alpha} &= \bar{y} - \beta\bar{x} \end{aligned}$$

Before we can advance further, we must solve for $\hat{\beta}$:

$$\begin{aligned} 0 &= \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i) x_i \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n x_i y_i - \alpha x_i - \beta x_i^2 \\ &= \frac{1}{\sigma^2} [n\bar{x}\bar{y} - n\alpha\bar{x} - n\beta\bar{x}\bar{x}] \\ \hat{\beta} &= \frac{\bar{x}\bar{y} - \alpha\bar{x}}{\bar{x}\bar{x}} \end{aligned}$$

Now we can plug $\hat{\beta}$ into $\hat{\alpha}$:

$$\begin{aligned}\hat{\alpha} &= \bar{y} - \left(\frac{\overline{xy} - \alpha \bar{x}}{\overline{xx}} \right) \bar{x} \\ \hat{\alpha} \left(1 - \frac{\bar{x}^2}{\overline{xx}} \right) &= \bar{y} - \frac{\overline{xx}\bar{y}}{\overline{xx}} \\ \hat{\alpha} (\overline{xx} - \bar{x}^2) &= \overline{xx}\bar{y} - \overline{xy}\bar{x} \\ \hat{\alpha} &= \frac{\overline{xx}\bar{y} - \overline{xy}\bar{x}}{\overline{xx} - \bar{x}^2}\end{aligned}$$

We can then plug $\hat{\alpha}$ back into $\hat{\beta}$:

$$\begin{aligned}\hat{\beta} &= \frac{\overline{xy} - \frac{\overline{xx}\bar{y} - \overline{xy}\bar{x}}{\overline{xx} - \bar{x}^2} \bar{x}}{\overline{xx}} \\ &= \frac{\frac{\overline{xy}(\overline{xx} - \bar{x}^2) - \overline{xx}\bar{y}\bar{x} + \overline{xy}\bar{x}^2}{\overline{xx} - \bar{x}^2}}{\overline{xx}} \\ &= \frac{\overline{xyxx} - \overline{xx}\bar{y}\bar{x} - \bar{x}^2\overline{xy} + \bar{x}^2\overline{xy}}{\overline{xx}(\overline{xx} - \bar{x}^2)} \\ &= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{xx} - \bar{x}^2}\end{aligned}$$

Now we can solve for σ^2 :

$$\begin{aligned}0 &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \\ \frac{n}{2\sigma^2} &= \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2\end{aligned}$$

9.4.3 Deriving the Observed Hessian

The observed Hessian $\mathcal{H}_{\theta,n}^o$ is defined as:

$$\mathcal{H}_{\theta,n}^o = \left[\frac{\partial^2 \ell_n(\theta)}{\partial \theta \partial \theta'} \right]$$

Let's go one at a time:

$$\begin{aligned}\frac{\partial^2 \ell_n(\theta)}{\partial \alpha^2} &= -\frac{n}{\sigma^2} \\ \frac{\partial^2 \ell_n(\theta)}{\partial \alpha \partial \beta} &= -\frac{1}{\sigma^2} \sum_{i=1}^n x_i\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \alpha \partial \sigma^2} &= -\frac{1}{\sigma^4} \sum_{i=1}^n (y_i - \alpha - \beta x_i) \\ \frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \beta^2} &= -\frac{1}{\sigma^2} \sum_{i=1}^n x_i^2 \\ \frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \beta \partial \sigma^2} &= -\frac{1}{\sigma^4} \sum_{i=1}^n (y_i - \alpha - \beta x_i) x_i \\ \frac{\partial^2 \ell_n(\boldsymbol{\theta})}{(\partial \sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\end{aligned}$$

9.4.4 Information Matrix

If we assume the information matrix equality holds, then the expectation of the negative of the observed Hessian is the same as the information matrix. Recall that we treat x_i as fixed data so that its mean is its expectation. Let's first take the negative expectation of the second derivatives we calculated:

$$\begin{aligned}\mathbb{E} \left[-\frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \alpha^2} \right] &= \frac{n}{\sigma^2} \\ \mathbb{E} \left[-\frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \alpha \partial \beta} \right] &= \frac{n}{\sigma^2} \bar{x} \\ \mathbb{E} \left[-\frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \alpha \partial \sigma^2} \right] &= 0 \\ \mathbb{E} \left[-\frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \beta^2} \right] &= \frac{n}{\sigma^2} \overline{xx} \\ \mathbb{E} \left[-\frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \beta \partial \sigma^2} \right] &= 0 \\ \mathbb{E} \left[-\frac{\partial^2 \ell_n(\boldsymbol{\theta})}{(\partial \sigma^2)^2} \right] &= \frac{n}{2\sigma^4}\end{aligned}$$

Note that we use both the property that the sum of the residuals is zero and that the sum of squared residuals is equal to the variance. So the information matrix is:

$$\mathcal{I}_{\boldsymbol{\theta},n} = \begin{bmatrix} \frac{n}{\sigma^2} & \frac{n}{\sigma^2} \bar{x} & 0 \\ \frac{n}{\sigma^2} \bar{x} & \frac{n}{\sigma^2} \overline{xx} & 0 \\ 0 & 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

So that was a lot of work! Know how to calculate information matrices quickly for the comprehensive exam in June.

9.5 Practice Problem 4: Hansen 13.3 Extended

Consider independent and identically distributed observations y_1, \dots, y_n from an exponential distribution with parameter λ . We define the distribution such that $\mathbb{E}[y] = \lambda$. We want a test for $H_0 : \lambda = 1$ against $H_1 : \lambda \neq 1$. For reference, an exponential distribution has pdf: $f(y) = \frac{1}{\lambda} e^{-\frac{1}{\lambda}y}$

- (a) Develop an asymptotic t -test based on the sample mean.
- (b) Derive the likelihood ratio statistic.
- (c) Derive the score test.
- (d) Derive the Wald test.

9.5.1 Part a

The asymptotic t -test takes the following form:

$$T = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sigma}$$

where σ is the asymptotic standard error. We know the true distribution of y , so we can use the variance of an exponential distribution:

$$\begin{aligned} VAR(\sqrt{n}\bar{y}) &= nVar(\bar{y}) \\ &= \frac{1}{n} \sum_{i=1}^n Var(y) \\ &= \lambda^2 \end{aligned}$$

Our asymptotic t -test will look as follows:

$$\begin{aligned} T &= \frac{\sqrt{n}(\hat{\lambda} - \lambda_0)}{\lambda} \\ &= \frac{\sqrt{n}(\mathbb{E}[y] - \lambda_0)}{\mathbb{E}[y]} \\ &= \frac{\sqrt{n}(\bar{y} - 1)}{\bar{y}} \\ &= \frac{\sqrt{n}(\bar{y} - 1)}{\bar{y}} \end{aligned}$$

9.5.2 Part b

The likelihood ratio test looks as follows:

$$LR = 2 \left[\ln \left(L(\hat{\lambda}|y) \right) - \ln \left(L(\lambda_0|y) \right) \right]$$

In our problem, our restricted estimator is $\hat{\lambda}_r = \lambda_0 = 1$. Plugging in the log-likelihood gives:

$$\begin{aligned} &= 2 \left[\sum_{i=1}^n -\ln(\hat{\lambda}) - \frac{y_i}{\hat{\lambda}} - \left(\sum_{i=1}^n -\ln(\lambda_0) - \frac{y_i}{\lambda_0} \right) \right] \\ &= -2 \left[-n \ln(\hat{\lambda}) - \frac{n}{\hat{\lambda}} \bar{y} + n\bar{y} \right] \\ &\xrightarrow{d} \chi_1^2 \end{aligned}$$

9.5.3 Part c

The score test looks as follows:

$$T_s(y) = \frac{1}{n} \frac{\partial \ell_n(\hat{\theta}_r)}{\partial \theta} \mathcal{I}_{\hat{\theta}_r, 1}^{-1} \frac{1}{n} \frac{\partial \ell_n(\hat{\theta}_r)}{\partial \theta}$$

First, we need to find the score:

$$\begin{aligned} \left. \frac{\partial \ell_n(\lambda)}{\partial \lambda} \right|_{\lambda=\lambda_0} &= \left. \sum_{i=1}^n -\frac{1}{\lambda} + \frac{y_i}{\lambda^2} \right|_{\lambda=\lambda_0} \\ &= -\frac{n}{\lambda_0} + \frac{n}{\lambda_0^2} \bar{y} \\ &= -n + n\bar{y} \end{aligned}$$

Now we calculate the information matrix, using the information matrix equality:

$$\begin{aligned} \mathcal{I}_{\lambda_0, 1} &= \mathbb{E} \left[\frac{\partial^2 \ell_n(\lambda)}{\partial \lambda \partial \lambda'} \right]_{\lambda=\lambda_0} \\ &= -\mathbb{E} \left[\frac{1}{\lambda^2} - 2 \frac{y_i}{\lambda^3} \right]_{\lambda=\lambda_0} \\ &= -\left[\frac{1}{\lambda^2} - \frac{2}{\lambda^2} \right]_{\lambda=\lambda_0} \\ &= \frac{1}{\lambda^2} \Big|_{\lambda=\lambda_0} \\ &= \frac{1}{\lambda_0^2} \\ &= 1 \end{aligned}$$

So the score statistic is:

$$\begin{aligned} T_s &= \frac{1}{n} (-n + n\bar{y})(1)^{-1} (-n + n\bar{y}) \\ &= n(\bar{y} - 1)^2 \\ &\xrightarrow{d} \chi_1^2 \end{aligned}$$

9.5.4 Part d

The Wald test takes the general form:

$$T_w(y) = n \left(R(\hat{\theta}) - q \right)' \left(\frac{\partial R(\hat{\theta})}{\partial \theta'} \mathcal{I}_{\hat{\theta},1}^{-1} \frac{\partial R(\hat{\theta})}{\partial \theta} \right)^{-1} (R(\hat{\theta}) - q)$$

In our case, $R(\hat{\theta}) = \hat{\lambda}$ and $q = 1$. Therefore, $\frac{\partial R(\hat{\theta})}{\partial \theta} = 1$. Plugging in everything yields:

$$\begin{aligned} T_w &= n \left(\hat{\lambda} - 1 \right) \left(\frac{1}{\hat{\lambda}^2} \right) \left(\hat{\lambda} - 1 \right) \\ &= \frac{n \left(\hat{\lambda} - 1 \right)^2}{\hat{\lambda}^2} \\ &= \frac{n(\bar{y} - 1)^2}{\bar{y}^2} \end{aligned}$$

This is the square of the t -test.

9.6 Practice Problem 5: Hansen 13.1 Extended

Take the Bernoulli model with probability parameter p . We want a test for $H_0 : p = 0.05$ against $H_1 : p \neq 0.05$.

- Develop a test based on the sample mean \bar{x}_n .
- Derive the likelihood ratio statistic. What is its asymptotic sampling distribution?
- Derive the score test. What is its asymptotic sampling distribution?
- Derive the Wald test. What is its asymptotic sampling distribution?

9.6.1 Part a

Because we know that the mean of a Bernoulli random variable is p , our estimator \hat{p} is simply \bar{x}_n . We use a t -test for (a):

$$\begin{aligned} T &= \frac{\sqrt{n}(\bar{x}_n - p_0)}{\hat{s}} \\ &= \frac{\sqrt{n}(\bar{x}_n - 0.05)}{\hat{s}} \end{aligned}$$

Because we know the distribution of our random variable, we replace \hat{s} with the estimated true variance. Bernoulli distributions have variance $p(1 - p)$, so our t -test becomes:

$$\begin{aligned}
T &= \frac{\sqrt{n}(\bar{x}_n - 0.05)}{\hat{s}} \\
&= \frac{\sqrt{n}(\bar{x}_n - 0.05)}{\sqrt{\hat{p}(1 - \hat{p})}} \\
&= \frac{\sqrt{n}(\bar{x}_n - 0.05)}{\sqrt{\bar{x}_n(1 - \bar{x}_n)}}
\end{aligned}$$

9.6.2 Part b

Our $\hat{p}_0 = 0.05$, so using the formula for the likelihood ratio statistic:

$$\begin{aligned}
lr_n &= 2 \left[\sum_{i=1}^n \ln(\hat{p}^{y_i}(1 - \hat{p})^{1 - y_i}) - \sum_{i=1}^n \ln(\hat{p}_0^{y_i}(1 - \hat{p}_0)^{1 - y_i}) \right] \\
&= 2 \left[\sum_{i=1}^n y_i \ln(\hat{p}) + (1 - y_i) \ln(1 - \hat{p}) - \sum_{i=1}^n y_i \ln(0.05) + (1 - y_i) \ln(1 - 0.05) \right] \\
&\xrightarrow{d} \chi_1^2
\end{aligned}$$

Where the one degree of freedom comes from the one restriction we impose (the null hypothesis).

9.6.3 Part c

To derive the score test, we first need to find the score vector evaluated at the restricted value:

$$\begin{aligned}
\left[\frac{\partial \ell_n(p)}{\partial p} \right]_{p=p_0} &= \left[\sum_{i=1}^n \frac{y_i}{p} - \sum_{i=1}^n \frac{1 - y_i}{1 - p} \right]_{p=p_0} \\
&= \sum_{i=1}^n \frac{y_i}{0.05} - \sum_{i=1}^n \frac{1 - y_i}{1 - 0.05} \\
&= \frac{n\bar{y}_n}{0.05} - \frac{n - n\bar{y}_n}{1 - 0.05} \\
&= n\bar{y}_n \left[20 + \frac{20}{19} \right] - \frac{20n}{19} \\
&= n\bar{y}_n \left[\frac{400}{19} \right] - \frac{20n}{19}
\end{aligned}$$

We now need the information matrix evaluated under the null hypothesis. Knowing that this parametric model is correctly specified (due to us knowing the underlying distribution) and regular, we can use the information matrix equality:

$$\begin{aligned}
\mathcal{I}_{\hat{p}_0,1} &= \mathbb{E} \left[-\frac{\partial \ell_n(p)}{\partial p} \right]_{p=p_0} \\
&= \mathbb{E} \left[\frac{\bar{y}}{p^2} - \frac{\bar{y}}{(1-p)^2} + \frac{1}{(1-p)^2} \right]_{p=p_0} \\
&= \left[\frac{p}{p^2} - \frac{p}{(1-p)^2} + \frac{1}{(1-p)^2} \right]_{p=p_0} \\
&= \left[\frac{1}{p} + \frac{1}{1-p} \right]_{p=p_0} \\
&= \left[\frac{1-p+p}{p(1-p)} \right]_{p=p_0} \\
&= \left[\frac{1}{p(1-p)} \right]_{p=p_0} \\
&= \frac{1}{0.05(1-0.05)} \\
&= \frac{400}{19}
\end{aligned}$$

We now have all of the pieces that we need. Plugging these into the score statistic formula:

$$\begin{aligned}
T_s &= \frac{1}{n} \left(n\bar{y} \left[\frac{400}{19} \right] - \frac{20n}{19} \right) \frac{19}{400} \left(n\bar{y} \left[\frac{400}{19} \right] - \frac{20n}{19} \right) \\
&\xrightarrow{d} \chi_1^2
\end{aligned}$$

9.6.4 Part d

We first need to set up our \mathbf{g} . As I noted last week, the easiest way to do this is linearly:

$$\begin{aligned}
\mathbf{g}(p) &= p - p_0 \\
&= p - 0.05 \\
&= 0
\end{aligned}$$

Next we need to take the derivative of $\mathbf{g}(p)$ with respect to p :

$$\frac{\partial \mathbf{g}(p)}{\partial p} = 1$$

Taking $\mathcal{I}_{\hat{p}_0,1}$ we estimated above, simply replace p with \hat{p} and we have the information matrix we need. Plugging everything into the Wald statistic formula:

$$\begin{aligned}
 T_w &= n(\hat{p} - 0.05)(1 \cdot \hat{p}(1 - \hat{p}) \cdot 1)^{-1}(\hat{p} - 0.05) \\
 &= \frac{n(\hat{p} - 0.05)^2}{\hat{p}(1 - \hat{p})} \\
 &\xrightarrow{d} \chi_1^2
 \end{aligned}$$

Notice that because $\hat{p} = \bar{x}_n$ and because $g(p)$ is linear, our Wald statistic is simply the square of our t-statistic.

9.7 Graphing MLE

```

%% Plotting the Likelihood

x = [5 0 1 1 0 3 2 3 4 1]';

% Likelihood Function
L = @(theta) prod((theta.*exp(-theta.*x)));
theta1=0.01:0.01:3.5;
Lf = L(theta1);

% Log-likelihood function
lnL = @(theta) log(prod((theta.*exp(-theta.*x))));
logf = lnL(theta1);
[~,index] = max(logf);
thetahat = theta1(index);

% Plot likelihood and log-likelihood
c=c+1;
figure(c)
yyaxis left
plot(theta1,Lf, 'LineWidth', 2)
xline(thetahat, 'LineWidth', 2, 'Color', "#EDB120")
xlabel('Values for $\theta$')
ylabel('Likelihood')
title('Plotting MLE')
hold on

yyaxis right
plot(theta1, logf, 'LineWidth', 2)
ylabel('Log-Likelihood')
hold off

```

The figure above gives an example of how to graph the likelihood and log-likelihood functions on the same plot. If you have any questions, feel free to email me.

The figure below gives an example of how to use Matlab's built-in numerical solver. The problem set asks you to use the Newton-Raphson algorithm, but you can use this code to check your work.

```
%% Numerically find the minimum

% Likelihood
initguess = 1; % Initial value
L1 = @(theta) -prod((theta.*exp(-theta.*x))); % Multiply likelihood by negative 1
estimate = fminsearch(L1,initguess); % Estimate using fminsearch
fprintf("\nThe MLE using the likelihood " + ...
        "function is %f\n", estimate);

% Log-likelihood
initguess = 1; % Initial value
lnL1 = @(theta) -log(prod((theta.*exp(-theta.*x)))); % Multiply log likelihood by negative 1
estimate = fminsearch(lnL1,initguess); % Estimate using fminsearch
fprintf("\nThe MLE using the log-likelihood " + ...
        "function is %f\n", estimate);
```

Chapter 10

Generalized Least Squares

10.1 GLS Theory

10.1.1 Purpose

Throughout this theory section, keep the classic regression equation in mind:

$$y = x\beta + \varepsilon$$

Under OLS, we made three main assumptions:

1. x is full rank
2. $\mathbb{E}[\varepsilon|x] = 0$
3. $Var(\varepsilon|x) = \sigma^2 I$

The last assumption is called homoskedasticity - the variance of the errors does not depend on the value of our right-hand side variables. Under GLS, we relax the assumption of homoskedasticity:

$$Var(\varepsilon_i|x_i) = \sigma_i^2$$

This type of variance is called heteroskedasticity. One way to correct for heteroskedasticity is to do GLS.

10.1.2 Derivation

Suppose the variance of the error term takes the following form:

$$\begin{aligned}
 \text{Var}(\varepsilon|x) &= \sigma^2 \Omega \\
 &= \sigma^2 \begin{bmatrix} \omega_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \omega_n \end{bmatrix} \\
 &= \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n^2 \end{bmatrix} \\
 &\equiv V
 \end{aligned}$$

Because this is a variance-covariance matrix, it must be positive definite. As such, it has an inverse, and that inverse has a Cholesky decomposition:

$$V^{-1} = P'P$$

Where P is an upper triangular matrix with real and positive diagonal entries.

Our goal in GLS is to “fix” the heteroskedasticity problem by reshaping the error’s variance to be homoskedastic. How? Premultiply the ε by P and take the variance:

$$\begin{aligned}
 \text{Var}(P\varepsilon) &= P\text{Var}(\varepsilon)P' \\
 &= P\sigma^2\Omega P' \\
 &= \sigma^2 P(P'P)^{-1}P' \\
 &= \sigma^2 PP^{-1}P'(P')^{-1} \\
 &= \sigma^2 I
 \end{aligned}$$

Therefore, under the transformation (premultiplication by P), the system has homoskedastic errors. Apply this transformation to the regression equation:

$$\begin{aligned}
 y &= x\beta + \varepsilon \\
 Py &= Px\beta + P\varepsilon
 \end{aligned}$$

Now solve for β :

$$\begin{aligned}
 x'P'Py &= x'P'Px\beta + x'P'P\varepsilon \\
 (x'P'Px)^{-1}(x'P'Py) &= \beta + (x'P'Px)^{-1}(x'P'P\varepsilon)
 \end{aligned}$$

$$\begin{aligned}(x'P'Px)^{-1}(x'P'Py) &= \hat{\beta}^{GLS} \\ (x'\Omega^{-1}x)^{-1}(x'\Omega^{-1}y) &= \hat{\beta}^{GLS}\end{aligned}$$

10.1.3 Bias and Variance

To find the bias, we take the expected value of the estimator:

$$\begin{aligned}\mathbb{E} \left[\hat{\beta}^{GLS} \mid x \right] &= \mathbb{E} \left[(x'\Omega^{-1}x)^{-1}(x'\Omega^{-1}(x\beta + \varepsilon)) \mid x \right] \\ &= \beta + (x'\Omega^{-1}x)^{-1}(x'\Omega^{-1}\mathbb{E}[\varepsilon \mid x]) \\ &= \beta\end{aligned}$$

So GLS is unbiased under the assumption of strong exogeneity. For the variance, relabel the GLS estimator as:

$$\begin{aligned}\hat{\beta}^{GLS} &= (x'\Omega^{-1}x)^{-1}(x'\Omega^{-1}y) \\ &= (x'P'Px)^{-1}(x'P'Py) \\ &= [(x^*)'x^*]^{-1}[(x^*)'y^*]\end{aligned}$$

We know that under homoskedasticity, the variance of the estimator is:

$$Var \left(\hat{\beta}^{OLS} \mid x \right) = \sigma^2(x'x)^{-1}$$

Applying this same logic to our transformed system gives the GLS variance:

$$\begin{aligned}Var \left(\hat{\beta}^{GLS} \mid x \right) &= \sigma^2 [(x^*)'x^*]^{-1} \\ &= \sigma^2 (x'\Omega^{-1}x)\end{aligned}$$

We have shown that the GLS estimator is unbiased. We can also note that the transformed model is a linear OLS model, so it must also have the least variance among unbiased estimators. Therefore, GLS is BLUE.

10.1.4 Feasible GLS

One problem with GLS is that we don't actually observe $\sigma^2\Omega$. As such, we need to estimate them. How do we do this? In two steps using weighted least squares.

First, run OLS on the model. Then get the error terms. Then estimate $\hat{\sigma}^2$ as follows:

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$$

We can then develop some function of these variances to form our weights (like $\hat{\omega}_i = \hat{\sigma}^2(x_i)$). We will

then compose our estimated Ω as:

$$\hat{\Omega} = \begin{bmatrix} x_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & x_n \end{bmatrix}$$

Second, calculate the GLS estimator:

$$\hat{\beta}^{GLS} = \left(\frac{1}{n} \sum_{i=1}^n x_i (\hat{\omega}_i)^{-1} x_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i (\hat{\omega}_i)^{-1} y_i \right)$$

Once we have the GLS estimator, we can iterate through these two steps (using the GLS estimator is step two to update $\hat{\sigma}^2$).

10.1.5 Consistency

Starting from the FGLS estimator:

$$\begin{aligned} \hat{\beta}^{GLS} &= \left(\frac{1}{n} \sum_{i=1}^n x_i' (\hat{\omega}_i)^{-1} x_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i' (\hat{\omega}_i)^{-1} y_i \right) \\ &= \beta + \left(\frac{1}{n} \sum_{i=1}^n x_i' (\hat{\omega}_i)^{-1} x_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i' (\hat{\omega}_i)^{-1} \varepsilon_i \right) \quad (*) \\ &\xrightarrow{P} \beta + \mathbb{E} [x_i' (\omega_i)^{-1} x_i]^{-1} \mathbb{E} [x_i' (\omega_i)^{-1} \varepsilon_i] \\ &= \beta \end{aligned}$$

This proof holds as long as $\hat{\omega}_i$ is consistent for ω_i .

10.1.6 Asymptotic Normality

Starting from (*):

$$\begin{aligned} \hat{\beta} &= \beta + \left(\frac{1}{n} \sum_{i=1}^n x_i' (\hat{\omega}_i)^{-1} x_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i' (\hat{\omega}_i)^{-1} \varepsilon_i \right) \\ \sqrt{n} (\hat{\beta} - \beta) &= \left(\frac{1}{n} \sum_{i=1}^n x_i' (\hat{\omega}_i)^{-1} x_i \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i' (\hat{\omega}_i)^{-1} \varepsilon_i \right) \end{aligned}$$

Brushing some technicalities on the estimation of the weights aside, we can send this to infinity:

$$\sqrt{n} (\hat{\beta} - \beta) \xrightarrow{d} N(0, (x' \Omega^{-1} x)^{-1} x' \Omega^{-1} \varepsilon \varepsilon' (\Omega^{-1})' x (x' \Omega^{-1} x)^{-1})$$

10.2 Heteroskedasticity

10.2.1 White's Standard Errors

We can correct for heteroskedasticity in OLS as well. Consider the OLS estimator:

$$\begin{aligned}\hat{\beta} &= (x'x)^{-1}(x'y) \\ &= \beta + (x'x)^{-1}(x'\varepsilon) \\ \sqrt{n}(\hat{\beta} - \beta) &= \left(\frac{1}{n} \sum_{i=1}^n x_i x_i'\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i \varepsilon_i\right) \\ &\xrightarrow{d} N\left(0, (x'x)^{-1} x' \varepsilon \varepsilon x (x'x)^{-1}\right)\end{aligned}$$

Because of heteroskedasticity, we cannot simplify $\varepsilon \varepsilon'$ to $\sigma^2 I$ as before. The simplification we usually make to $\sigma^2 (x'x)^{-1}$ cannot occur. The standard errors that result from this unsimplified variance are called **White's standard errors**.

To estimate the asymptotic variance:

$$\widehat{AVAR} = (x'x)^{-1} \left(\frac{n}{n-k} \sum_{i=1}^n x_i \hat{\varepsilon}_i' \hat{\varepsilon}_i x_i' \right) (x'x)^{-1}$$

10.2.2 White's Heteroskedasticity Test

Without going too much into the theory, White's test for heteroskedasticity is calculated by finding the R^2 value of a regression of the squared OLS residuals, $\hat{\varepsilon}^2$, on x and the cross-products of every variable in x . Then, using the Lagrange Multiplier test:

$$LM = nR^2$$

For example, suppose $x = [1 \text{ age } \text{age}^2]$. Then the steps are:

1. Run $y = x\beta + \varepsilon$.
2. Find $\hat{\varepsilon}$.
3. Build $x^* = [x \text{ age}^3 \text{ age}^4]$.
4. Run $\hat{\varepsilon}^2 = x^* \beta + \varepsilon^*$.
5. Calculate the R^2 from this second regression.
6. Calculate $LM = nR^2$.
7. Compare to the asymptotic critical value from χ_{k-1}^2 , where k is the number of regressors in the second regression.

10.2.3 Breusch-Pagan Test

The first few steps are similar to White's test. First, run the OLS model. Then obtain the estimated residuals. In the Breusch-Pagan test, we then regress the square of the estimated residuals on x again. Lastly, calculate the Lagrange Multiplier statistic.

Let's use the example from above again. Suppose $x = [1 \text{ age } \text{age}^2]$. Then:

1. Run $y = x\beta + \varepsilon$.
2. Find $\hat{\varepsilon}$.
3. Run $\hat{\varepsilon}^2 = x\beta + \varepsilon^*$.
4. Calculate $LM = \frac{1}{2}(TSS - SSR)$, using values from the second regression.
5. Compare the LM statistic to the critical value from χ_{k-1}^2 , where k is the number of regressors in the second equation.

10.3 Example

Consider the following population model: $y_i \sim N(2\pi x_i, 100) + \varepsilon_i$, where $x_i \stackrel{iid}{\sim} Unif(0, 1)$ and $\varepsilon_i | x_i \stackrel{iid}{\sim} N(0, x_i^2)$.

- (a) Simulate 10,000 samples with 10,000 observations each. For each sample, take the first 500, 1000, and 10,000 observations. Store the slope coefficients from OLS.
- (b) Take the mean and variance of each slope coefficient.
- (c) Now conduct FGLS by assigning $\omega_i = x_i$. Repeat parts (a) and (b).

```

mean ols by sample size
      6.5085      6.3541      6.3442

variance ols by sample size
    244.7322    121.4102    12.2218

mean gls by sample size
      2.9177      6.0302      6.1091

variance gls by sample size
    1.0e+05 *

      2.2510      2.6458      4.5025

```

We can see from the figure above that OLS converges to an estimate around 2π quickly. This result makes sense, as if you take the probability limit of the OLS estimator, you will get 2π analytically.

FGLS, though, takes much longer to converge to 2π . Why might this be? Unlike GLS, we cannot guarantee that FGLS will be unbiased. As such, it may take a much larger sample size for FGLS to converge to zero. In this exercise, the variance of the FGLS estimator is also much larger.

Chapter 11

Generalized Method of Moments

11.1 Previous Problem: PS 7, Question 2

Consider the following joint probability function of (x, y) :

$$f(x, y|\beta) = \frac{1}{\beta + x} e^{-\frac{y}{\beta+x}}$$

- (a) Write the log-likelihood function for β .
- (b) Using “DataHw7.2.xlsx”, obtain the MLE for β using the Newton-Raphson algorithm.
- (c) Consider joint probability function:

$$f(x, y|\beta, \rho) = \frac{(\beta + x)^{-\rho}}{\Gamma(\rho)} y^{\rho-1} e^{-\frac{y}{\beta+x}}$$

Use MLE to find $\hat{\beta}$ and $\hat{\rho}$.

- (d) Test $H_0 : \rho = 1$ using an asymptotic t-test, the log-likelihood ratio test, the Wald test, and the score test.

11.1.1 Part a

To obtain the log-likelihood function, multiply the joint pdf together n times:

$$L(\beta) = \prod_{i=1}^n \frac{1}{\beta + x_i} e^{-\frac{y_i}{\beta+x_i}}$$

Then take the natural log:

$$\ell_n(\beta) = \sum_{i=1}^n \left(-\ln(\beta + x_i) - \frac{y_i}{\beta + x_i} \right)$$

11.1.2 Part b

To solve this in Matlab, we need to define the log-likelihood. I do so as a function handle:

```
%% Define the two log-likelihood functions

loga = @(beta) sum(-log(beta + x) - y./(beta+x));
```

I then pass the log-likelihood to the Newton-Raphson algorithm:

```
%% Find beta for Part a

b0 = 1;
k = 1;
[betaa, stda, vara] = nr(loga, n, k, b0);
```

where the variance is calculated using the negative expected Hessian. The algorithm returns:

```
=====
Variables      Estimates      Std. Err      t-stat      p-value
-----
Var 0          15.602724      6.794225      2.296469      0.033192
=====
```

The estimate for β is 15.60 with a standard error of 6.79. For convenience, define:

$$\hat{\theta}_a = \begin{bmatrix} \hat{\beta}_a & 1 \end{bmatrix}'$$

11.1.3 Part c

To obtain the log-likelihood, multiply the joint pdf together n times:

$$L(\beta, \rho) = \prod_{i=1}^n \frac{(\beta + x)^{-\rho}}{\Gamma(\rho)} y^{\rho-1} e^{-\frac{y}{\beta+x}}$$

Then take the natural log:

$$\ell_n(\beta, \rho) = \sum_{i=1}^n \left(-\rho \ln(\beta + x_i) - \ln(\Gamma(\rho)) + (\rho - 1) \ln(y_i) - \frac{y_i}{\beta + x_i} \right)$$

In Matlab, I once again define the log-likelihood function:


```
logc = @(theta) sum(-theta(2).*log(theta(1)+x) - log(gamma(theta(2))) + ...
    (theta(2)-1)*log(y) - y./(theta(1)+x));
```

where $\theta = \begin{bmatrix} \beta & \rho \end{bmatrix}'$ in the figure.

I call the Newton-Raphson algorithm to solve the MLE problem again:

```
%% Find beta for Part c

b0 = [1;1];
k = 2;
[thetac, stdc, varc] = nr(logc, n, k, b0);
```

getting the following results:

Variables	Estimates	Std. Err	t-stat	p-value
Var 0	-4.718504	2.344866	-2.012270	0.059402
Var 1	3.150896	0.794248	3.967145	0.000904

By not restricting ρ to 1, I find that $\hat{\beta}$ is now -4.72 with a standard error 2.34. I also find that $\hat{\rho} = 3.15$ with a standard error of 0.79. Going forward, let:

$$\hat{\theta}_c = \begin{bmatrix} \hat{\beta}_c & \hat{\rho} \end{bmatrix}'$$

11.1.4 Part d

Here, I test the hypothesis that $\rho = 1$ in four ways.

t-Test

The asymptotic t-test looks as follows:

$$t = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$$

In my case, this looks like:

$$\begin{aligned} t &= \frac{\hat{\rho} - 1}{SE(\hat{\rho})} \\ &= \frac{3.15 - 1}{0.79} \\ &\approx 2.71 \end{aligned}$$

The critical value from the asymptotic t -test is 1.96. So I reject the null.

Likelihood-Ratio Test

The likelihood-ratio test looks as follows:

$$LR = -2 \left[\ell_n \left(\hat{\theta}_r \right) - \ell_n \left(\hat{\theta}_{ur} \right) \right]$$

In my case, this looks like:

$$LR = -2 \left[\ell_n \left(\hat{\beta}_a \right) - \ell_n \left(\hat{\theta}_c \right) \right]$$

Using Matlab and plugging the estimates for $\hat{\beta}_a$ and $\hat{\theta}_c$ from above into the log-likelihoods gives:

$$\begin{aligned} LR &= -2 [-88.44 + 82.92] \\ &\approx 11.04 \end{aligned}$$

The critical value for the likelihood ratio statistic is 3.84, as we have one degree of freedom. I reject the null.

Wald Test

The Wald test looks as follows:

$$W = \left(R \left(\hat{\theta}_{ur} \right) - q \right)' \left(r \left(\hat{\theta}_{ur} \right) \mathcal{I} \left(\hat{\theta}_{ur} \right)^{-1} r \left(\hat{\theta}_{ur} \right)' \right)^{-1} \left(R \left(\hat{\theta}_{ur} \right) - q \right)$$

Because I assume the model is correctly specified and regular, the inverse of the information matrix is the asymptotic variance of my unrestricted estimator. In my case, the Wald statistic therefore looks like:

$$\begin{aligned} W &= (\hat{\rho} - 1)' \left(\begin{bmatrix} 0 & 1 \end{bmatrix} \text{Var} \left(\hat{\theta}_c \right) \begin{bmatrix} 0 & 1 \end{bmatrix}' \right)^{-1} (\hat{\rho} - 1) \\ &= (3.15 - 1)(0.63)^{-1}(3.15 - 1) \\ &\approx 7.33 \end{aligned}$$

The critical value for the Wald statistic is the same as the critical value for the likelihood-ratio statistic, as they both have the same asymptotic distributions. I reject the null.

Score Test

The score test looks as follows:

$$S = \left(\frac{\partial \ell_n \left(\hat{\theta}_r \right)}{\partial \theta} \right)' \mathcal{I} \left(\hat{\theta}_r \right)^{-1} \left(\frac{\partial \ell_n \left(\hat{\theta}_r \right)}{\partial \theta} \right)$$

Note that the information matrix is evaluated at the restricted estimates. To do this, I must first calculate the information matrix *assuming no restrictions*. Then I can plug in the restricted estimates. In Matlab, I evaluate the Hessian using the log-likelihood function from part c and then plug in $\hat{\theta}_a$. This gives a score statistic of:

$$S = \begin{bmatrix} 0 & 7.91 \end{bmatrix} \begin{bmatrix} 124.02 & -2.52 \\ -2.52 & 0.08 \end{bmatrix} \begin{bmatrix} 0 \\ 7.91 \end{bmatrix} \\ \approx 5.12$$

Once again, the critical value is the same as the one for the likelihood-ratio statistic. I reject the null hypothesis.

```
%% Asymptotic Tests

% T-test

% Use the standard errors from the unrestricted model
tstat = (thetac(2) - 1)/sqrt(varc(2,2));
fprintf("\nThe t-stat is %f\n", tstat)

% Likelihood ratio

% Calculate the two likelihood ratios:
lrstat = -2*(loga(betaa) - logc(thetac));
fprintf("\nThe LR-stat is %f\n", lrstat)

% Wald Statistic

R=[0 1]; % Choose rho from the beta vector
q = 1;

W=(R*thetac - q)'*inv(R*varc*R')*(R*thetac-q);
fprintf("\nThe Wald-stat is %f\n", W)

% Score test

scorestat = gradf(logc, thetaa)*inv(-Hessi(logc, thetaa))...
*gradf(logc, thetaa)';

fprintf("\nThe score stat is %f\n", scorestat)
```

The above figure displays the code I used to generate the asymptotic test statistics.

11.2 Generalized Method of Moments Theory

Suppose we have a vector of moment conditions with length L that must be satisfied in our system (whichever system that may be). We write these moments in such a way that:

$$\mathbb{E}[m(\beta)] = 0$$

Our goal is to find a vector of size K , β , that satisfies all the moment conditions that we have.

Identification

β is identified if $\exists! \beta : \mathbb{E}[g(\beta)] = 0$. That is, we can find a solution to the moment conditions if there exists a unique β such that the conditions are satisfied. The system of moment equations has three levels of identification:

- (1) Under-identification: $L < K$, meaning we have more unknowns than equations.
- (2) Just-identified: $L = K$, meaning we have just enough information to find β .
- (3) Over-identified: $L > K$, meaning we have more than enough information to find β .

If $L = K$, then we can use the regular method of moments estimation technique. Consider two examples. First, suppose $\beta = \mu_y$, the sample mean of scalar y . Then our moment condition is:

$$\begin{aligned} m(\mu) &= \mathbb{E}[y_i - \mu] = 0 \\ \mathbb{E}[y_i] &= \mu \\ \frac{1}{n} \sum_{i=1}^n y_i &= \hat{\mu} \end{aligned}$$

Secondly, consider OLS:

$$m(\beta, \sigma^2) = \mathbb{E} \begin{bmatrix} x\varepsilon \\ \varepsilon^2 - \sigma^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Solving this moment conditions will give us our OLS estimates for β and σ^2 .

The crucial assumption behind instrumental variables also fits into the moment conditions framework, where the moment is:

$$m(\beta) = \mathbb{E}[z'u] = 0$$

But what happens when $L > K$, or when the moment condition in the just-identified case is impossible to solve analytically? Take 2SLS. Our moment condition here is the same as under IV:

$$\begin{aligned} m(\beta) &= \mathbb{E}[z'u] = 0 \\ \frac{1}{n} \sum_{i=1}^n z_i(y_i - x_i\beta) &= 0 \end{aligned}$$

$$(z'y)_{L \times 1} = (z'x)_{L \times K} \beta_{K \times 1}$$

Now, instead of exactly identified, we are over-identified. The matrix dimensions will not work for us

to identify β . How would we estimate this? By minimizing the sum of squares:

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} (z'(y - xb))' z'(y - xb)$$

Here is where GMM comes in - we can improve the precision of this estimator by weighting observations:

$$\hat{\beta}^{GMM} = \underset{b}{\operatorname{argmin}} (z'(y - xb))' W z'(y - xb)$$

Minimizing the Criterion Function

The criterion function, in general, looks like:

$$q(\beta) = n\bar{m}(\beta)' W \bar{m}(\beta)$$

where W is a positive definite, symmetric, weighting matrix.

As an example, consider the 2SLS case again:

$$m(\beta) = z'(y - x\beta) \qquad q(\beta) = n(z'(y - x\beta))' W (z'(y - x\beta))$$

Taking the FOC of $q(\beta)$:

$$\begin{aligned} \frac{\partial q(\beta)}{\partial \beta} &= -2nx'zWz' (y - x\hat{\beta}) = 0 \\ x'zWz' (y - x\hat{\beta}) &= 0 \\ x'zWz'y &= x'zWz'x\hat{\beta} \\ (x'zWz'x)^{-1}(x'zWz'y) &= \hat{\beta}^{GMM} \end{aligned}$$

What should W be? In this case, choose $W = (z'z)^{-1}$. Then:

$$\begin{aligned} \hat{\beta}^{GMM} &= (x'z(z'z)^{-1}z'x)^{-1}(x'z(z'z)^{-1}z'y) \\ &= (x'P_zx)^{-1}(x'P_zy) \\ &= \hat{\beta}^{2SLS} \end{aligned}$$

In the just-identified case where $K = L$:

$$\begin{aligned} \hat{\beta}^{GMM} &= (z'x)^{-1}(z'z)(x'z)^{-1}(x'z)(z'z)^{-1}(z'y) \\ &= (z'x)^{-1}(z'y) \\ &= \hat{\beta}^{IV} \end{aligned}$$

GMM Properties

Usually I would include two sections on proving consistency and asymptotic normality. In this case, however, I am already at 9 pages and you will cover this again in Marinho's class. As such, I will simply cite a theorem:

Theorem 6 (Chamberlain (1987)). *GMM is asymptotically efficient among all \sqrt{n} -consistent estimators if all we know is $\mathbb{E}[m(\beta)] = 0$.*

where \sqrt{n} -consistent estimator are estimators, $\hat{\theta}$ such that:

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n f(\text{data}) + R$$

where R converges to zero in probability.

This theorem tells us that GMM estimators are consistent and efficient. By appealing to the central limit theorem, seeing that GMM estimators are asymptotically normal should not be too much of a leap.

Chapter 12

Time Series

12.1 ARMA(1,1)

Consider the ARMA(1,1) model:

$$y_t = d + \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1} \quad \varepsilon_t \sim WN(0, \sigma^2)$$

12.1.1 Part a

Derive the unconditional mean.

Take the expectation:

$$\begin{aligned} \mathbb{E}[y_t] &= \mathbb{E}[y_t = d + \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}] \\ \mu &= d + \phi \mathbb{E}[y_{t-1}] + \mathbb{E}[\varepsilon_t] + \theta \mathbb{E}[\varepsilon_{t-1}] \\ &= d + \phi \mu \end{aligned}$$

This gives us:

$$\mu = \frac{d}{1 - \phi}$$

12.1.2 Part b

Derive the unconditional variance.

Sub-in for d to demean the process:

$$\begin{aligned}
 y_t &= \mu + \phi(y_{t-1} - \mu) + \varepsilon_t + \theta\varepsilon_{t-1} \\
 y_t - \mu &= \phi(y_{t-1} - \mu) + \varepsilon_t + \theta\varepsilon_{t-1} \\
 (y_t - \mu)^2 &= (\phi(y_{t-1} - \mu) + \varepsilon_t + \theta\varepsilon_{t-1})^2 \\
 \mathbb{E}[(y_t - \mu)^2] &= \mathbb{E}[(\phi(y_{t-1} - \mu) + \varepsilon_t + \theta\varepsilon_{t-1})^2] \\
 \gamma_0 &= \phi^2\gamma_0 + \mathbb{E}[\varepsilon_t^2] + \theta^2\mathbb{E}[\varepsilon_{t-1}^2] + 2\phi\theta\mathbb{E}[\varepsilon_t\varepsilon_{t-1}] \\
 \gamma_0 &= \frac{1 + \theta^2 + 2\phi\theta}{1 - \phi^2}\sigma^2
 \end{aligned}$$

12.1.3 Part c

Derive the first and second-order autocovariances.

$$\begin{aligned}
 (y_t - \mu)(y_{t-1} - \mu) &= (\phi(y_{t-1} - \mu) + \varepsilon_t + \theta\varepsilon_{t-1})(y_{t-1} - \mu) \\
 \mathbb{E}[(y_t - \mu)(y_{t-1} - \mu)] &= \phi\mathbb{E}[(y_{t-1} - \mu)^2] + \theta\mathbb{E}[\varepsilon_{t-1}(y_{t-1} - \mu)] \\
 \gamma_1 &= \phi\gamma_0 + \theta\mathbb{E}[\varepsilon_{t-1}(\phi(y_{t-2} - \mu) + \varepsilon_{t-1} + \theta\varepsilon_{t-2})] \\
 &= \phi\gamma_0 + \theta\sigma^2
 \end{aligned}$$

We know γ_0 , so we have a closed-form solution for γ_1 . Now for γ_2 :

$$\begin{aligned}
 (y_t - \mu)(y_{t-2} - \mu) &= (\phi(y_{t-1} - \mu) + \varepsilon_t + \theta\varepsilon_{t-1})(y_{t-2} - \mu) \\
 \mathbb{E}[(y_t - \mu)(y_{t-2} - \mu)] &= \phi\mathbb{E}[(y_{t-1} - \mu)(y_{t-2} - \mu)] + \theta\mathbb{E}[\varepsilon_{t-1}(y_{t-2} - \mu)] \\
 \gamma_2 &= \phi\gamma_1
 \end{aligned}$$

We know γ_1 , so we have solved for γ_2 .

12.1.4 Part d

Given the information set F_t , find the forecasts for the conditional expectations of y_{t+i} for $i \in \{1, 2, 3, 4\}$.

$$\begin{aligned}
 \mathbb{E}_t[y_{t+1}] &= \mathbb{E}_t[d + \phi y_t + \varepsilon_{t+1} + \theta\varepsilon_t] \\
 &= d + \phi y_t + \theta\varepsilon_t
 \end{aligned}$$

$$\begin{aligned}
\mathbb{E}_t[y_{t+2}] &= \mathbb{E}_t[d + \phi y_{t+1} + \varepsilon_{t+2} + \theta \varepsilon_{t+1}] \\
&= d + \phi \mathbb{E}_t[y_{t+1}] \\
&= d + \phi(d + \phi y_t + \theta \varepsilon_t) \\
&= (1 + \phi)d + \phi^2 y_t + \phi \theta \varepsilon_t
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_t[y_{t+3}] &= \mathbb{E}_t[d + \phi y_{t+2} + \varepsilon_{t+3} + \theta \varepsilon_{t+2}] \\
&= d + \phi \mathbb{E}_t[y_{t+2}] \\
&= d + \phi((1 + \phi)d + \phi^2 y_t + \phi \theta \varepsilon_t) \\
&= (1 + \phi + \phi^2)d + \phi^3 y_t + \phi^2 \theta \varepsilon_t
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_t[y_{t+4}] &= \mathbb{E}_t[d + \phi y_{t+3} + \varepsilon_{t+4} + \theta \varepsilon_{t+3}] \\
&= d + \phi \mathbb{E}_t[y_{t+3}] \\
&= d + \phi((1 + \phi + \phi^2)d + \phi^3 y_t + \phi^2 \theta \varepsilon_t) \\
&= (1 + \phi + \phi^2 + \phi^3)d + \phi^4 y_t + \phi^3 \theta \varepsilon_t
\end{aligned}$$

12.1.5 Part e

Derive the forecast $\mathbb{E}_t[y_{t+h}]$ and its limit as $h \rightarrow \infty$.

We can see from Part d that the forecast at horizon h is:

$$\mathbb{E}_t[y_{t+h}] = (1 + \phi + \phi^2 + \dots + \phi^{h-1})d + \phi^h y_t + \phi^{h-1} \theta \varepsilon_t$$

Take the limit:

$$\begin{aligned}
\lim_{h \rightarrow \infty} \mathbb{E}_t[y_{t+h}] &= \lim_{h \rightarrow \infty} (1 + \phi + \phi^2 + \dots + \phi^{h-1})d + \phi^h y_t + \phi^{h-1} \theta \varepsilon_t \\
&= \frac{d}{1 - \phi} \\
&= \mu
\end{aligned}$$

12.1.6 Part f

Calculate the variance of the forecast for the next two time periods. Is the variance increasing or decreasing as the horizon grows larger?

$$\begin{aligned}
 \text{Var}_t(y_{t+1}) &= \text{Var}(d + \phi y_t + \varepsilon_{t+1} + \theta \varepsilon_t) \\
 &= \text{Var}(\varepsilon_{t+1}) \\
 &= \sigma^2
 \end{aligned}$$

$$\begin{aligned}
 \text{Var}_t(y_{t+2}) &= \text{Var}(d + \phi y_{t+1} + \varepsilon_{t+2} + \theta \varepsilon_{t+1}) \\
 &= \text{Var}(d + \phi[d + \phi y_t + \varepsilon_{t+1} + \theta \varepsilon_t] + \varepsilon_{t+2} + \theta \varepsilon_{t+1}) \\
 &= \text{Var}([\phi + \theta]\varepsilon_{t+1} + \varepsilon_{t+2}) \\
 &= [\phi + \theta]^2 \sigma^2 + \sigma^2
 \end{aligned}$$

The variance grows larger as the horizon increases.

12.2 AR(2) Process

An AR(2) process relates a random variable located in time period t to two lags of that random variable. We usually write this as: $y_t = d + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$ where $\varepsilon_t \sim WN(0, \sigma^2)$. Recall that white noise implies that ε_t has no autocorrelation but does not imply that ε_t is not independent from its past.

- (a) Find $\mathbb{E}[y_t]$
- (b) Find $\text{Var}(y_t)$
- (c) Find γ_0, γ_1 , and γ_2
- (d) Find the impulse responses for a shock ε_t for $k \in 0, 1, 2, 3, 4$
- (e) Find $\mathbb{E}[y_{t+3}|t]$

12.2.1 Solution: Part a

We just take the expectation of the AR(2) equation:

$$\begin{aligned}
 \mathbb{E}[y_t] &= \mathbb{E}[d + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t] \\
 \mu_y &= d + \phi_1 \mathbb{E}[y_{t-1}] + \phi_2 \mathbb{E}[y_{t-2}] + 0 \\
 \mu_y &= d + (\phi_1 + \phi_2) \mu_y \\
 \mu_y(1 - \phi_1 - \phi_2) &= d \\
 \mu_y &= \frac{d}{1 - \phi_1 - \phi_2}
 \end{aligned}$$

12.2.2 Solution: Part b

Similarly, we take the variance of the AR(2) equation:

$$\begin{aligned} \text{Var}(y_t) &= \text{Var}(d + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t) \\ \gamma_0 &= \phi_1^2 \text{Var}(y_{t-1}) + \phi_2^2 \text{Var}(y_{t-2}) + 2\phi_1 \phi_2 \text{Cov}(y_{t-1}, y_{t-2}) + \sigma^2 \\ \gamma_0 &= (\phi_1^2 + \phi_2^2)\gamma_0 + 2\phi_1 \phi_2 \gamma_1 + \sigma^2 \end{aligned}$$

12.2.3 Solution: Part c

In class, Drew proved that the autocorrelation function for an AR(2) is given by: $\rho_j = \phi_1 \rho_{j-1} + \phi_2 \rho_{j-2}$. Using this equation we can find what we need. First, let's find γ_1 :

$$\begin{aligned} \gamma_1 &= \phi_1 \gamma_0 + \phi_2 \gamma_{-1} && \text{Recall that } \gamma_{-i} = \gamma_i \\ \gamma_1 &= \phi_1 \gamma_0 + \phi_2 \gamma_1 \\ \gamma_1 &= \frac{\phi_1 \gamma_0}{1 - \phi_2} \end{aligned}$$

Plugging this into the equation we found for γ_0 in part a:

$$\begin{aligned} \gamma_0 &= (\phi_1^2 + \phi_2^2)\gamma_0 + 2\phi_1 \phi_2 \frac{\phi_1 \gamma_0}{1 - \phi_2} + \sigma^2 \\ \gamma_0(1 - \phi_1^2 - \phi_2^2 - 2\phi_1 \phi_2 \frac{\phi_1}{1 - \phi_2}) &= \sigma^2 \\ \gamma_0 &= \frac{\sigma^2}{1 - \phi_1^2 - \phi_2^2 - 2\phi_1 \phi_2 \frac{\phi_1}{1 - \phi_2}} \end{aligned}$$

Plug this into the expression for γ_1 :

$$\begin{aligned} \gamma_1 &= \frac{\phi_1 \gamma_0}{1 - \phi_2} \\ \gamma_1 &= \frac{\phi_1}{1 - \phi_2} \left[\frac{\sigma^2}{1 - \phi_1^2 - \phi_2^2 - 2\phi_1 \phi_2 \frac{\phi_1}{1 - \phi_2}} \right] \end{aligned}$$

Lastly, use the ACF to find γ_2 :

$$\begin{aligned} \gamma_2 &= \phi_1 \gamma_1 + \phi_2 \gamma_0 \\ \gamma_2 &= \frac{\phi_1^2}{1 - \phi_2} \left[\frac{\sigma^2}{1 - \phi_1^2 - \phi_2^2 - 2\phi_1 \phi_2 \frac{\phi_1}{1 - \phi_2}} \right] + \frac{\phi_2 \sigma^2}{1 - \phi_1^2 - \phi_2^2 - 2\phi_1 \phi_2 \frac{\phi_1}{1 - \phi_2}} \end{aligned}$$

What if we did not have the ACF provided? We could always calculate γ_0 , γ_1 , and γ_2 by brute force. First, we rewrite the AR(2):

$$\begin{aligned}y_t &= d + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t \\y_t &= \mu_y(1 - \phi_1 - \phi_2) + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t \\(y_t - \mu_y) &= \phi_1(y_{t-1} - \mu_y) + \phi_2(y_{t-2} - \mu_y) + \varepsilon_t\end{aligned}$$

If we wanted to calculate the unconditional variance γ_0 , for example, then:

$$\begin{aligned}(y_t - \mu_y)^2 &= [\phi_1(y_{t-1} - \mu_y) + \phi_2(y_{t-2} - \mu_y) + \varepsilon_t](y_t - \mu_y) \\ \mathbb{E}[(y_t - \mu_y)^2] &= \mathbb{E}[(\phi_1(y_{t-1} - \mu_y) + \phi_2(y_{t-2} - \mu_y) + \varepsilon_t)(y_t - \mu_y)] \\ &= \gamma_0\end{aligned}$$

What if we wanted to calculate γ_1 ?

$$\begin{aligned}(y_t - \mu_y)(y_{t-1} - \mu_y) &= [\phi_1(y_{t-1} - \mu_y) + \phi_2(y_{t-2} - \mu_y) + \varepsilon_t](y_{t-1} - \mu_y) \\ \mathbb{E}[(y_t - \mu_y)(y_{t-1} - \mu_y)] &= \mathbb{E}[(\phi_1(y_{t-1} - \mu_y) + \phi_2(y_{t-2} - \mu_y) + \varepsilon_t)(y_{t-1} - \mu_y)] \\ &= \gamma_1\end{aligned}$$

Similarly for γ_2 :

$$\begin{aligned}(y_t - \mu_y)(y_{t-2} - \mu_y) &= [\phi_1(y_{t-1} - \mu_y) + \phi_2(y_{t-2} - \mu_y) + \varepsilon_t](y_{t-2} - \mu_y) \\ \mathbb{E}[(y_t - \mu_y)(y_{t-2} - \mu_y)] &= \mathbb{E}[(\phi_1(y_{t-1} - \mu_y) + \phi_2(y_{t-2} - \mu_y) + \varepsilon_t)(y_{t-2} - \mu_y)] \\ &= \gamma_2\end{aligned}$$

12.2.4 Solution: Part d

Impulses are defined as $\frac{\partial y_{t+k}}{\partial \varepsilon_t}$. Let's find this for $k = 0$ first:

$$\begin{aligned}y_t &= d + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t \\ \frac{\partial y_t}{\partial \varepsilon_t} &= 1\end{aligned}$$

Now for $k = 1$:

$$\begin{aligned}
 y_{t+1} &= d + \phi_1 y_t + \phi_2 y_{t-1} + \varepsilon_{t+1} \\
 y_{t+1} &= d + \phi_1 (d + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t) + \phi_2 y_{t-1} + \varepsilon_{t+1} \\
 \frac{\partial y_{t+1}}{\partial \varepsilon_t} &= \phi_1
 \end{aligned}$$

For $k = 2$:

$$\begin{aligned}
 y_{t+2} &= d + \phi_1 y_{t+1} + \phi_2 y_t + \varepsilon_{t+2} \\
 \frac{\partial y_{t+2}}{\partial \varepsilon_t} &= \phi_1 \cdot \frac{\partial}{\partial \varepsilon_t} y_{t+1} + \phi_2 \cdot \frac{\partial}{\partial \varepsilon_t} y_t \\
 &= \phi_1^2 + \phi_2
 \end{aligned}$$

And for $k = 3$:

$$\begin{aligned}
 y_{t+3} &= d + \phi_1 y_{t+2} + \phi_2 y_{t+1} + \varepsilon_{t+3} \\
 \frac{\partial y_{t+3}}{\partial \varepsilon_t} &= \phi_1 \cdot \frac{\partial}{\partial \varepsilon_t} y_{t+2} + \phi_2 \cdot \frac{\partial}{\partial \varepsilon_t} y_{t+1} \\
 &= \phi_1^3 + 2\phi_1\phi_2
 \end{aligned}$$

Finally for $k = 4$:

$$\begin{aligned}
 y_{t+4} &= d + \phi_1 y_{t+3} + \phi_2 y_{t+2} + \varepsilon_{t+4} \\
 \frac{\partial y_{t+4}}{\partial \varepsilon_t} &= \phi_1 \cdot \frac{\partial}{\partial \varepsilon_t} y_{t+3} + \phi_2 \cdot \frac{\partial}{\partial \varepsilon_t} y_{t+2} \\
 &= \phi_1^4 + 3\phi_1^2\phi_2 + \phi_2^2
 \end{aligned}$$

As you can see, calculating impulse responses for an AR(2) by hand is much more labor-intensive than calculating them for an AR(1).

12.2.5 Solution: Part e

This is similar to calculating the unconditional moment. Now, though, we know all the variables up to time t . So:

$$\begin{aligned}
\mathbb{E}[y_{t+3}|t] &= \mathbb{E}[d + \phi_1 y_{t+2} + \phi_2 y_{t+1} + \varepsilon_{t+3}|t] \\
&= d + \phi_1 \mathbb{E}[y_{t+2}|t] + \phi_2 \mathbb{E}[y_{t+1}|t] + 0 \\
&= d + \phi_1 \mathbb{E}[d + \phi_1 y_{t+1} + \phi_2 y_t + \varepsilon_{t+2}|t] + \phi_2 \mathbb{E}[d + \phi_1 y_t + \phi_2 y_{t-1} + \varepsilon_{t+1}|t] \\
&= d + \phi_1 (d + \phi_2 y_t + \phi_1 \mathbb{E}[d + \phi_1 y_t + \phi_2 y_{t-1} + \varepsilon_{t+1}|t]) + \phi_2 (d + \phi_1 y_t + \phi_2 y_{t-1}) \\
&= d + \phi_1 d + \phi_1 \phi_2 y_t + \phi_1^2 d + \phi_1^3 y_t + \phi_1^2 \phi_2 y_{t-1} + \phi_2^2 y_{t-1} + \phi_2 d + \phi_2 \phi_1 y_t + \phi_2^2 y_{t-1} \\
&= d(1 + \phi_1 + \phi_1^2 + \phi_2) + y_t(\phi_1^3 + 2\phi_1 \phi_2) + y_{t-1}(\phi_1^2 \phi_2 + \phi_2^2)
\end{aligned}$$

Notice how we treated any variable dated at time t or before as if they were constants. That's the key to solving conditional expectations or variances with respect to time.

12.3 Deriving the MA(∞) Form for an AR(1)

Drew went over this in class, but this is an important wrench to have in your toolbox. The MA(∞) form splits your autoregressive process into the stationary mean and sum of impulse shocks. We start with a basic AR(1):

$$\begin{aligned}
y_t &= d + \phi y_{t-1} + \varepsilon_t && \text{Recursively sub-in:} \\
&= d + \phi(d + \phi y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\
&= d(1 + \phi) + \phi^2 y_{t-2} + \phi \varepsilon_{t-1} + \varepsilon_t \\
&= d(1 + \phi) + \phi^2 (d + \phi y_{t-3} + \varepsilon_{t-2}) + \phi \varepsilon_{t-1} + \varepsilon_t \\
&= d(1 + \phi + \phi^2) + \phi^3 y_{t-3} + \phi^2 \varepsilon_{t-2} + \phi \varepsilon_{t-1} + \varepsilon_t \\
&= d(1 + \phi + \phi^2) + \phi^3 (d + \phi y_{t-4} + \varepsilon_{t-3}) + \phi^2 \varepsilon_{t-2} + \phi \varepsilon_{t-1} + \varepsilon_t \\
&= d(1 + \phi + \phi^2 + \phi^3) + \phi^4 y_{t-4} + \phi^3 \varepsilon_{t-3} + \phi^2 \varepsilon_{t-2} + \phi \varepsilon_{t-1} + \varepsilon_t
\end{aligned}$$

We see a pattern emerging here. Using induction, we can write this process as:

$$= d \sum_{i=1}^{\infty} \phi^{i-1} + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$$

Using the fact that $\sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}$, and making the assumption that $\phi < 1$, this becomes:

$$= \frac{d}{1-\phi} + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$$

We can simplify further. Solving for the mean of an AR(1) process:

$$\begin{aligned}\mathbb{E}[y_t] &= \mathbb{E}[d + \phi y_{t-1} + \varepsilon_t] \\ \mu_y &= d + \phi \mathbb{E}[y_{t-1}] + 0 \\ \mu_y &= d + \phi \mu_y \\ \mu_y &= \frac{d}{1 - \phi}\end{aligned}$$

Therefore, the MA(∞) form becomes:

$$y_t = \mu_y + \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}$$

Where μ_y is the unconditional mean of Y and the sum contains the impulse shocks of ε .

Chapter 13

Final Review

13.1 Asymptotic Properties of Non-Linear Least Squares

The general formula for a regression equation is:

$$y_i = h(x_i, \beta) + \varepsilon_i$$

where $h(x_i, \beta)$ is some function of our right-hand side variables. To find the least squares β :

$$\begin{aligned} \min_{\beta} S(\beta) &= \min_{\beta} \sum_{i=1}^n (y_i - h(x_i, \beta))^2 \\ \frac{\partial S(\beta)}{\partial \beta} &= -2 \sum_{i=1}^n (y_i - h(x_i, \beta)) \frac{\partial h}{\partial \beta} \end{aligned}$$

We then set this equation equal to zero and solve, usually using a numerical solver like the Newton-Raphson algorithm.

Consistency

Denote $\hat{\beta}_{NL}$ as the non-linear least squares estimator. Then by definition:

$$\frac{\partial S(\hat{\beta}_{NL})}{\partial \beta} = 0$$

Assume that there is exactly one β that satisfies this condition (that is, that β is identified). Then we know that $\hat{\beta}_{NL}$ is consistent if:

$$plim \frac{1}{n} \frac{\partial S(\hat{\beta}_{NL})}{\partial \beta} = plim \frac{1}{n} \frac{\partial S(\beta_0)}{\partial \beta}$$

where β_0 denotes the true β . Now, recall the continuous mapping theorem:

Theorem 7 (Continuous Mapping Theorem). *Let x_n be a sequence of real-valued random vectors and let $h : \mathcal{R}^k \rightarrow \mathcal{R}^m$. Define the set of discontinuous points as:*

$$D_h = \{x \in \mathcal{X} : h(\cdot) \text{ is discontinuous at } x\}$$

Now, if $P(x \in D_h) = 0$ and:

$$(i) \text{ if } x_n \xrightarrow{P} x, \text{ then } h(x_n) \xrightarrow{P} h(x)$$

$$(ii) \text{ if } x_n \xrightarrow{d} x, \text{ then } h(x_n) \xrightarrow{d} h(x)$$

$$(iii) \text{ if } x_n \xrightarrow{a.s.} x, \text{ then } h(x_n) \xrightarrow{a.s.} h(x)$$

The converse of this theorem holds true when $h(\cdot)$ is an injective (one-to-one) function. Stated another way: [Converse] Let $h(\cdot)$ be a continuous, injective function such that $h(x_n) \xrightarrow{P} h(x)$. Then:

$$x_n \xrightarrow{P} x$$

How do we plan on applying this converse? x_n in our scenario is $\hat{\beta}_{NL}$. We want to show that $\hat{\beta}_{NL} \xrightarrow{P} \beta$. We also have a continuous, one-to-one function in $\frac{\partial S(\hat{\beta}_{NL})}{\partial \beta}$. We already know that:

$$\frac{\partial S(\hat{\beta}_{NL})}{\partial \beta} = 0$$

If we take the probability limit:

$$\frac{1}{n} \frac{\partial S(\hat{\beta}_{NL})}{\partial \beta} \xrightarrow{P} 0$$

Now we need to show that $\frac{\partial S(\beta_0)}{\partial \beta} = 0$ too. We begin with the sample mean of the derivative:

$$\begin{aligned} \frac{1}{n} \frac{\partial S(\beta_0)}{\partial \beta} &= \frac{1}{n} \sum_{i=1}^n (y_i - h(x_i, \beta)) \frac{\partial h(x_i, \beta)}{\partial \beta} \\ &\xrightarrow{P} \mathbb{E} \left[\varepsilon_i \frac{\partial h(x_i, \beta)}{\partial \beta} \right] \\ &= \mathbb{E}_x \left[\mathbb{E}[\varepsilon_i | x] \frac{\partial h(x_i, \beta)}{\partial \beta} \right] \\ &= 0 \end{aligned}$$

Therefore:

$$plim \frac{1}{n} \frac{\partial S(\hat{\beta}_{NL})}{\partial \beta} = plim \frac{1}{n} \frac{\partial S(\beta_0)}{\partial \beta}$$

By the converse of the continuous mapping theorem then:

$$\hat{\beta}_{NL} \xrightarrow{P} \beta_0$$

We have proven that the non-linear least squares estimator is consistent.

Asymptotic Normality

Start with the first order condition:

$$\frac{\partial S(\hat{\beta}_{NL})}{\partial \beta} = -2 \sum_{i=1}^n (y_i - h(x_i, \beta)) \frac{\partial h(x_i, \beta)}{\partial \beta}$$

Denote the right-hand side as $g(\hat{\beta})$. Doing a first-order Taylor approximation around the true β delivers:

$$g(\hat{\beta}_{NL}) = g(\beta) + \mathbf{H}(\tilde{\beta}_{NL})(\hat{\beta}_{NL} - \beta) \quad \text{for } \tilde{\beta}_{NL} \in (\hat{\beta}_{NL}, \beta) \quad (13.1)$$

Looking just at the Hessian (in red):

$$\begin{aligned} H(\tilde{\beta}_{NL}) &= \frac{\partial^2 S(\tilde{\beta}_{NL})}{\partial \beta \partial \beta'} \\ &= 2 \sum_{i=1}^n \frac{\partial h(x_i, \tilde{\beta}_{NL})}{\partial \beta} \frac{\partial h(x_i, \tilde{\beta}_{NL})}{\partial \beta'} - 2 \sum_{i=1}^n (y_i - h(x_i, \tilde{\beta}_{NL})) \frac{\partial^2 h(x_i, \tilde{\beta}_{NL})}{\partial \beta \partial \beta'} \\ \frac{1}{n} H(\tilde{\beta}_{NL}) &\xrightarrow{P} 2\mathbb{E} \left[\frac{\partial h(x_i, \tilde{\beta}_{NL})}{\partial \beta} \frac{\partial h(x_i, \tilde{\beta}_{NL})}{\partial \beta'} \right] \\ &= 2Q_0 \end{aligned}$$

Going back to equation (1) and dividing by \sqrt{n} :

$$\underbrace{\frac{1}{\sqrt{n}} g(\hat{\beta}_{NL})}_{=0} = \frac{1}{\sqrt{n}} g(\beta) + \frac{1}{n} H(\tilde{\beta}_{NL}) \sqrt{n} (\hat{\beta}_{NL} - \beta) \quad (13.2)$$

Now let's look at the gradient (in blue):

$$\begin{aligned} \frac{1}{\sqrt{n}} g(\beta) &= \frac{-2}{\sqrt{n}} \sum_{i=1}^n (y_i - h(x_i, \beta)) \frac{\partial h(x_i, \beta)}{\partial \beta} \\ &= \frac{-2}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \frac{\partial h(x_i, \beta)}{\partial \beta} \end{aligned}$$

$$\begin{aligned}
& \xrightarrow{d} N\left(0, 4\mathbb{E}\left[\varepsilon_i \frac{\partial h(x_i, \beta)}{\partial \beta} \frac{\partial h(x_i, \beta)'}{\partial \beta} \varepsilon_i'\right]\right) \\
& = N\left(0, 4\mathbb{E}_x\left[\frac{\partial h(x_i, \beta)'}{\partial \beta} \mathbb{E}[\varepsilon_i \varepsilon_i' | x] \frac{\partial h(x_i, \beta)'}{\partial \beta}\right]\right) \\
& = N(0, 4\sigma^2 Q_0)
\end{aligned}$$

Rearranging equation (2) gives:

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_{NL} - \beta) &= -\frac{1}{n}H(\tilde{\beta}_{NL})^{-1} \frac{1}{\sqrt{n}}g(\beta) \\
&\xrightarrow{d} \frac{1}{2}Q_0^{-1}N(0, 4\sigma^2 Q_0) \\
&= N(0, \sigma^2 Q_0^{-1})
\end{aligned}$$

This completes the asymptotic normality proof.

13.2 Cochrane-Orcutt Procedure

The Cochrane-Orcutt procedure is a way to estimate AR(1) processes. In class, you generally dealt with an AR(1) of the following form:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad (13.3)$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad (13.4)$$

Lag the regression equation:

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1}$$

Now multiply by ρ :

$$\rho y_{t-1} = \rho \beta_0 + \rho \beta_1 x_{t-1} + \rho \varepsilon_{t-1} \quad (13.5)$$

Subtract equations (5) from equation (3) to get:

$$\begin{aligned}
y_t - \rho y_{t-1} &= \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + \varepsilon_t - \rho \varepsilon_{t-1} \\
y_t - \rho y_{t-1} &= \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + u_t \\
y_t^* &= \beta_0^* + \beta_1 x_t^* + u_t
\end{aligned} \quad (13.6)$$

Recall that u_t is white noise, so it satisfies all of the classical regression assumptions. We can use OLS.

In reality, we do not know ρ . So we must estimate ρ by running OLS on equation (4). Then we can start the Cochrane-Orcutt iteration step:

1. Estimate equation (6) via OLS to obtain $\hat{\beta}$ and OLS residuals $\hat{\varepsilon}$.
2. Estimate new $\hat{\rho}$.

3. Transform original data into y_t^* and x_t^* .
4. Re-estimate $\hat{\beta}$.
5. Repeat until $\hat{\beta}_i$ converges to $\hat{\beta}_i$.

13.2.1 Prais-Winsten

Recall that the variance of the error term is (for a 3×3 for simplicity):

$$\begin{aligned} \text{Var}(\varepsilon) &= \sigma^2 \Omega \\ &= \sigma_u^2 \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix} \end{aligned}$$

Because Ω is a positive definite matrix (by definition of the variance-covariance matrix), it has an inverse and Cholesky decomposition:

$$\Omega^{-1} = P'P$$

where The triangular matrix P takes the following form:

$$P = \begin{bmatrix} \sqrt{1 - \rho^2} & 0 & 0 \\ -\rho & 1 & 0 \\ 0 & -\rho & 1 \end{bmatrix}$$

Now, premultiply y by P :

$$\begin{aligned} Py &= \begin{bmatrix} \sqrt{1 - \rho^2} & 0 & 0 \\ -\rho & 1 & 0 \\ 0 & -\rho & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\ y^* &= \begin{bmatrix} \sqrt{1 - \rho^2} y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \end{bmatrix} \end{aligned}$$

Then premultiply x by P :

$$\begin{aligned} Px &= \begin{bmatrix} \sqrt{1 - \rho^2} & 0 & 0 \\ -\rho & 1 & 0 \\ 0 & -\rho & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ x^* &= \begin{bmatrix} \sqrt{1 - \rho^2} x_1 \\ x_2 - \rho x_1 \\ x_3 - \rho x_2 \end{bmatrix} \end{aligned}$$

Then run OLS on:

$$y^* = x^* \beta + \varepsilon^*$$

13.3 Reverse Regression

We now have in mind the classic regression equation:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Reverse regression is a way to solve measurement error in the right-hand side variable. Suppose x_i can be written as:

$$x_i = x_i^* + u_i$$

Then we can rewrite the standard regression equation as:

$$\begin{aligned} y_i &= \alpha + \beta(x_i - u_i) + \varepsilon_i \\ &= \alpha + \beta x_i + \varepsilon_i - \beta u_i \end{aligned}$$

We can see that the classical regression assumptions do not necessarily hold here:

$$\begin{aligned} \text{Cov}(x_i, \varepsilon_i - \beta u_i) &= \text{Cov}(x_i^* + u_i, \varepsilon_i - \beta u_i) \\ &\neq 0 \end{aligned}$$

Instead, we can run a reverse regression:

$$x_i^* = \gamma_0 + \gamma_1 y_i + \delta_i$$

δ_i satisfies all the classical regression assumptions, so $\hat{\gamma}_1$ will be unbiased and consistent. $\hat{\gamma}_1$ gives us an upper bound on the value of $\hat{\beta}_1$. Why? Recall the Cauchy-Schwartz Inequality:

$$\left(\sum_{i=1}^n x_i y_i \right)^2 \leq \left(\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2 \right)$$

Notice that:

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \\ \hat{\gamma} &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^2} \end{aligned}$$

Multiply $\hat{\gamma}$ and $\hat{\beta}$:

$$\hat{\beta}\hat{\gamma} = \frac{(\sum_{i=1}^n x_i y_i)^2}{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}$$

Using the Cauchy-Schwartz inequality:

$$\begin{aligned}\hat{\beta}\hat{\gamma} &\leq \frac{(\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2)}{(\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2)} \\ \hat{\beta} &\leq \frac{1}{\hat{\gamma}}\end{aligned}$$

So if $\hat{\beta}$ is positive, $\hat{\gamma}$ gives us an upper bound on $\hat{\beta}$.