Contents

(Last Updated September 22, 2024)

LECTURE 1

Matrix and Statistics Concepts

1.1 What is the Role of Matrices in Econometrics

- Matrices serve several very practical purposes

1) In data analysis we store large amounts of data, which we typically organize into tables
   - Rows are observations and columns are variables
   - Tables are just matrices with labels
2) In data analysis we need to perform computations on huge lists of numbers
   - This is easily accomplished with matrix algebra
3) In theoretical statistics we need to represent and understand the statistical properties of huge groups of numbers. This is possible using matrices

1.2 Organizing Numbers in Matrices

- An $m \times n$ matrix has $m * n$ elements stored in m rows and n columns
  - In this class, and typically outside of this class, columns are filled first, e.g., a 3×2 matrix with elements 1, 2, 3, 4, 5 ,6 will look like $\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$
- Transposing a matrix involves flipping the dimensions. If we call the matrix above A, the transpose of A in this class is denoted A' and takes the form $A' = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$

1.3 Matrix Multiplication

- If two matrices have the same adjacent dimensions we can apply matrix multiplication
- This will create a new matrix with dimension equal to the outer dimension
- If $A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$, then A is 3×2 and A' is 2×3
- $A'A$ denotes multiplying transpose of A by itself. This is a 2×3 multiplied by 3×2
  - The adjacent dimensions are the same (3) so matrix multiplication is possible
  - The new dimensions are determined by the outer dimensions 2×2

1.4 Vectors

- Vectors are special cases of matrices where one dimension is 1
- A column vector has many rows but only 1 column, for example $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$
- A row vector has only 1 row but many columns
  - There are no row vectors, only transposed column vectors
  - For example $X' = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$

1.5 Vector Inner Product

- If $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$
- Consider $X'X$ this is 1×3 multiplying 3×1, so the result is scaler
- $X'X = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1^2 + x_2^2 + x_3^2 = \sum_{i=1}^{3} x_i^2$
- Thus if I want to take the sum of squared values of a list of numbers, put them in a column vector X and compute $X'X$

1.6 Vector Outer Product

- If $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$
- Consider XX' this is 3×1 multiplying 1×3, so the result is a 3×3 matrix
- $XX' = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} = \begin{bmatrix} x_1^2 & x_1x_2 & x_1x_3 \\ x_2x_1 & x_2^2 & x_2x_3 \\ x_3x_1 & x_3x_2 & x_3^2 \end{bmatrix}$
- Notice this matrix square and symmetric and has all positive numbers along the diagonal

1.7 Scalar Random Variables

- If x_1 is a random variable we have $\mathrm{E}(x_1)$ is the expected value
- $\mathrm{Var}(x_1)$ is its variance. Two ways to express the variance
  - $\mathrm{Var}(x_1) = \mathrm{E}\left[(x_1 - \mathrm{E}(x_1))^2\right] = \mathrm{E}(x_1^2) - \mathrm{E}(x_1)^2$

1.8 Vectors of Random Variables

- If $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$
- We have $\mathrm{E}(X) = \mathrm{E}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \mathrm{E}(x_1) \\ \mathrm{E}(x_2) \\ \mathrm{E}(x_3) \end{bmatrix}$
- Then what is $\mathrm{Var}(X)$? $=?$ $\begin{bmatrix} \mathrm{Var}(x_1) \\ \mathrm{Var}(x_2) \\ \mathrm{Var}(x_3) \end{bmatrix}$? NO!

1.9 Variance-Covariance Matrix

- If $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$
- $\mathrm{Var}(X) = \mathrm{E}(XX') - \mathrm{E}(X)\,\mathrm{E}(X)$
- Recall $XX' = \begin{bmatrix} x_1^2 & x_1x_2 & x_1x_3 \\ x_2x_1 & x_2^2 & x_2x_3 \\ x_3x_1 & x_3x_2 & x_3^2 \end{bmatrix}$
- So $\mathrm{E}(XX') = \begin{bmatrix} \mathrm{E}(x_1^2) & \mathrm{E}(x_1x_2) & \mathrm{E}(x_1x_3) \\ \mathrm{E}(x_2x_1) & \mathrm{E}(x_2^2) & \mathrm{E}(x_2x_3) \\ \mathrm{E}(x_3x_1) & \mathrm{E}(x_3x_2) & \mathrm{E}(x_3^2) \end{bmatrix}$

1.10 Variance-Covariance Matrix

- The other part $\mathrm{E}(X)\,\mathrm{E}(X) = \begin{bmatrix} \mathrm{E}(x_1) \\ \mathrm{E}(x_2) \\ \mathrm{E}(x_3) \end{bmatrix} \begin{bmatrix} \mathrm{E}(x_1) & \mathrm{E}(x_2) & \mathrm{E}(x_3) \end{bmatrix}$
- Which is $\mathrm{E}(X)\,\mathrm{E}(X) = \begin{bmatrix} \mathrm{E}(x_1)^2 & \mathrm{E}(x_1)\,\mathrm{E}(x_2) & \mathrm{E}(x_1)\,\mathrm{E}(x_3) \\ \mathrm{E}(x_2)\,\mathrm{E}(x_1) & \mathrm{E}(x_2)^2 & \mathrm{E}(x_2)\,\mathrm{E}(x_3) \\ \mathrm{E}(x_3)\,\mathrm{E}(x_1) & \mathrm{E}(x_3)\,\mathrm{E}(x_2) & \mathrm{E}(x_3)^2 \end{bmatrix}$
- Combining with results above, $\mathrm{Var}(X) = \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_2) & \mathrm{Cov}(x_1, x_3) \\ \mathrm{Cov}(x_2, x_1) & \mathrm{Var}(x_2) & \mathrm{Cov}(x_2, x_3) \\ \mathrm{Cov}(x_3, x_1) & \mathrm{Cov}(x_3, x_2) & \mathrm{Var}(x_3) \end{bmatrix}$
- Because $\mathrm{E}(x_1^2) - \mathrm{E}(x_1)^2 = \mathrm{Var}(x_1)$ and $\mathrm{E}(x_1x_3) - \mathrm{E}(x_1)\,\mathrm{E}(x_3) = \mathrm{Cov}(x_1, x_3)$

1.11 Functions of scaler Random Variables

- x is random with $\mathrm{E}(x)$ and $\mathrm{Var}(x)$
- Define $z = b \cdot x$
- Then the following properties hold

  - $\mathrm{E}(z) = \mathrm{E}(b \cdot x) = b\,\mathrm{E}(x)$
  - $\mathrm{Var}(z) = \mathrm{Var}(b \cdot x) = b^2\,\mathrm{Var}(x)$

1.12 Common Function of Random Variables is to Standardize Them

- If x is random with $\mathrm{E}(x)$ and $\mathrm{Var}(x)$ given
- Define $z = \dfrac{x-a}{\sqrt{b}}$, where $a = \mathrm{E}(X)$ and $b = \mathrm{Var}(X)$
- In this case z has mean zero an variance 1
- $\mathrm{E}(z) = \mathrm{E}\left(\dfrac{x-a}{\sqrt{b}}\right) = \left(\dfrac{1}{\sqrt{b}}\right)(\mathrm{E}\,(x) - \mathrm{E}\,(a)) = 0$
- $\mathrm{Var}(z) = \mathrm{Var}\left(\dfrac{x-a}{\sqrt{b}}\right) = \left(\dfrac{1}{\sqrt{b}}\right)^2 \mathrm{Var}(x-a) = 1$

1.13 Functions of Vector Random Variables

- $X = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}'$ is random with $\mathrm{E}(X)$ and $\mathrm{Var}(X)$
- If $A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}'$ vector of constants
- Define $z = A'X$ (notice z is a scaler
- Then the following properties hold
  - $\mathrm{E}(z) = \mathrm{E}(A'X) = A'\,\mathrm{E}(X)$
  - $\mathrm{Var}(z) = \mathrm{Var}(A'X) = A'\,\mathrm{Var}(X)A$

1.14 Bi-Variate Normal

- Two random variables, x_1 and x_2
  - $x_1 \sim N(\mu_1, \sigma_1^2)$
  - $x_2 \sim N(\mu_2, \sigma_2^2)$
  - $\mathrm{Cov}(x_1, x_2) = \sigma_{12}$
- Define $X = \begin{bmatrix} x_1 & x_2 \end{bmatrix}'$ a vector of random variables

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right)$$

- $\mathrm{E}(x_1|x_2) = \mu_1 + \dfrac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2)$
- $\mathrm{Var}(x_1|x_2) = \sigma_1^2 - \dfrac{(\sigma_{12})^2}{\sigma_2^2}$

Lecture 2

Statistics and Data in R

2.1 Working with R

- We are going to review some R syntax and commands
- You should put your code into a script for reproducibility and debugging
- Later we will focus on formatting output in markdown files

2.2 Data Types in R

```
# In R, the number sign is a comment line
#   Comment lines are ignored when the code is run
#   and are a good place to document your code

# A scalar (single number)
x = 1

# Create a list of numbers
x = c(1,2,3,4,5,6,7,8)

# Accessing the third element
x[3]

# Convenience Functions (list containing all ones)
x = rep(1,3)

# Populate integers
x = 1:3
```

2.3 Matrices in R

```
# the code c(1,2,3,4,5,6,7,8) creates a sequence of numbers
#   running from 1 to 8
# There are 3 ways that we can put these numbers into a 4 x 2 matrix

# METHOD 1: The matrix function
A = matrix(c(1,2,3,4,5,6,7,8),nrow=4,ncol=2)

# METHOD 2: Make list of numbers, assign dimension
A = c(1,2,3,4,5,6,7,8)
dim(A) = c(4,2)

# METHOD 3: Column bind 2 separate lists
A = cbind(c(1,2,3,4),c(5,6,7,8))
```

2.4 Matrix Operations

```r
# A couple of useful functions for matrices

A = matrix(c(1,2,3,4,5,6,7,8),nrow=4,ncol=2)

# Get the number of rows or columns
nrow(A)
ncol(A)

# transpose the matrix
t(A)
# notice that in R, t is a function
#   this means never assign a variable the name t
#   because it will override this important function

# matrix multiplication A'A
t(A) %*% A
# notice * by itself does element-wise multiplication

# matrix inverse inv(A'A)
solve(t(A) %*% A)
```

2.5 Working with Distributions in R

- The outcomes of a continuous random variable are defined by its probability density function (PDF) $f(x)$
- In many cases the PDF has a functional form
  - For example if $x \sim N(\mu, \sigma^2)$ then

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

- We usually visualize these by sketching them
- The area under the PDF is the cumulative distribution function (CDF) $Pr(x \leq a) = \int_{-\infty}^{a} f(x)dx$
  - The integral usually does not have a functional form and the computer solves it

2.6 Working with Probability Distributions

```
# if we need to evaluate the pdf we could calculate it
#   by hand, or we can use built-in functions

# For example if the random variable is normal
#   with mean 4 an variance 10, then
#   the PDF evaluated at 5
dnorm(5,mean=4,sd=sqrt(10))

# The CDF evaluated at 5
pnorm(5,mean=4,sd=sqrt(10))

# Sometimes we want to compute Pr(X ge 5)
1 - pnorm(5,mean=4,sd=sqrt(10))
pnorm(5,mean=4,sd=sqrt(10),lower.tail=FALSE)

# There are functions for the inverse CDF
qnorm(.1,mean=4,sd=sqrt(10))
```

2.7 Packages Will Be Important Later On

```
#Installing packages
install.packages("dplyr")
install.packages("readr")

#load the libraries each session
library(dplyr)
library(readr)
```

2.8 Working With Data

```
library(readr)
# Load in data from Greene's Webpage
#   This data on Monet Paintings sale price
# http://www.stern.nyu.edu/~wgreene/Text/Edition7/TableF4-1.csv
data = read_csv('http://www.stern.nyu.edu/~wgreene/Text/Edition7/TableF
```

```
## Rows: 430 Columns: 6
## -- Column specification --------------------------------------
## Delimiter: ","
## dbl (6): PRICE, HEIGHT, WIDTH, SIGNED, PICTURE, HOUSE
##
## i Use `spec()` to retrieve the full column specification for this da
## i Specify the column types or set `show_col_types = FALSE` to quiet
```

```
# you need to be able to do a couple of things with data frames
# (1) put data in matrices
# (2) add variables
# (3) select observations

# Make a matrix X that has price and whether it is signed as columns
X = cbind(data$PRICE,data$SIGNED)
# notice the dollar sign accesses the variables in 'data'

# Make a matrix X identaical to above, but also has a column of 1's
X = cbind(rep(1,nrow(data)),data$PRICE,data$SIGNED)
```

2.9 Adding Variables

```
suppressMessages(library(dplyr))
# dplyr loads with a bunch of messages
# data does not have area or aspect ratio

# two ways to add variables
# (1) the old-fashion way
data$AREA = data$WIDTH*data$HEIGHT
data$ASPECTRATIO = data$WIDTH/data$HEIGHT

# (2) dplyr mutate function
data = mutate(data,
              AREA = WIDTH*HEIGHT,
              ASPECTRATIO = WIDTH/HEIGHT)
# don't have to use $ or repeatedly write 'data'
```

2.10 Selecting Data

```
# create new data set that selects only rows
#   where the painting sold for more than $2 million

# two ways to add variables
# (1) the old-fashion way
data2 = data[data$PRICE>2,]

# (2) dplyr filter function
data2 = filter(data,PRICE>2)

# With both approaches we are creating logical variables
# to select the rows we want
```

Lecture 3

Prediction and Least Squares

3.1 Prediction Versus Causality

- What can we do with data?
  1) Use data to *predict* outcomes
  2) Use data to understand *causal* relationship
- Before any data analysis first consider what is the purpose
- Econometrics tends to center on causal effects
- These are *not* unrelated (double negative)
  - The process for prediction is mostly the same as the process for causal relationships
  - If we are trying to establish causal effects there is just a bunch of other stuff we have to worry about

3.2 Prediction

- Prediction is straightforward
- We have some (scalar) outcome y we would like to predict
- Usually also observe (vector) x to help form prediction
- Since x is a vector, denote x_i as some combination of elements of x.
  - For example if x as three elements then we define

$$x_i = \begin{bmatrix} x_{i1} & x_{i2} & x_{i3} \end{bmatrix}'$$

- For a given combination of x_i we seek a prediction for y_i

3.3 Best Predictors

- Best predictors have small prediction errors
- Let $\hat{y}_i = g(x_i)$ be some prediction/prediction function

$$y_i = \underbrace{\hat{y}_i}_{Prediction} + \underbrace{e_i}_{Error/Residual}$$

- We should choose $g(x)$ to minimize the variance (amount) of the prediction error $\text{Var}(e) = \text{E}(e^2) - \text{E}(e)^2$

$$\text{E}\left[(y - g(x))^2\right]$$

- What function, $g(x)$ should we choose to minimize this expression

3.4 Proof: Best Predictor ($g(x)$) is Conditional Expectation

- Decompose y: $y = m(x) + u$
- What is $\mathrm{E}(u|x)$?
  - $\mathrm{E}(y - m(x)|x) = \mathrm{E}(y|x) - \mathrm{E}(m(x)|x) = 0$

$$\begin{aligned}\mathrm{E}\left[(y - g(x))^2\right] &= \mathrm{E}\left[(m(x) + u - g(x))^2\right] \\ &= \mathrm{E}\left[u^2\right] + 2\,\mathrm{E}\left[u\,(m(x) - g(x))\right] + \mathrm{E}\left[(m(x) - g(x))^2\right] \\ &= \mathrm{E}\left[u^2\right] + \mathrm{E}\left[(m(x) - g(x))^2\right]\end{aligned}$$

Minimized when $E\left[(m(x) - g(x))^2\right] = 0$, so set $g(x) = m(x)$

- Law of iterated expectations $\mathrm{E}(v) = \mathrm{E}_x\left[\mathrm{E}(v|x)\right]$
- $\mathrm{E}\left[uh(x)\right] = \mathrm{E}_x\left[\mathrm{E}\left[uh(x)|x\right]\right] = \mathrm{E}_x\left[h(x)\,\mathrm{E}\left(u|x\right)\right]$

3.5 Prediction in Practice

- Suppose we observe y_i for $i = 1, 2, \ldots, n$
- and scalar x_i for $i = 1, 2, \ldots, n$
- We consider a simple linear prediction function $\hat{y}_i = g(x_i) = a + b \cdot x_i$
- Define the prediction error $e_i = y_i - (a + bx_i)$
- And choose a and b to minimize the sample variance of the prediction error

$$\widehat{\mathrm{Var}(e)} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Use data to choose parameters of prediction function to minimize the sum of squared errors (SSE) $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

3.6 Prediction in Practice

- The proof shows that the function that minimizes the prediction error is the *conditional expectation*
- Thus, if we take data and find parameters that minimize the sum of squared errors (SSE), the resulting output will be an estimate of the conditional expectation function from the observed data
  - For example if we choose a and b in $g(x) = a + b \cdot x$ to minimize SSE, then

$$g(x) = a + b \cdot x \approx \mathrm{E}(y|x)$$

3.7 Minimizing SSE (Prediction Error)

$$SSE = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Strategy: take two partial derivatives, set them equal to zero, and solving the system of equations

$$\frac{\partial SSE}{\partial a} = -2 \times \sum_{i=1}^{n} (y_i - a - bx_i)$$
$$\frac{\partial SSE}{\partial b} = -2 \times \sum_{i=1}^{n} x_i (y_i - a - bx_i)$$

3.8 Minimizing SSE

System of equations

$$0 = \sum_{i=1}^{n} (y_i - a - bx_i) \tag{1}$$
$$0 = \sum_{i=1}^{n} x_i (y_i - a - bx_i) \tag{2}$$

Re-writing Eq. (1)

$$\begin{aligned} n \times a &= \textstyle\sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i \\ a &= \tfrac{1}{n} \textstyle\sum_{i=1}^{n} y_i - b \tfrac{1}{n} \sum_{i=1}^{n} x_i \\ \boxed{a &= \overline{y} - b\overline{x}} \end{aligned} \tag{3}$$

3.9 Minimizing SSE

Plug Eq. (3) into Eq. (2) and solve for b

$$\begin{aligned} 0 &= \textstyle\sum_{i=1}^{n} x_i (y_i - a - bx_i) \\ 0 &= \textstyle\sum_{i=1}^{n} x_i (y_i - \overline{y} + b\overline{x} - bx_i) \\ 0 &= \textstyle\sum_{i=1}^{n} x_i [y_i - \overline{y} - b(x_i - \overline{x})] \\ 0 &= \textstyle\sum_{i=1}^{n} x_i (y_i - \overline{y}) - b \sum_{i=1}^{n} x_i (x_i - \overline{x}) \\ 0 &= \textstyle\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) - b \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x}) \\ \boxed{b &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})}} \end{aligned} \tag{4}$$

Written in terms of the sample covariance and variance

$$b = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

3.10 Summarizing

- For $g(x) = a + b \cdot x$
  - Where $a = \overline{y} - b\overline{x}$ and $b = \text{Cov}(x, y)/\text{Var}(x)$
  - Is referred to as the Best Linear Predictor (BLP) of y given x
- This approximates the conditional expectation
- Notice that this is the *actual conditional expectation if y and x are normally distributed
  - Recall $\text{E}(y|x) = \mu_y + \dfrac{\text{Cov}(y, x)}{\text{Var}(x)}(x - \mu_x)$
  - Rearrange $\text{E}(y|x) = \left[\mu_y - \dfrac{\text{Cov}(y, x)}{\text{Var}(x)}\mu_x\right] + \left(\dfrac{\text{Cov}(y, x)}{\text{Var}(x)}\right)(x)$

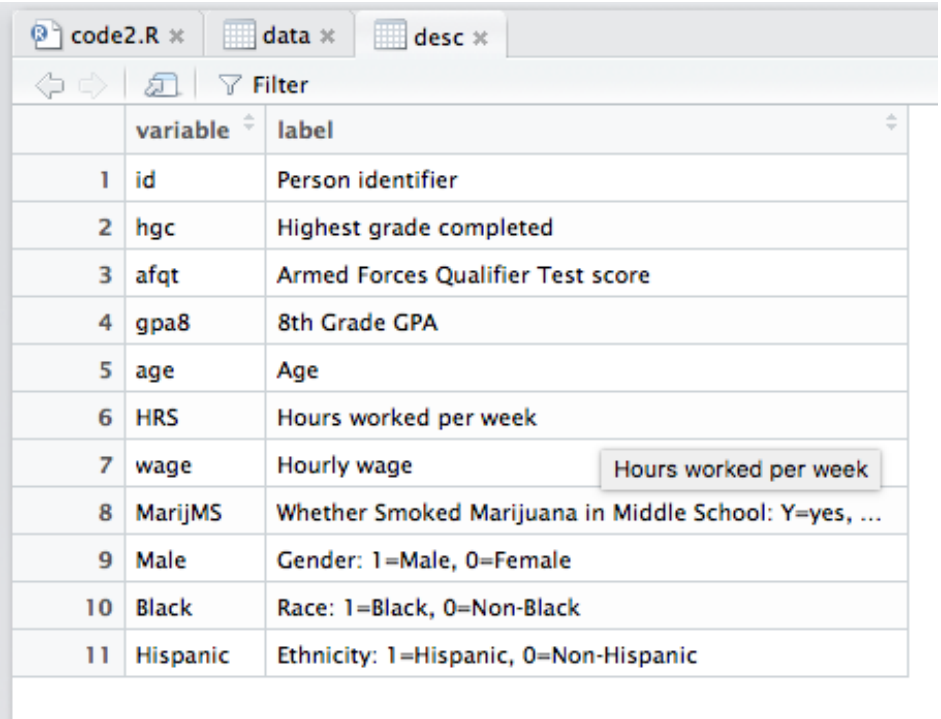
3.11 'NLSY97r13.RData

code2.R × data ×

Filter

	id	hgc	afqt	gpa8	age	HRS	wage	MarijMS	Male	Black	Hispanic
1	1	16	45.070	3.0	32	40	36.06	N	0	0	0
2	2	14	58.483	3.5	31	40	41.96	N	1	0	1
3	4	13	37.012	4.0	32	40	21.68	N	0	0	1
4	6	14	22.001	1.5	31	35	7.40	N	0	0	1
5	11	16	30.583	3.5	31	40	18.83	N	0	0	1
6	13	10	67.533	2.5	29	28	7.25	N	1	0	1
7	16	16	44.451	2.5	31	NA	NA	N	1	0	1
8	18	13	9.339	2.5	32	35	15.80	N	1	1	0
9	21	8	3.505	3.0	31	NA	NA	Y	1	0	1
10	22	13	10.443	1.5	31	40	46.49	N	1	0	1
11	23	15	6.423	2.0	30	35	30.11	N	0	0	1
12	24	8	6.120	1.0	29	9	50.00	N	1	0	1
13	25	8	5.403	2.5	30	NA	NA	N	0	0	1
14	26	11	12.866	3.0	33	NA	NA	N	1	1	0
15	28	17	17.058	3.5	30	40	15.00	N	0	1	0
16	31	16	62.806	3.0	31	40	32.56	N	1	0	0
17	32	18	39.949	4.0	32	40	67.00	N	0	0	0
18	33	16	96.655	4.0	32	15	8.33	N	0	0	0

3.12 'NLSY97r13.RData



	variable	label
1	id	Person identifier
2	hgc	Highest grade completed
3	afqt	Armed Forces Qualifier Test score
4	gpa8	8th Grade GPA
5	age	Age
6	HRS	Hours worked per week
7	wage	Hourly wage
8	MarijMS	Whether Smoked Marijuana in Middle School: Y=yes, ...
9	Male	Gender: 1=Male, 0=Female
10	Black	Race: 1=Black, 0=Non-Black
11	Hispanic	Ethnicity: 1=Hispanic, 0=Non-Hispanic

Hours worked per week

3.13 Predicting Wages with 8th Grade GPA

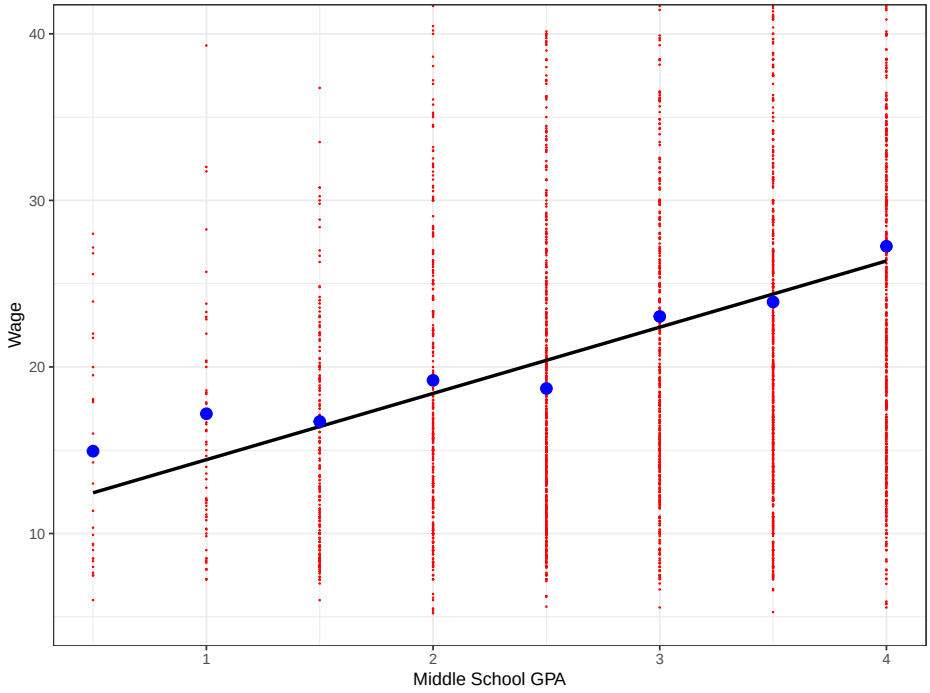
```
load(file.path(rdata_loc,'NLSY97r13.RData'))

# Keep only rows with valid wage
data = data[!is.na(data$wage),]

# Predicting wage with gpa8
# Slope Parameter
b_gpa = cov(data$wage,data$gpa8)/var(data$gpa8)
b_gpa
```

```
## [1] 3.977438
```

```
# Intercept Parameter
a_gpa = mean(data$wage) - b_gpa*mean(data$gpa8)
a_gpa
```

```
## [1] 10.45583
```

3.14 Prediction Line: $\widehat{wage} = 10.45 + 3.98gpa$



3.15 Summary

1) The best predictor of y given x is $\mathrm{E}(y|x)$ generalizes to when x_i is a vector
   - Best predictor of y given $x = \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{iK} \end{bmatrix}$ is $\mathrm{E}(y|x_{i1}, x_{i2}, \dots, x_{iK})$
2) Whenever we minimize the SSE we are finding parameters that summarizes all of the conditional expectations in the data in a convenient way
3) We considered linear prediction function $\widehat{y} = a + bx$
   - Minimizing prediction error with linear prediction function is called ordinary least squares (OLS)
   - Later extend OLS to multiple x's, i.e. $\widehat{y}_i = a + bx_{i1} + cx_{i2} + \cdots$

3.16 Goodness of Fit, R^2

- Characterize how well we are able to fit/predict the outcome variable
- R^2 says what fraction of the sample variation in y can be predicted (is explained) by x
- SST is sum of squared total, total variability in y
  - $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$
- SSE is sum of squared error, total variability in y after accounting for x
  - $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$$\boxed{R^2 = 1 - \frac{SSE}{SST}} \in [0, 1]$$

3.17 R Squared

```
# fitted wage for each observation
data$fittedwage = a_gpa + b_gpa*data$gpa8

# residual for each observation
data$residwage = data$wage - data$fittedwage

# Compute Rsquared
SST = sum( (data$wage - mean(data$wage))^2 )
SSE = sum( data$residwage^2 )
1 - SSE/SST
```

```
## [1] 0.0627139
```

Lecture 4

Inferring Causality From Covariability

4.1 Causality

- Consider the relationship: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- β_1 is a causal/partial/marginal effect of x_1

$$\begin{aligned}\frac{\partial y}{\partial x_1} =& \frac{\partial \beta_0}{\partial x_1} + \frac{\partial \beta_1 x_1}{\partial x_1} + \frac{\partial \beta_2 x_2}{\partial x_1} \\ =& 0 + \beta_1 \frac{\partial x_1}{\partial x_1} + 0 \\ =& \beta_1\end{aligned}$$

- Resources are scarce. Firms want to minimize costs, governments only have so much money
- Causal effects are of central importance in decision making

4.2 Deterministic Verus Stochastic Functions

- Consider the relationship: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
  - β_1 is a causal/partial/marginal effect of x_1
  - If x_1 and x_2 is observed then y is completely determined (it is deterministic)
  - Observing a handful (3) values of y, x_1, x_2 can determine β_0, β_1, β_2
- What if x_2 is not observed
  - Need to treat $\beta_2 \cdot x_2$ as a random variable
- Consider the stochastic function $y = \beta_0 + \beta_1 x_1 + \varepsilon$
  - Cannot determine β_1 only observing values of x_1 and y

4.3 Stochastic Functions Cont.

- A variable that is a function of (nearly) infinite explanatory variables

$$y = \beta_0 + \beta_1 x_1 + \underbrace{\beta_2 x_2 + \cdots + \beta_\infty x_\infty}_{\varepsilon}$$

- β_1 is a causal effect, effect of x_1 holding everything else fixed: $\partial y / \partial x_1$
- If only observe y and x_1 left with stochastic function

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- Under what conditions can we determine (estimate) β_1 only observing y and x_1?

4.4 Econometric Modeling

- Objective: Characterize the relationship between a variable and a set of 'related' variables
  - Specifically causal effects
- There are aspects of the relationship that are *not* observed
  - Treat these as random variables
- Definition: A model is a set of assumptions (restrictions) on the joint distribution of the variables (observed and unobserved)

4.5 Objective of Econometric Modeling

- Suppose $y = \beta_0 + \beta_1 x_1 + \varepsilon$ w/ β_1 the causal effect
- If we observe n observations on y and x_1, $\{y_i, x_{i1}\}_{i=1}^n$
  - We can estimate the prediction function $\widehat{y} = a + bx_1$
  - $b = \text{Cov}(y, x_1)/\text{Var}(x_1)$
- The purpose of the econometric model is to connect what we can calculate, b, with what we want β_1
- It lays out the conditions under which $b = \beta_1$
- It does not guarantee conditions will be met, but makes clear what is needed

4.6 Inferring Causality From Covariability

- Our econometric model is assumptions about $y = \beta_0 + \beta_1 x_1 + \varepsilon$
- We estimate the CEF $\text{E}(y|x_1) = a + bx_1$
- For $b = \beta_1$ we need to outline conditions in which the CEF is a linear function: $\text{E}(y|x_1) = \beta_0 + \beta_1 x_1$
- *Main Condition:* This will be true if ε is mean independent of x_1

$$\text{E}(\varepsilon|x_1) = \text{E}(\varepsilon) = 0$$

- If ε is mean independent of x_1, then $\text{E}(y|x_1)$

$$\begin{aligned}\text{E}(\beta_0 + \beta_1 x_1 + \varepsilon|x_1) =& \text{E}(\beta_0|x_1) + \text{E}(\beta_1 x_1|x_1) + \text{E}(\varepsilon|x_1)\\ =& \beta_0 + \beta_1 \text{E}(x_1|x_1) + \text{E}(\varepsilon|x_1)\\ =& \beta_0 + \beta_1 x_1\end{aligned}$$

4.7 Example: Effect of Education on Wages

- Wage Equation: $wage = \beta_0 + \beta_1 hgc + \varepsilon$
  - $wage$ is hourly wage, hgc is highest grade completed
  - ε is everything that determines wages *except* hgc
  - β_1 is the causal effect of education on wages, holds fixed ε
- With any dataset can estimate conditional expectation function (CEF), $\mathrm{E}(y|x)$
  - $\mathrm{E}(wage|hgc) = a + b \cdot hgc$
  - $b = \mathrm{Cov}(wage, hgc)/\mathrm{Var}(hgc)$
- When is $b = \beta_1$?

4.8 Example: Effect of Education on Wages

- Define $\mathrm{E}(\varepsilon|hgc) = \pi_0 + \pi_1 hgc$

$$\begin{aligned}\mathrm{E}(wage|hgc) =& \mathrm{E}(\beta_0 + \beta_1 hgc + \varepsilon|hgc)\\ =& \beta_0 + \beta_1 \mathrm{E}(hgc|hgc) + \mathrm{E}(\varepsilon|hgc)\\ =& \beta_0 + \beta_1 \mathrm{E}(hgc|hgc) + \pi_0 + \pi_1 hgc\\ =& (\beta_0 + \pi_0) + (\beta_1 + \pi_1) hgc\end{aligned}$$

- $b = (\beta_1 + \pi_1)$, If $\pi_1 = 0$ then success!
- $\pi_1 = 0 \Rightarrow \mathrm{E}(\varepsilon|hgc) = \mathrm{E}(\varepsilon)$ (Mean Independence)
  - Of all of the other things that influence wages except hgc, I cannot predict anything about these other things knowing hgc

4.9 Summarizing Econometric Modeling

- If we want to say something about underlying relationships between variables (causality) we need more than just data
- We need to understand the context and circumstance of the *observed* and *unobserved* data
- We formalize the context with a list of assumptions
- We never have a guarantee that an assumption is true (that's why it is called and assumption)
- Listing assumptions does not make them true
- We need to take proactive steps ensure the assumptions are *plausibly* true

Lecture 5

Introduction to Linear Regression Model

5.1 First Econometric Model: LRM

- Definition: A model is a set of assumptions (restrictions) on the joint distribution of the variables
- Introduce (Classical) Linear Regression Model (LRM)
  - The most widely applied econometric tool
  - Defined by 5 assumptions on the joint distribution of variables
- Describes the relationship between a dependent variable y and set of observed independent (explanatory) variables $x_1, x_2, \ldots, x_K$

$$y = f(x_1, x_2, x_3, \ldots, x_K) + \underbrace{\varepsilon}_{\text{Disturbance}}$$

- We observe y and x's, don't observe ε. Like to learn about $f(\cdot)$

5.2 Assumption 1: Linearity

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_K x_K + \varepsilon$$

- Usually set $x_1 = 1$ (constant term)
  - Constant allows us to explain overall levels
  - always include a constant unless good reason not to
- Linear in parameters
  - $y = \beta_1 f_1(\cdot) + \beta_2 f_2(\cdot) + \cdots + \varepsilon$
  - $f_k(\cdot)$ can be any function of $x_1, x_2, \ldots, x_K$
- Valid: $\ln(y) = \beta_1 + \beta_2 sine(x_1 + x_2) + \varepsilon$
- Invalid: $y = \beta_1 + x^{\beta_2} + \varepsilon$

5.3 Data Organization Notation

Data for individual i

- $x_i = \begin{bmatrix} 1 \\ x_{i2} \\ \vdots \\ x_{iK} \end{bmatrix}_{K \times 1}$ $y_i = x_i'\beta + \varepsilon_i$, where $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}_{K \times 1}$

Data for n observations

- $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}$ $X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}_{n\times K}$ $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$
- $Y = X\beta + \boldsymbol{\varepsilon}$
- Note: First column in X is all 1's

5.4 Assumption 2: Full Rank of X

- No exact *linear* relationship among any of the independent variable
- Necessary for identification
- Violation
  - $C_i = \beta_1 + \beta_2 TI_i + \beta_3 II_i + \beta_4 LI_i + \varepsilon_i$
  - C=Consumption, TI= Total Income, II= Investment Income, LI= Labor Income
  - $TI = II + LI$
  - Unique values for β_2, β_3, and β_4 do not exist
- Non-linear relationships are OK $C_i = \beta_1 + \beta_2 \ln(TI_i) + \beta_3 II_i + \beta_4 LI_i + \varepsilon_i$

5.5 Assumption 3: Mean Independence $\mathrm{E}(\boldsymbol{\varepsilon}|X) = \mathrm{E}(\boldsymbol{\varepsilon}) = 0$

- The x's (all of them) are not informative about the expected value of ε (any of them)
  - $wage_i = \beta_1 + \beta_2 hgc_i + \varepsilon_i$
  - ε_i includes everything else that determines i's wage except hgc
  - Knowing hgc_i does not give us any information about ε_i
- Not an assumption about $\mathrm{Var}(\varepsilon|X)$

5.6 Useful Results with $\mathrm{E}(\boldsymbol{\varepsilon}|X) = 0$

1) $\mathrm{E}(\varepsilon_i) = 0$, from Law of Iterated Expectations
   - $\mathrm{E}(\varepsilon_i) = \mathrm{E}_X\left[\mathrm{E}(\varepsilon_i|X)\right] = 0$
2) $\mathrm{E}(x_{ik}\varepsilon_j) = 0$ for any $i = 1, \ldots, n$ and $j = 1, \ldots, n$

$$\begin{aligned} \mathrm{E}(x_{ik}\varepsilon_j) &= \mathrm{E}_X\left[\mathrm{E}(x_{ik}\varepsilon_j|X)\right] \\ &= \mathrm{E}_X\left[x_{ik}\,\mathrm{E}(\varepsilon_j|X)\right] \\ &= \mathrm{E}_X\left[x_{ik}\cdot 0\right] = 0 \end{aligned}$$

3) $\mathrm{Cov}(x_{ik}, \varepsilon_j) = 0$, from (1) and (2) above
   - $\mathrm{Cov}(x_{ik}, \varepsilon_j) = \mathrm{E}(x_{ik}\varepsilon_j) - \mathrm{E}(x_{ik})\,\mathrm{E}(\varepsilon_j) = 0$

- Mean independence implies Cov = 0. Other way is not necessarily true

5.7 Implications for Assumptions 1 and 3

Linear in parameters: $y = x'\beta + \varepsilon$ and $\mathrm{E}(\varepsilon|X) = 0$

- Linear CEF: $\mathrm{E}(y|x) = x'\beta$

$$\begin{aligned}\mathrm{E}(y|x) =& \mathrm{E}(x'\beta + \varepsilon|x)\\ =& \mathrm{E}(x'\beta|x) + \mathrm{E}(\varepsilon|x)\\ =& x'\beta\end{aligned}$$

  - Definition of Regression is conditional expectation

- Causality: parameters of the CEF,β_k, are causal effect

$$\begin{aligned}\frac{\partial \mathrm{E}(y|x)}{\partial x_k} =& \frac{\partial x'\beta}{\partial x_k} + \frac{\partial \mathrm{E}(\varepsilon|x)}{\partial x_k}\\ =& \beta_k + 0\end{aligned}$$

5.8 Assumption 4: Homosekedasticty and Non-Autocorrelation

Homoskedasticty $\mathrm{Var}(\varepsilon|X) = \mathrm{Var}(\varepsilon) = \sigma^2$

- Variance of the disturbance is independent of X
- Heteroskedasticity $\mathrm{Var}(\varepsilon|x_i) = \sigma_i^2$
- Example: $wage_i = \beta_0 + \beta_{1i} hgc_i + \varepsilon_i$
  - Random coefficient $\beta_{1i} = \beta_1 + r_i$
  - DGP: $wage_i = \beta_0 + \beta_1 hgc_i + \widetilde{\varepsilon}_i$
    - where $\widetilde{\varepsilon}_i = r_i hgc_i + \varepsilon_i$
    - $\mathrm{Var}(\widetilde{\varepsilon}_i|hgc_i) = hgc_i^2 \mathrm{Var}(r_i) + \sigma^2$

5.9 Assumption 4: Homosekedasticty and Non-Autocorrelation

Non-autocorrelation $\mathrm{Cov}(\varepsilon_i, \varepsilon_j|X) = 0$

- No common disturbances amongst observations
- Auto-correlation

$$\begin{aligned}Profits_i =& \beta_1 + \beta_2 FirmSize_i + \varepsilon_i\\ Profits_j =& \beta_1 + \beta_2 FirmSize_j + \varepsilon_j\end{aligned}$$

  - If firm i and j are in same city then deviations of profits from the expected value are likely correlated

5.10 Implications for Assumption 4

- $\text{Var}(\boldsymbol{\varepsilon}|X) = \text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|X) - \text{E}(\boldsymbol{\varepsilon}|X)\,\text{E}(\boldsymbol{\varepsilon}|X)' = \sigma^2 I_{n\times n}$
- $\text{E}(\boldsymbol{\varepsilon}|X) = 0$

$$\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|X) = \begin{bmatrix} \text{E}(\varepsilon_1^2|X) & \text{E}(\varepsilon_1\varepsilon_2|X) & \cdots & \text{E}(\varepsilon_1\varepsilon_n|X) \\ \text{E}(\varepsilon_2\varepsilon_1|X) & \text{E}(\varepsilon_2^2|X) & \cdots & \text{E}(\varepsilon_2\varepsilon_n|X) \\ \vdots & \vdots & \ddots & \vdots \\ \text{E}(\varepsilon_n\varepsilon_1|X) & \text{E}(\varepsilon_n\varepsilon_2|X) & \cdots & \text{E}(\varepsilon_n^2|X) \end{bmatrix}$$
$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I_{n\times n}$$

- $\text{Var}(\boldsymbol{\varepsilon}) = \text{E}\left[\text{Var}(\boldsymbol{\varepsilon}|X)\right] + \text{Var}\left[\text{E}(\boldsymbol{\varepsilon}|X)\right] = \sigma^2 I_{n\times n}$

5.11 Summarizing Assumptions of LRM

- Assumptions Required to Estimate β
  1) Linearity: $Y = X\beta + \boldsymbol{\varepsilon}$
  2) Full rank of X
  3) Mean independence: $\text{E}(\boldsymbol{\varepsilon}|X) = 0$
- Assumptions Required for Model Inferenece
  4) Homosekedasticty and Non-Autocorrelation: $\text{Var}(\boldsymbol{\varepsilon}|X) = \sigma^2 I_{n\times n}$
  5) Normality of disturbances: $\boldsymbol{\varepsilon}|X \sim N(0, \sigma^2 I_{n\times n})$
     - Replace by central limit theorem later

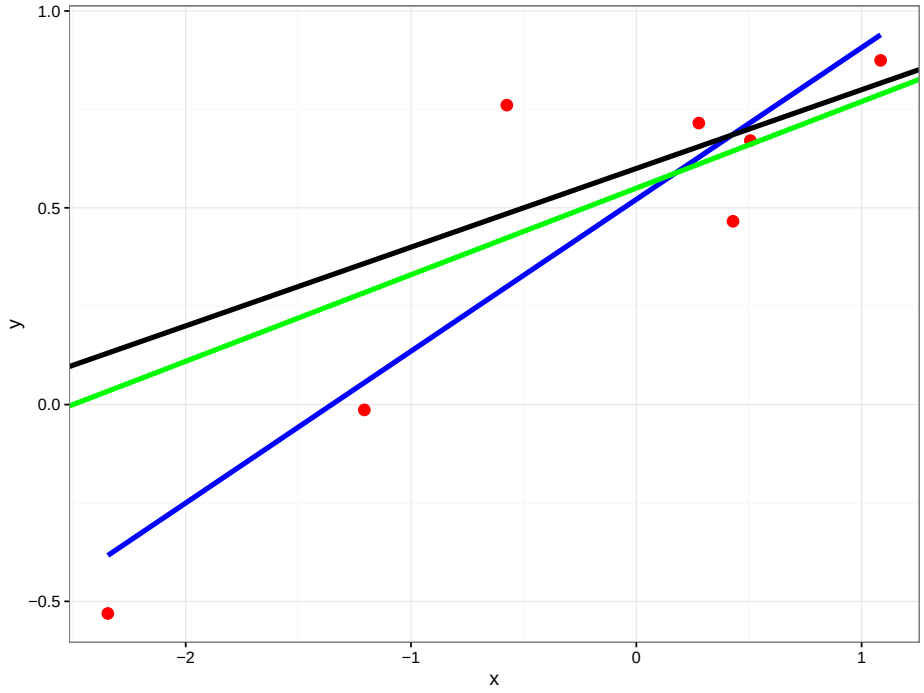
LECTURE 6

Estimating the Linear Regression Model

6.1 Finding b, an estimate of β in LRM

- Vocabulary:
  - Population regression (CEF): $\mathrm{E}(y|x) = x'\beta$
  - Estimated regression (CEF): $\mathrm{E}(y|x) = x'b$
  - Predicted value: $\hat{y} = x'b$
  - Disturbance: $\varepsilon = y - x'\beta$
  - Residual: $e = y - x'b$
- Identity: $y = x'b + e = x'\beta + \varepsilon$
- A good estimate is not $y = x'b$, but $\boxed{e = \varepsilon}$
  - Recall if ε is known then we have deterministic function and β is easy to get
  - if $e = \varepsilon$ then $b = \beta$

6.2 Black Line Is Population, Which b Better? Blue or Green?



6.3 Finding b, an estimate of β in LRM

- Assumptions Required to Estimate β

  1) Linearity: $Y = X\beta + \boldsymbol{\varepsilon}$

2) Full rank of X
3) Mean independence: $\mathrm{E}(\boldsymbol{\varepsilon}|X) = 0$

- Least Squares. Minimize $SSE = \sum_{i=1}^{n}(y_i - x_i'b)^2$
  - Assumptions 1-3 imply $\mathrm{E}(Y|X)$ is of the form $X\beta$
  - Minimizing the prediction error produces conditional expectation functions of the form Xb
- All approaches to estimation lead to the same estimates: MLE, MOMS, OLS

6.4 Estimating β with OLS

- Minimize the SSE, $b = \text{argmin} \sum_{i=1}^{n}(y_i - x_i'b)^2$
- Strategy: take K partial derivatives, set them equal to zero, and solve the system of equations

$$\begin{aligned}
\frac{\partial SSE}{\partial b_1} &= -2 \times \sum_{i=1}^{n} x_{i1}(y_i - x_i'b) = 0 \\
\frac{\partial SSE}{\partial b_2} &= -2 \times \sum_{i=1}^{n} x_{i2}(y_i - x_i'b) = 0 \\
&\vdots \\
\frac{\partial SSE}{\partial b_K} &= -2 \times \sum_{i=1}^{n} x_{iK}(y_i - x_i'b) = 0
\end{aligned}$$

6.5 Least Squares Normal Equations

- System of equations: $X'_{n\times K}(Y - Xb)_{n\times 1} = \mathbf{0}_{K\times 1}$

$$\begin{aligned}
\mathbf{0} &= X'(Y - Xb) \\
&= X'Y - X'Xb \\
X'Xb &= X'Y \\
(X'X)^{-1}(X'X)b &= (X'X)^{-1}X'Y \\
I_{K\times K}b &= (X'X)^{-1}X'Y \\
\boxed{b = (X'X)^{-1}X'Y}
\end{aligned}$$

- Assumption 2 (X is full rank) implies $(X'X)$ is invertible

6.6 Estimating LRM in R

```
load(file.path(rdata_loc,'NLSY97r13.RData'))

# Keep only rows with valid wage
data = data[!is.na(data$wage),]

n = nrow(data)
```

6.7 Simple Linear Regression

```
# linear model wage = b1 + b2 gpa
X = cbind(rep(1,n),data$gpa)
Y = data$wage
b = solve(t(X) %*% X) %*% (t(X)%*% Y)
b
```

```
##           [,1]
## [1,] 10.455831
## [2,]  3.977438
```

```
#predicted wage for 3.0 GPA
cbind(1,3)%*%b
```

```
##          [,1]
## [1,] 22.38814
```

6.8 Multiple Linear Regression

```
# linear model wage = b1 + b2 gpa + b3 hgc + b4 age
X = cbind(rep(1,n),data$gpa,data$hgc,data$age)
Y = data$wage
b = solve(t(X) %*% X) %*% (t(X)%*% Y)
b
```

```
##             [,1]
## [1,] -18.6565798
## [2,]   1.8852002
## [3,]   1.1273001
## [4,]   0.6119431
```

```
#predicted wage for 3.0 GPA, 12 hgc, 30 age
cbind(1,3,12,30)%*%b
```

```
##          [,1]
## [1,] 18.88491
```

6.9 Built in OLS in R: lm Function

```
# lm is missing the 'R', lm=linear model
lm(wage ~ gpa8 + hgc + age, data=data)
```

```
##
## Call:
## lm(formula = wage ~ gpa8 + hgc + age, data = data)
##
## Coefficients:
## (Intercept)         gpa8          hgc          age
##    -18.6566       1.8852       1.1273       0.6119
```

6.10 lm Output

```
# more information than just the estimates
mod = lm(wage ~ gpa8 + hgc + age, data=data)
attributes(mod)
```

```
## $names
##  [1] "coefficients"  "residuals"     "effects"
##  [4] "rank"          "fitted.values" "assign"
##  [7] "qr"            "df.residual"   "xlevels"
## [10] "call"          "terms"         "model"
##
## $class
## [1] "lm"
```

```
nobs(mod)
```

```
## [1] 3104
```

```
mod$coefficients
```

```
## (Intercept)        gpa8         hgc         age
## -18.6565798   1.8852002   1.1273001   0.6119431
```

```
mod$call
```

```
## lm(formula = wage ~ gpa8 + hgc + age, data = data)
```

```
sum(mod$residuals)
```

```
## [1] 2.470135e-12
```

6.11 Decomposing Variation

- SST $= \sum_{i=1}^{n}(y_i - \overline{y})^2$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i + e_i - \overline{y})^2$$
$$= \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}_{\text{sum of squares regression (SSR)}} + \underbrace{2\sum_{i=1}^{n}(\hat{y}_i - \overline{y})e_i}_{0} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{sum of squared errors (SSE)}}$$

- $\sum_{i=1}^{n}(b'x_i - b'\overline{x})e_i = b'\sum_{i=1}^{n} x_i e_i - b'\overline{x}\sum_{i=1}^{n} e_i = 0$
- SST = SSR + SSE

6.12 Model Fit: R-squared

- SST = SSR + SSE
- R^2 says what fraction of the sample variation in y is explained by x

$$\boxed{R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}} \in [0, 1]$$

- Is large R^2 always good?

  - R^2 is only comparable when SST is same
  - Adding additional regressors never degrades fit
  - Dropping observations can increase R^2!!!
    * Overfitting. $R^2 = 1$ if $n = K$

6.13 The Adjusted-R^2

- Trade-off of better model fit and loss in degrees of freedom when variables are added
- $R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$

$$\boxed{\text{adj-}R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2/(n-K)}{\sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)}}$$

- Discussion
  - Adjusted-R^2 can be negative
  - If adding a regressor increases Adjusted-R^2, then good reason to keep
  - There are other measures of fit that impose larger penalty for lossed degrees of freedom

6.14 R-Squared and adjusted-R-Squared

```
summary(mod)$r.squared
```

```
## [1] 0.1070757
```

```
summary(mod)$adj.r.squared
```

```
## [1] 0.1062115
```

Lecture 7

Logarithms in the Linear Regression Model

7.1 Using Logs in the Regression Function

- Quick review
- 'ln' is the natural log. Differs from log_{10} log base 10
- $\ln(\exp(x)) = x$
  - exp is the exponential function
- $d\ln(x)/dx = 1/x$
- $d\exp(x)/dx = \exp(x)$

7.2 Using Logs in the Regression Function

$$\boxed{y = \exp(\beta_1 + \beta_2 x + \varepsilon)}$$

$$\boxed{\ln(y) = \beta_1 + \beta_2 x + \varepsilon}$$

- What is meaning of β_2?
- $\partial y/\partial x = \exp(\beta_1 + \beta_2 x + \varepsilon) \times \beta_2$
  - Does this simplify?
  - $\partial y/\partial x = y \times \beta_2 \Rightarrow (\partial y/y)/\partial x = \beta_2$
  - $\%\Delta y/\Delta x \approx \beta_2$

7.3 Log-Wage Equation

```
load(file.path(rdata_loc,'NLSY97r13.RData'))

# Keep only rows with valid wage
data = data[!is.na(data$wage),]

data$MariJ = +(data$MarijMS=='Y')
```

7.4 Log-Wage Equation Model 1

```
## ----  Sec02
# short, ln(wage) = b1 + b2 age + b3 HRS + b4 Male + b5 MariJ
mod_short = lm(log(wage)~age+HRS+Male+MariJ,data=data)
round(mod_short$coefficients,3)
```

```
## (Intercept)         age         HRS        Male       MariJ
##       2.007       0.024       0.004       0.104      -0.095
```

7.5 Log-Wage Equation Model 2

```
# long, ln(wage) = b1 + b2 age + b3 HRS + b4 Male + b5 MariJ + b6 HGC
mod_long = lm(log(wage)~age+HRS+Male+MariJ+hgc,data=data)
round(mod_long$coefficients,3)
```

```
## (Intercept)         age         HRS        Male       MariJ
##       1.074       0.022       0.003       0.174       0.008
##         hgc
##       0.069
```

7.6 Summary of Regressions with Logs

- $y = \beta_1 + \beta_2 x$: β_2 is the level-change in y given level-change in x (level-level function)
- Functions with natural logs tell us about (approximate) percentage changes
  - $\ln(y) = \beta_1 + \beta_2 x$: $100 \times \beta_2$ is the percent change in y given a 1 unit change in x
  - $y = \beta_1 + \beta_2 \ln(x)$: $\beta_2/100$ is the level change in y given a 1 percent change in x
  - $\ln(y) = \beta_1 + \beta_2 \ln(x)$: β_2 is the percent change in y given a 1 percent change in x
- Use logs when want effects as percent, or data is heavily skewed (but must be positive)

LECTURE 8

Linear Regression Model Algebra

8.1 I: OLS is a Linear Function of Random Variables

- $b = (X'X)^{-1}X'Y$ is our estimate of $Y = X\beta + \boldsymbol{\varepsilon}$
- Plugging in an re-arranging

$$\begin{aligned} b =& (X'X)^{-1}X'Y \\ =& (X'X)^{-1}X'(X\beta + \boldsymbol{\varepsilon}) \\ =& (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\boldsymbol{\varepsilon} \\ =& \beta + (X'X)^{-1}X'\boldsymbol{\varepsilon} \end{aligned}$$

8.2 II: Projection Matrix: Predicted Values

- Predicted values: $\hat{Y} = Xb = \underbrace{X(X'X)^{-1}X'}_{\text{Projection Matrix } \mathbf{P}_X} Y$
  - Projection matrix $\mathbf{P}_X$ creates predicted values of Y with linear combination of columns in X
  - $\mathbf{P}_X X = X$, can predict X perfectly with X

8.3 III: Residual Maker: Residuals

- Residuals: $e = Y - \hat{Y} = Y - \mathbf{P}_X Y = \underbrace{(I_{n\times n} - \mathbf{P}_X)}_{\text{Residual Maker } \mathbf{M}_X} Y$
  - Residual maker $\mathbf{M}_X$ creates component of Y that cannot be predicted by linear combination of X
  - $\mathbf{M}_X X = 0$, nothing left when predicting X with X

8.4 How Does Least Squares Work

- The effects identified with least squares are interpreted as "holding the other observed variables fixed"
  - How does this work?
- What Determines Value of Single Coefficient

$$Y = X\beta + z\gamma + \boldsymbol{\varepsilon}$$

- γ is the partial effect of z on Y holding fixed X
- Normally include z in X, but want to see how $\hat{\gamma}$ (estimate of γ) is determined

8.5 What Determines Value of Single Coefficient

$$Y = X\beta + z\gamma + \boldsymbol{\varepsilon}$$

- Multiple both sides by residual maker $\mathbf{M}_X$

$$\begin{aligned}\mathbf{M}_X Y =& \mathbf{M}_X X\beta + \mathbf{M}_X z\gamma + \mathbf{M}_X \boldsymbol{\varepsilon} \\ Y^* =& z^*\gamma + \boldsymbol{\varepsilon}\end{aligned}$$

- $Y^* = \mathbf{M}_X Y$ residual of Y not explained by linear combination of X
- $z^* = \mathbf{M}_X z$ residuals of z not explained by linear combination of X
- This is called partial-ling out

8.6 Partial-ling Out

- From $Y = X\beta + z\gamma + \boldsymbol{\varepsilon}$
- Partial-out X: $Y^* = z^*\gamma + \boldsymbol{\varepsilon}$
- $\hat{\gamma}$ is estimated by minimizing SSE

$$\hat{\gamma} = \left[(z^*)'(z^*)\right]^{-1}\left[(z^*)'Y^*\right]$$

- Residual on residual regression, Greene p. 34
- If X and z are related, then partial effects of z are found by

  1) purging the effect of X from the variation in Y and z (holding fixed X)
  2) determining the effect of the residual variation in z on the residual variation in Y

LECTURE 9

Properties of Least Squares I: Is b a good estimate of β

9.1 Properties of Estimators

- $b = (X'X)^{-1}X'Y$ is our estimate of $Y = X\beta + \boldsymbol{\varepsilon}$
- b is a linear combination of a vector of random variables

$$\begin{aligned} b =& (X'X)^{-1}X'Y \\ =& (X'X)^{-1}X'(X\beta + \boldsymbol{\varepsilon}) \\ =& (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\boldsymbol{\varepsilon} \\ =& \underbrace{\beta}_{\substack{unknown \\ parameters}} + \underbrace{(X'X)^{-1}X'\boldsymbol{\varepsilon}}_{\substack{random \\ variable}} \end{aligned}$$

- b will never exactly equal β
- From different samples, we will get different b's

9.2 Properties of Estimators

- All estimators should be viewed as a random variable because they are a function of random data
- In OLS $b = \underbrace{\beta}_{\substack{unknown \\ parameters}} + \underbrace{(X'X)^{-1}X'\boldsymbol{\varepsilon}}_{\substack{random \\ variable}}$
- To understand if b is a good estimate of β depends on the properties of the random variable $(X'X)^{-1}X'\boldsymbol{\varepsilon}$. This is the estimators *sampling properties*
- Properties of great estimators

  1) Unbiassed: $\mathrm{E}(b) = \beta$, (show $\mathrm{E}((X'X)^{-1}X'\boldsymbol{\varepsilon}) = 0$)
  2) Efficient: $\mathrm{Var}(b)$ AS(mall)AP
  3) Normally distributed

9.3 How does b relate to β

- $b = (X'X)^{-1}X'Y$ is our estimate of $Y = X\beta + \boldsymbol{\varepsilon}$
  - b is vector of coefficients if from trying to predict Y with X
- The OLS estimate is $b = \beta + (X'X)^{-1}X'\boldsymbol{\varepsilon}$
- $(X'X)^{-1}X'\boldsymbol{\varepsilon}$ is vector of coefficients if tried to predict $\boldsymbol{\varepsilon}$ with X
- In reality $(X'X)^{-1}X'\boldsymbol{\varepsilon}$ will *NEVER* be zero
- If mean independence holds, $(X'X)^{-1}X'\boldsymbol{\varepsilon}$ will be zero in expectation
- In this case, we call $(X'X)^{-1}X'\boldsymbol{\varepsilon}$ *sampling error* so

$$b = \beta + \text{sampling error}$$

9.4 Failure of Mean Independence: Bias

- If mean independence does *NOT* hold $(X'X)^{-1}X'\boldsymbol{\varepsilon}$ represents bias (and sampling error)
- Our main concern is the bias
- This means in expectation $(X'X)^{-1}X'\boldsymbol{\varepsilon}$ is NOT zero in expectation
  - We can predict $\boldsymbol{\varepsilon}$ with X
- In this case, we call $(X'X)^{-1}X'\boldsymbol{\varepsilon}$ *bias* so

$$b = \beta + \text{bias}$$

9.5 Under Mean Independence, b is an Unbiased Estimate of β, .i.e., $\mathrm{E}(b) = \beta$

- Note: b is a random variables, β is a constant

$$b|X = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \boldsymbol{\varepsilon})$$

$$\begin{aligned}
\mathrm{E}(b|X) &= \mathrm{E}\left[(X'X)^{-1}X'(X\beta + \boldsymbol{\varepsilon})|X\right] \\
&= \mathrm{E}\left[(X'X)^{-1}X'X\beta|X\right] + \mathrm{E}\left[(X'X)^{-1}X'\boldsymbol{\varepsilon})|X\right] \\
&= \mathrm{E}\left[\beta|X\right] + (X'X)^{-1}X'\,\mathrm{E}\left[\boldsymbol{\varepsilon}|X\right] \\
&= \beta \qquad \boxed{\mathrm{E}\left[\boldsymbol{\varepsilon}|X\right] = 0 \text{ by ASM 3}}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(b) &= \mathrm{E}_X\left[\mathrm{E}(b|X)\right] \\
&= \mathrm{E}_X(\beta) \\
&= \beta \qquad \boxed{b \text{ is unbiased}}
\end{aligned}$$

9.6 Key Ingredient for OLS's Unbiasedness: Mean Independence

- Only ASM1-3 are needed for unbias (this is very important)
- Without mean independence (FAIL!)
  - OLS produces biased estimates
  - Only best linear predictor (BLP) not causal effect
- The most common critique of any econometric analysis is BIAS. Sources of failure of mean independence
  - Omitting important variables
  - Endogeneity: $\boldsymbol{\varepsilon}$ (jointly) determines X
  - Measurement error in X

Lecture 10

Bias From Violating Mean Independence

10.1 Multiple Sources of Violating Mean Independence

- LRM: $Y = X\beta + \boldsymbol{\varepsilon}$
- Sources of Mean Independence Violation
  1) Omitting important variables
     - Some variable Z is in $\boldsymbol{\varepsilon}$ that we can predict with X
  2) Endogeneity
     - X takes values that depend on $\boldsymbol{\varepsilon}$
  3) Measurement error in X
     - Mechanical relationship between $\boldsymbol{\varepsilon}$ and X

10.2 Bias From Omitting Important Variables

- Suppose $Y = X_1\beta_1 + X_2\beta_2 + \boldsymbol{\varepsilon}^*$, where $\mathrm{E}(\varepsilon_i^*|X_1, X_2) = 0$
- What are the consequences of omitting X_2?
  - If $\beta_2 = 0$ then nothing, X_2 does not belong
- What if $\beta_2 \neq 0$, does omitting X_2 create problems for estimating β_1?
  - Maybe, depends on the relationship between X_2 and X_1

10.3 Omitting X_2

- Define $\mathbf{P}_{X_1}$ as the projection matrix for X_1
- $\widehat{X}_2 = \mathbf{P}_{X_1}X_2$ component of X_2 that can be predicted by X_1
- Then X_2 can be decomposed as

$$X_2 = \underbrace{\widehat{X}_2}_{\substack{\text{can be}\\ \text{predicted}\\ \text{by } X_1}} + \underbrace{X_2^*}_{\substack{\text{can NOT be}\\ \text{predicted}\\ \text{by } X_1}}$$

- Plugging in

$$Y = X_1\beta_1 + (\widehat{X}_2 + X_2^*)\beta_2 + \boldsymbol{\varepsilon}^*$$

10.4 A New ERROR

- OLS will estimate parameters where the ERROR cannot be predicted by X
- Consider $Y = X_1\beta_1 + \boldsymbol{\varepsilon}$
  - Is β_1 the same β_1?

$$
\begin{aligned}
Y =& X_1\beta_1 + (\widehat{X}_2 + X_2^*)\beta_2 + \boldsymbol{\varepsilon}^* \\
=& X_1\beta_1 + (X_1(X_1'X_1)^{-1}X_1'X_2 + X_2^*)\beta_2 + \boldsymbol{\varepsilon}^* \\
=& \left[X_1\beta_1 + X_1(X_1'X_1)^{-1}X_1'X_2\beta_2\right] + X_2^*\beta_2 + \boldsymbol{\varepsilon}^* \\
=& X_1 \underbrace{\left[\beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2\right]}_{\substack{\text{NOT } \beta_1 \\ \text{This is } \beta_1 \text{ plus BIAS}}} + \underbrace{X_2^*\beta_2 + \boldsymbol{\varepsilon}^*}_{\substack{=\boldsymbol{\varepsilon} \\ \text{can NOT be} \\ \text{predicted} \\ \text{by } X_1}}
\end{aligned}
$$

10.5 BIAS From Omitting X_2

- $$Y = X_1\left[\beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2\right] + \boldsymbol{\varepsilon}$$
- β_1 is the partial effect of X_1 on Y holding fixed X_2
- β_2 is the partial effect of X_2 on Y holding fixed X_1
- What is $(X_1'X_1)^{-1}X_1'X_2$?
  - Suppose we write $\widehat{X}_2 = X_1\hat{\pi}$
  - $\hat{\pi}$ is the coefficients that we multiply X_1 to predict X_2
  - $\hat{\pi} = (X_1'X_1)^{-1}X_1'X_2$

10.6 Bias From Omitting Important Variables

$$
Y = X_1\beta_1 + \underbrace{(X_1\pi + X_2^*)}_{X_2}\beta_2 + \varepsilon^*
$$

$$
\begin{aligned}
b_1 =& (X_1'X_1)^{-1}(X_1'Y) \\
=& (X_1'X_1)^{-1}(X_1'(X_1\beta_1 + (X_1\pi + X_2^*)\beta_2 + \varepsilon^*)) \\
=& (X_1'X_1)^{-1}(X_1'X_1\beta_1 + X_1'X_1\pi\beta_2 + X_1'(X_2^*\beta_2 + \varepsilon^*)) \\
=& (X_1'X_1)^{-1}X_1'X_1\beta + (X_1'X_1)^{-1}X_1'X_1\pi\beta_2 \\
& \quad + (X_1'X_1)^{-1}X_1'(X_2^*\beta_2 + \varepsilon^*) \\
=& \beta_1 + \underbrace{\pi\beta_2}_{\text{bias}} + \underbrace{(X_1'X_1)^{-1}X_1'(X_2^*\beta_2 + \varepsilon^*)}_{\text{Sampling Error}}
\end{aligned}
$$

$$
\mathrm{E}(b_1|X_1) = \beta + \underbrace{\pi}_{\substack{\text{Coefficients} \\ \text{of regression of all} \\ \text{variables in } X_1 \text{ on } X_2}} \overbrace{\beta_2}^{\substack{\text{Effect of} \\ X_2 \text{ on } Y}}
$$

10.7 Uses of the Bias Equation

This formula is very useful $\mathrm{E}(b_1|X_1) = \beta_1 + \pi\beta_2$

1) It tells us exactly how parameters will change when we include or exclude variables
   - Define long regression $\hat{y}^{(l)} = b_1^{(l)} + b_2^{(l)}x_2 + \cdots + b_K^{(l)}x_K + \hat{\gamma}z$
   - Intermediate regression $\hat{z} = \hat{\pi}_1 + \hat{\pi}_2 x_2 + \cdots + \hat{\pi}_K x_K$
   - Short regression (excludes z) $\hat{y}^{s} = b_1^{(s)} + b_2^{(s)}x_2 + \cdots + b_K^{(s)}x_K$
   - $b_k^{(s)} - b_k^{(l)} = \hat{\pi}_k\hat{\gamma}$ EXACTLY!
2) We can sign the bias even with observing z

10.8 Understanding Why Parameters Change

```
load(file.path(rdata_loc,'NLSY97r13.RData'))

# Keep only rows with valid wage
data = data[!is.na(data$wage),]

data$MariJ = +(data$MarijMS=='Y')
```

10.9

```
## ----  Sec02
# short, ln(wage) = b1 + b2 age + b3 HRS + b4 Male + b5 MariJ
mod_short = lm(log(wage)~age+HRS+Male+MariJ,data=data)
round(mod_short$coefficients,3)
```

```
## (Intercept)         age         HRS        Male       MariJ
##       2.007       0.024       0.004       0.104      -0.095
```

10.10

```
# long, ln(wage) = b1 + b2 age + b3 HRS + b4 Male + b5 MariJ + b6 HGC
mod_long = lm(log(wage)~age+HRS+Male+MariJ+hgc,data=data)
round(mod_long$coefficients,3)
```

```
## (Intercept)         age         HRS        Male       MariJ
##       1.074       0.022       0.003       0.174       0.008
##         hgc
##       0.069
```

```
mod_short$coefficients["MariJ"] - mod_long$coefficients["MariJ"]
```

```
##      MariJ
## -0.1031756
```

10.11

```
# intermediate, hgc = p1 + p2 age + p3 HRS + p4 Male + p5 MariJ
mod_inter = lm(hgc~age+HRS+Male+MariJ,data=data)
round(mod_inter$coefficients,3)
```

```
## (Intercept)         age         HRS        Male       MariJ
##      13.532       0.021       0.027      -1.007      -1.497
```

```
mod_inter$coefficients["MariJ"]*mod_long$coefficients["hgc"]
```

```
##      MariJ
## -0.1031756
```

10.12 Signing the Bias

- Using omitted variable bias formula to predict how coefficient would change for a control we don't actually have
- Example, return to attending private school.
  - $\ln(wage_i) = \beta_1 + \beta_2 a_i + \varepsilon_i$
  - a_i is 1 if i attended private school
  - Estimates $\widehat{\ln(wage)} = b_1 + b_2 \cdot a_i \quad ; \quad b_2 = 0.13$
- Is $\beta_2 \lessgtr b_2 = 0.13$

10.13 Summarizing

- If our model satisfies mean independence then OLS is an unbiased estimator
- If our model does *not* satisfy mean independence then OLS is biased
  - Bias occurs because we are not holding fixed relevant variables
  - The effects we are identifying are not causal
- We can correct the bias by including relevant controls, often we don't have the data we would like
  - Signing the bias provides some insight

LECTURE 11

Properties of Least Squares: Is b a precise estimate of β

11.1 Properties of Estimators

- Unbiased $\mathrm{E}(b) = \beta$, if $\mathrm{E}(\boldsymbol{\varepsilon}|X) = 0$ is a useful result
  - When we produce estimate from data in expectation we get the right answer
  - Does not tell us how close we are to right answer for any given sample (precision)
- Next we look at the variance of b
  - $\mathrm{Var}(b)$ indicates the precision of our estimate
  - Does the data give us a clear answer?
  - Consistent as $n \to \infty$ then $\mathrm{Var}(b) \to 0$ thus $b \to \beta$

11.2 Variances of Estimators

- Our model has many parameters
  - A vector of random variables
- When referring to variance of a vector of random variables really are talking about variance-covariance matrix
  - e.g., $\mathrm{Var}(b_1)$, but also $\mathrm{Cov}(b_1, b_2)$
- With K parameters we have $K \times K$ covariance matrix
  - Diagonal elements are the variances and off-diagonal are covariances
  - Covariance matrix is symmetric

11.3 The OLS Estimator

The true value plus linear combination of a random variable

$$\begin{aligned} b =& (X'X)^{-1}X'Y \\ =& (X'X)^{-1}X'(X\beta + \boldsymbol{\varepsilon}) \\ =& \beta + (X'X)^{-1}X'\boldsymbol{\varepsilon} \\ =& \beta + A\boldsymbol{\varepsilon} \end{aligned}$$

11.4 What is $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|X)$?

$$\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|X) = \mathrm{E}\left[\begin{array}{cccc|c} \varepsilon_1\varepsilon_1 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_n & \\ \varepsilon_2\varepsilon_1 & \varepsilon_2\varepsilon_2 & \cdots & \varepsilon_2\varepsilon_n & X \\ \vdots & \vdots & \ddots & \vdots & \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \cdots & \varepsilon_n\varepsilon_n & \end{array}\right]$$

- Non-autocorrelation says $\mathrm{E}(\varepsilon_i \varepsilon_j) = 0$ for all $j \neq i$
- Homoskedasticity says $\mathrm{E}(\varepsilon_i^2) = \sigma^2$ for all i
- Together $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|X) = \sigma^2 I_{n\times n}$

11.5 Variance of b

- $b = \beta + A\boldsymbol{\varepsilon}$, where $A = (X'X)^{-1}X'$

$$\begin{aligned}
\mathrm{Var}(b|X) =& \mathrm{E}\left[(b - \mathrm{E}(b))(b - \mathrm{E}(b))'|X\right] \\
=& \mathrm{E}\left[(\beta + A\boldsymbol{\varepsilon} - \beta)(\beta + A\boldsymbol{\varepsilon} - \beta)'|X\right] \\
=& \mathrm{E}\left[(A\boldsymbol{\varepsilon})(A\boldsymbol{\varepsilon})'|X\right] \\
=& \mathrm{E}\left[A\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'A'|X\right] \\
=& A\,\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|X\right]A' \\
=& A\sigma^2 I_{n\times n}A' \\
=& \sigma^2 AA' \\
\boxed{} =& \boxed{\sigma^2(X'X)^{-1}}
\end{aligned}$$

- Note: $AA' = (X'X)^{-1}X'X(X'X)^{-1} = (X'X)^{-1}$
- Law of total variance $\mathrm{Var}(b) = \sigma^2\,\mathrm{E}((X'X)^{-1})$

11.6 Estimating $\mathrm{Var}(b)$

- $\mathrm{Var}(b) = \sigma^2\,\mathrm{E}((X'X)^{-1})$
  - Need to estimate σ^2 and $\mathrm{E}((X'X)^{-1})$
  - $\widehat{\mathrm{Var}(b)} = s^2(X'X)^{-1}$
- Since $\sigma^2 = \mathrm{E}(\varepsilon^2)$, a natural, but biased estimator is sample analog $\mathrm{E}(e^2) = (1/n)\sum_{i=1}^n e_i^2$
- An unbiased estimator is $s^2 = \dfrac{1}{n-K}\sum_{i=1}^n e_i^2$
- $\boxed{\widehat{\mathrm{Var}(b|X)} = s^2(X'X)^{-1}}$

11.7 s^2 is an unbiased Estimate of σ^2

- $\mathbf{M}_X$ is residual maker: $\mathbf{M}_X Y = \mathbf{M}_X X + \mathbf{M}_X \varepsilon = e$

$$
\begin{aligned}
\mathrm{E}(e'e) &= \mathrm{E}\left[(\mathbf{M}_X\varepsilon)'(\mathbf{M}_X\varepsilon)\right] \\
&= \mathrm{E}\left[\varepsilon'\mathbf{M}_X'\mathbf{M}_X\varepsilon\right] \quad \text{def. of transpose} \\
&= \mathrm{E}\left[\mathrm{tr}(\varepsilon'\mathbf{M}_X\varepsilon)\right] \quad \text{thing being E'd is scalar} \\
&= \mathrm{E}\left[\mathrm{tr}(\mathbf{M}_X\varepsilon\varepsilon')\right] \quad \text{prop. of trace} \\
&= \mathrm{tr}(\mathbf{M}_X\,\mathrm{E}\left[\varepsilon\varepsilon'\right]) \quad \text{prop. of trace :)} \\
&= \mathrm{tr}(\mathbf{M}_X\sigma^2 I_n) \quad \text{look at prev. slide} \\
&= \sigma^2\,\mathrm{tr}(\mathbf{M}_X) \\
&= \sigma^2\,\mathrm{tr}(I_n - X(X'X)^{-1}X') \\
&= \sigma^2(\mathrm{tr}(I_n) - \mathrm{tr}(X(X'X)^{-1}X')) \\
&= \sigma^2(n - \mathrm{tr}((X'X)^{-1}XX')) = \sigma^2(n - \mathrm{tr}(I_K)) \\
\boxed{= \sigma^2(n-K)} &\Rightarrow \qquad\qquad \boxed{\mathrm{E}((e'e)/(n-K)) = \sigma^2}
\end{aligned}
$$

11.8 Gauss-Markov Theorem

- Gauss-Markov theorem states why we use OLS to estimate LRM
- Theorem: Under the assumptions of the linear regression model, least squares is the minimum variance linear unbiased estimator (MVLUE) of β
  - In the class of linear unbiased estimator, OLS has the smallest variance
  - Linear means estimator is linear function of the data
  - Least squares is **efficient** (best use of the data)

11.9 Implications of Gauss-Markov Theorem

- If we have heteroskedasticity instead of homoskedasticity OLS is unbiased, but not efficient
- An alternative linear unbiased estimator is OLS ignoring some observations
  - This estimator will have a larger variance
- There are other linear BIASed estimators that have smaller variance
- Sometimes we don't know if an estimator is efficient, if we can frame the estimator under the LRM assumptions, then the theorem tells us it is efficient

LECTURE 12

Determinants of Variance of the OLS Estimator

12.1 Variance of Single Parameter

- Recall: $Y = Z\delta + x\beta + \varepsilon$, where Z is a matrix of controls, x is a $n \times 1$ vector of data, with $\mathrm{E}(\varepsilon|Z, x) = 0$
- Let M_Z be the residual maker $(I - Z(Z'Z)^{-1}Z')$

$$Y^* = x^*\beta + \varepsilon$$

- Where $Y^* = \mathrm{M}_Z Y$ and $x^* = \mathrm{M}_Z x$
- b the estimate of β

$$b = (x^{*'}x^*)^{-1}x^{*'}Y^*$$

- This is just OLS with different data, we know the sampling distribution (note b is a scalar)
  - b is unbiased $\mathrm{E}(b) = \beta$, $\mathrm{Var}(b) = \dfrac{\sigma^2}{x^{*'}x^*}$

12.2 Variance for a Single parameter

- $\mathrm{Var}(b) = \dfrac{\sigma^2}{x^{*'}x^*}$
- $x^* = \mathrm{M}_Z x$
  - Vector of residuals from Z on x
  - $x^{*'}x^*$ is the SSE from regression of Z on x
- $R^2_{x|Z} = 1 - \frac{x^{'*}x^*}{\sum_{i=1}^n (x_i - \overline{x})^2}$ (please don't confuse with $R^2_{Y|Z,x}$)
- $x^{'*}x^* = (1 - R^2_{x|Z}) \sum_{i=1}^n (x_i - \overline{x})^2$

$$\boxed{\mathrm{Var}(b) = \frac{\sigma^2_\varepsilon}{(1 - R^2_{x|Z}) \sum_{i=1}^n (x_i - \overline{x})^2}}$$

12.3 Results: Smaller Variance When ...

$$\boxed{\mathrm{Var}(b) = \frac{\sigma^2_\varepsilon}{(1 - R^2_{x|Z}) \sum_{i=1}^n (x_i - \overline{x})^2}}$$

$$SST_x = \sum_{i=1}^n (x_i - \overline{x})^2 = n \cdot \frac{(x_i - \overline{x})^2}{n} = n \cdot \widehat{\mathrm{Var}}(x)$$

1) ... Sample Size Increases
   - SST_x always gets larger as n increases
2) ... More Observed Variability in x increases SST_x
3) ... Adding Controls to Reduce $\mathrm{Var}(\sigma^2)$

12.4 Adding More Controls May Increase or Decrease the Variance of Other Parameters (it depends)

$$\boxed{\text{Var}(b) = \frac{\sigma_\varepsilon^2}{(1 - R_{x|Z}^2)\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

- Adding controls always reduces $\text{Var}(\sigma^2) \downarrow$
- But new controls might *covary* with x, so $R_{x|Z}^2 \uparrow$
- Net effect depends on which effect dominates
- Variance inflation factor $VIF(x) = \dfrac{1}{(1 - R_{x|Z}^2)}$
  - By how much is variance inflated because x covaries with Z compared to ideal world where x and Z are uncorrelated

12.5 Mulitcollinearity

- Multicollinearity is near violation of full rank
  - $Y = Z\delta + z\beta + \boldsymbol{\varepsilon}$
  - Violation of full rank $R_{x|Z}^2 = 1$
  - Multicollinearity $R_{x|Z}^2 \approx 1$

$$\boxed{\text{Var}(b) = \frac{\sigma_\varepsilon^2}{(1 - R_{x|Z}^2)\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

- Multicollinearity produces huge standard errors
- Wikipedia says(?) if $VIF > 10$ then multicollinearity is high
  - But σ^2 could be small or SST_x huge, so may not matter

12.6 Multicollinearity, It gets Worse

- Warning signs
  - Coefficients with high standard errors (variances) but high R^2
  - Small changes in the data produce huge changes in estimates
  - Coefficients have the wrong sign or implausible magnitude
- What to do
  - Be Aware of it
  - No real solution. Dropping variable will usually lead to bias

12.7 Bias/Variance Tradeoff

$$\boxed{\text{Var}(b) = \frac{\sigma_\varepsilon^2}{(1 - R_{x|Z}^2)\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

- Consider
  - $Y = Z\delta + x_1\beta_1 + x_2 \underbrace{\beta_2}_{\approx 0} + \boldsymbol{\varepsilon}$
  - $\text{E}(\boldsymbol{\varepsilon}|Z, x_1, x_2) = 0$, but $\text{E}(x_2|x_1) \neq 0$, and $R_{x_1|Z}^2 \ll R_{x_1|Z,x_2}^2$
- INCLUDING x_2: b_1 is unbiased, but $\text{Var}(b_1)$ is large
- EXCLUDING x_2: b_1 is *biased*, but $\text{Var}(b_1)$ is small
- We need to judge on a case by case basis, which is worse

12.8 Working With and Estimator's Variance

- $\text{Var}(b) = \sigma^2(X'X)^{-1}$ is a variance-covariance matrix
- The kth diagonal element is the variance of b_k
  - Let $v_k = \text{Var}(b_k) = \left[\sigma^2(X'X)^{-1}\right]_{(k,k)}$
- An estimator's *standard error* is the square root of its variance
  - $\text{SE}(b_k) = \sqrt{v_k} = \sqrt{\text{Var}(b_k)} = \sqrt{[\sigma^2(X'X)^{-1}]_{(k,k)}}$
  - Standard Error (versus Standard Deviation), acknowledges variance in estimator due to sampling error
- Since we have to estimate σ^2, need to acknowledge that SE is estimated: $\widehat{\text{SE}(b_k)}$

12.9 Does Job Training Make Manufacturing Firms More Productive

$$\ln(scrap) = \beta_1 + \beta_2 hrsemp + \beta_3 \ln(sale) + \beta_4 \ln(employ) + \varepsilon$$

- *scrap* rate for a manufacturing firm is the number of defective items out of every 100 produced
- *hrsemp* is hours of training per employee

12.10

```
suppressMessages(library(dplyr))
load(file.path(rdata_loc,'jtrain.RData'))
data = data %>%
  filter(year==1987)

summary(data$scrap)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.010   1.000   1.675   4.612   6.000  30.000     103
```

```
summary(data$hrsemp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.000   0.000   8.887  10.000 100.000      28
```

12.11

```
mod = lm(log(scrap)~hrsemp+log(sales)+log(employ) ,data=data)
mod$coefficients
```

```
## (Intercept)      hrsemp  log(sales) log(employ)
## 11.74425473 -0.04218282 -0.95063527  0.99213418
```

```
nobs(mod)
```

```
## [1] 43
```

```
round(vcov(mod),4)
```

```
##             (Intercept)  hrsemp log(sales) log(employ)
## (Intercept)     20.9279  0.0185    -1.6684      1.2170
## hrsemp           0.0185  0.0003    -0.0018      0.0017
## log(sales)      -1.6684 -0.0018     0.1368     -0.1112
## log(employ)      1.2170  0.0017    -0.1112      0.1274
```

```
SEhat = sqrt(diag(vcov(mod)))
round(SEhat,4)
```

```
## (Intercept)      hrsemp  log(sales) log(employ)
##      4.5747      0.0187      0.3698      0.3569
```

Lecture 13

Properties of Least Squares: Consistency and Sampling Distribution

13.1 Properties of the OLS Estimator

$$\begin{aligned} b = (X'X)^{-1}X'Y =& (X'X)^{-1}X'(X\beta + \boldsymbol{\varepsilon}) \\ =& \beta + (X'X)^{-1}X'\boldsymbol{\varepsilon} \\ =& \beta + A'\boldsymbol{\varepsilon} \end{aligned}$$

1) Unbias $\mathrm{E}(b) = \beta$, if $\mathrm{E}(\boldsymbol{\varepsilon}|X) = 0$
2) $\mathrm{Var}(b|X) = \sigma^2(X'X)^{-1}$
   - $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|X) = \sigma^2 I_{n\times n}$
3) Consistency $\operatorname{plim} b = \beta$
   - For unbiased estimator $n \to \infty$ then $\mathrm{Var}(b) \to 0$
   - For biased estimator, $n \to \infty$ then $bias(b) \to 0$ and $\mathrm{Var}(b) \to 0$
4) Sampling distribution for conducting inference

13.2 Probability Limit

- Let b_1 be an estimator of β only using first observation, b_m be an estimator of β using all m observations, b_{m+1} be an estimator of β adding one additional observations to m
- $b_1, b_2, \ldots, b_m, b_{m+1}, b_{m+2}, \ldots$ create a sequence of random variables each adding a random data point
- We would like to say $\operatorname{plim}_{n\to\infty} b_n = \beta$
- Probability limit, define $D(b_n, \beta)$ as the distance between b_n and β, for any η

$$\lim_{n\to\infty} \Pr(D(b_n, \beta) \geq \eta) = 0$$

- Example $\operatorname{plim} \overline{x} = \mathrm{E}(x)$

13.3 Asymptotic Large Sample Properties

- First consistency. Show $\operatorname{plim} b = \beta$

$$\begin{aligned} b =& \beta + (X'X)^{-1}X'\boldsymbol{\varepsilon} \\ =& \beta + (X'X)^{-1}\frac{(n^{-1})^{-1}}{n}X'\boldsymbol{\varepsilon} \\ =& \beta + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\boldsymbol{\varepsilon}}{n}\right) \\ \operatorname{plim} b =& \beta + Q^{-1}\operatorname{plim}\left(\frac{X'\boldsymbol{\varepsilon}}{n}\right) \end{aligned}$$

- $Q = \mathrm{E}(xx')$, i.e. the covariance of the x's
- $\operatorname{plim}\left(\frac{X'\boldsymbol{\varepsilon}}{n}\right) = \operatorname{plim}\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i = \mathrm{E}(x\epsilon) = 0$
- $\operatorname{plim} b = \beta + Q\,\mathrm{E}(x\epsilon) = \beta + Q \times 0 = \beta$ (Consistency)

13.4 Sampling Distribution From Assumption 5

Assumption 5: Normality of disturbances: $\varepsilon \sim N(0, \sigma^2)$

- The distribution of a linear combination of a normally distributed random variable is also normally distributed.

$$b = \beta + A\boldsymbol{\varepsilon}$$

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

$$\boxed{b_k|X \sim N\left(\beta_k, \left[\sigma^2(X'X)^{-1}\right]_{(k,k)}\right)}$$

13.5 Relaxing Assumption 5

- Assumption 5 is that ε is normally distributed
  - Generally, the distribution is unknown
- Assumption 5 is only necessary in small samples
- Knowing the distribution of ε is not necessary
  - Actually only need to know distribution of $\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i$
  - Sample average is an estimator for expected value
  - CLT states that as $n \to \infty$ sample averages are guaranteed to be normally distributed
- Showing b is normally distributed by showing it is a function of sample average

13.6 The OLS Estimator is a Function of a Sample Average

$$\begin{aligned}
b =& \beta + (X'X)^{-1}X'\boldsymbol{\varepsilon} \\
=& \beta + (X'X)^{-1}\frac{(n^{-1})^{-1}}{n}X'\boldsymbol{\varepsilon} \\
=& \beta + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\boldsymbol{\varepsilon}}{n}\right) \\
=& \beta + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i\right) \\
=& \beta + \left(\frac{X'X}{n}\right)^{-1}(\overline{w})
\end{aligned}$$

- CLT: $\overline{w} \sim N\left[\mathrm{E}\left(\overline{w}\right), \mathrm{Var}\left(\overline{w}\right)\right]$

13.7 Consistency of OLS

- $\mathrm{E}(\overline{w}) = \mathrm{E}\left(\frac{X'\boldsymbol{\varepsilon}}{n}\right) = \mathrm{E}(x\varepsilon) = 0$ from mean independence assumption

$$\begin{aligned}
\mathrm{Var}(\overline{w}|X) &= \mathrm{E}(\overline{w}\ \overline{w}') - \mathrm{E}(\overline{w})\,\mathrm{E}(\overline{w})' \\
&= \mathrm{E}\left[\left(\frac{X'\boldsymbol{\varepsilon}}{n}\right)\left(\frac{X'\boldsymbol{\varepsilon}}{n}\right)'|X\right] \\
&= \left(\frac{X'}{n}\right)\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|X\right]\left(\frac{X}{n}\right) \\
&= \left(\frac{\sigma^2}{n}\right)\left(\frac{X'X}{n}\right) \\
\mathrm{Var}(\overline{w}) &= \left(\frac{\sigma^2}{n}\right)\mathrm{E}\left(\frac{X'X}{n}\right)
\end{aligned}$$

13.8 Asymptotic Normality

- Central limit theorem states that sample average of independently drawn data is approximately normal
- $b|X = \beta + \left(\frac{X'X}{n}\right)^{-1}\overline{w}$
- $\mathrm{E}(\overline{w}) = 0$, $\mathrm{Var}(\overline{w}|X) = \left(\frac{\sigma^2}{n}\right)\left(\frac{X'X}{n}\right)$
- CLT: $\overline{w}|X \sim N\left[0, \left(\frac{\sigma^2}{n}\right)\left(\frac{X'X}{n}\right)\right]$

$$b|X \sim N\left[\beta, \left(\frac{X'X}{n}\right)^{-1}\left(\frac{\sigma^2}{n}\right)\left(\frac{X'X}{n}\right)\left(\frac{X'X}{n}\right)^{-1}\right]$$

$$\boxed{b|X \sim N\left[\beta, \sigma^2(X'X)^{-1}\right]}$$

Lecture 14

Using the Sampling Distribution for Interval Estimation

14.1 Interval Estimation

- b is our best guess of β
- If I have 100 different samples, $\Pr(\beta = b) = 0$ for all 100 samples
- Interval estimation uses the data to identify a range of plausible values (a confidence interval)
  - $\Pr(b^{lower} \leq \beta \leq b^{upper}) = C$ for confidence level C
- With 100 different samples C fraction will contain β
- Interpretation: Probability that a random interval contains β is C. Not Probability β is in the interval
- I carry out a procedure, which produces a random number. C fraction of the instances I carry out this procedure the random interval will contain the truth.

14.2 Interval Estimation

- Let $v_k = \text{Var}(b_k) = \left[\sigma^2(X'X)^{-1}\right]_{(k,k)}$
- By Normality $b_k \sim N(\beta_k, v_k)$
- Re-arranging

$$\frac{(b_k - \beta_k)}{\sqrt{v_k}} \sim N(0,1)$$
$$\frac{(b_k - \beta_k)}{\text{SE}(b_k)} \sim N(0,1)$$

14.3 Interval Estimation Math

- If Z is standard normal then $\Pr(Z < -1.96) = .025$
- If β_k was known and b_k was estimated
  - 2.5% of the time $\frac{b_k - \beta_k}{\text{SE}(b_k)} \leq -1.96$
  - 2.5% of the time $\frac{b_k - \beta_k}{\text{SE}(b_k)} \geq +1.96$

$$\Pr\left(-1.96 \leq \frac{b_k - \beta_k}{\text{SE}(b_k)} \leq 1.96\right) = 0.95$$
$$\Pr\left(-1.96\,\text{SE}(b_k) \leq b_k - \beta_k \leq 1.96\,\text{SE}(b_k)\right) = 0.95$$
$$\Pr\left(-b_k - 1.96\,\text{SE}(b_k) \leq -\beta_k \leq -b_k + 1.96\,\text{SE}(b_k)\right) = 0.95$$
$$\Pr\left(b_k + 1.96\,\text{SE}(b_k) \geq \beta_k \geq b_k - 1.96\,\text{SE}(b_k)\right) = 0.95$$

$$\boxed{\Pr\left(b_k - 1.96\,\text{SE}(b_k) \leq \beta_k \leq b_k + 1.96\,\text{SE}(b_k)\right) = 0.95}$$

14.4 Interval Estimation in Practice

- $\dfrac{b_k - \beta_k}{\text{SE}(b_k)} \sim N(0,1)$
- We need to estimate the standard error, $\widehat{\text{SE}(b_k)}$

$$\widehat{\text{SE}(b_k)} = \sqrt{\widehat{\text{Var}(b_k)}} = \sqrt{[s^2(X'X)^{-1}]_{(k,k)}}$$

  - $s^2 = \sum_{i=1}^{n} e_i^2/(n-K)$
- t-distribution with $n-K$ DF

$$\boxed{\frac{b_k - \beta_k}{\widehat{\text{SE}(b_k)}} \sim t[n-K]}$$

14.5 Steps to Interval Estimation

(1) Choose level of significance, α (i.e. 0.05 significance for 95% confidence interval)
  - $\Pr\left(b_k^{lower} \leq \beta_k \leq b_k^{upper}\right) = 1-\alpha$

(2) Choose critical value (CV) from $t[n-K]$ distribution $\Pr(t < \text{CV}) = \alpha/2$
  - Will be very close to critical value from standard normal when $n-K > 100$

(3) Construct interval
  - $\left[b_k - \text{CV} \times \widehat{SE(b_k)}, b_k + \text{CV} \times \widehat{SE(b_k)}\right]$

$$\Pr\left(b_k - \text{CV} \times \widehat{SE(b_k)} \leq \beta_k \leq b_k + \text{CV} \times \widehat{SE(b_k)}\right) = 1-\alpha$$

14.6 Does Job Training Make Manufacturing Firms More Productive

$$\ln(scrap) = \beta_1 + \beta_2 hrsemp + \beta_3 \ln(sale) + \beta_4 \ln(employ) + \varepsilon$$

- *scrap* rate for a manufacturing firm is the number of defective items out of every 100 produced
- *hrsemp* is hours of training per employee

14.7

```
suppressMessages(library(dplyr))
load(file.path(rdata_loc,'jtrain.RData'))
data = data %>%
  filter(year==1987)
mod = lm(log(scrap)~hrsemp+log(sales)+log(employ) ,data=data)
mod$coefficients
```

```
## (Intercept)      hrsemp  log(sales) log(employ)
## 11.74425473 -0.04218282 -0.95063527  0.99213418
```

```
nobs(mod)
```

```
## [1] 43
```

```
round(vcov(mod),4)
```

```
##             (Intercept)  hrsemp log(sales) log(employ)
## (Intercept)     20.9279  0.0185    -1.6684      1.2170
## hrsemp           0.0185  0.0003    -0.0018      0.0017
## log(sales)      -1.6684 -0.0018     0.1368     -0.1112
## log(employ)      1.2170  0.0017    -0.1112      0.1274
```

```
SEhat = sqrt(diag(vcov(mod)))
round(SEhat,4)
```

```
## (Intercept)      hrsemp  log(sales) log(employ)
##      4.5747      0.0187      0.3698      0.3569
```

14.8

```
C = 1 - .05
DF = nobs(mod) - length(mod$coefficients)
DF
```

```
## [1] 39
```

```
#or
mod$df.residual
```

```
## [1] 39
```

```
CV = abs(qt((1-C)/2,DF))
CV
```

```
## [1] 2.022691
```

```
mod$coefficients['hrsemp'] + c(-1,1)*CV*SEhat['hrsemp']
```

```
## [1] -0.079958211 -0.004407433
```

```
confint(mod,level=c(.95))
```

```
##                   2.5 %       97.5 %
## (Intercept)  2.49104365 20.997465810
## hrsemp      -0.07995821 -0.004407433
## log(sales)  -1.69869878 -0.202571758
## log(employ)  0.27019216  1.714076198
```

LECTURE 15

Hypothesis Testing For A Single Parameter

15.1 Hypothesis Testing: The Main Idea

- The estimator is a random variable

$$\underbrace{b_k}_{\text{We observe this}} \sim N\Big(\underbrace{\beta_k}_{\substack{\text{Trying to learn about} \\ \text{location of this}}} , \underbrace{\widehat{SE(b_k)}}_{\substack{\text{We can} \\ \text{estimate this}}} \Big)$$

- Using a draw on a random variable to learn about the location that produced it

15.2 Ethnic Favoritism In Kenya

- Do Kenyan Presidents show favoritism to districts that share their ethnicity?
- (LATER) Do democratic elections force Kenyan presidents to be more equitable?
- Analyze road expenditures in districts in Kenya 1963-2011
  - *road_share* is (%) of national road budget received by district, divided by (%) of national population in district
  - *coethnic* indicator if ≥ 50 percent district i's has same ethnicity as president

15.3 Ethnic Favoritism In Kenya

$$road\_share_i = \beta_1 + \beta_2 coethnic_i + \beta_3 area + \beta_4 pop + \varepsilon_i$$

- A hypothesis is about the value of the parameters (it is a set of restrictions on the possible values)
- Hypothesis Testing
  - Is a more restricted version of our model consistent with the data? If so, then then their is evidence that the restriction (the hypothesis) is correct.
  - Does the data 'support' or reject the hypothesis? At what significance level (precision)?
- Is there ethnic favoritism ... Is $\beta_2 = 0$?

15.4 Review of Hypothesis Testing

- The data and the null hypothesis produce a single observation of a random variable (a test statistic)
- Under the null, we know the distribution of this random variable
- We look to see if it is likely that the random variable we observed actually came from this distribution
  - If the random variable has extremely large value, it is unlikely to come from the the distribution

    - Reject the the null hypothesis
  - If the random variable 'appears' to come from the distribution
    - Fail to reject the null

15.5 Steps to Hypothesis Testing

1) State the Hypothesis
2) Compute the test statistic, assuming null is true
3) Define rejection region
   - Choose significance level α: Frequency we are comfortable erroneously rejecting the null when it is true (usually $\alpha = 0.05$)
   - Combining significance level and distribution provides a critical value for the random variable and a rejection region
4) See if test statistic falls in or out of rejection region and conclude the test

15.6 Hypothesis Testing for Single Parameter

1) State the Hypothesis
   - $H_0 : \beta_k = q$
   - $H_1 : \beta_k \neq q$
2) Compute the test statistic, assume null is true
   - $\dfrac{b_k - \beta_k}{\widehat{\text{SE}}(b_k)} \sim t[n - K]$
   - Test statistic: $t = \dfrac{b_k - q}{\widehat{\text{SE}}(b_k)}$

15.7 Hypothesis Testing for Single Parameter

3) Define rejection region by choosing significance level α
   - Under the null, the test statistic is a random draw from t-distribution with $n - K$ DF
   - α is frequency we are comfortable erroneously rejecting the null when it is true
     - Usually 5% level (1 time out of 20) we will reject the null when null is true
   - The significance level and the distribution imply a critical value for the random variable and a rejection region
4) See if test statistic falls in or out of rejection region

15.8

```
suppressMessages(library(dplyr))
suppressMessages(library(lmtest))
load(file.path(rdata_loc,'KENYA.RData'))
data = data %>%
  mutate(pop = pop1962/1000000,
         area = area/100000) %>%
  rename(coethnic=president)
mod = lm(exp_dens_share~coethnic+area+pop,data=data)
paste('Number of Observations: ',nobs(mod))
```

```
## [1] "Number of Observations:  2009"
```

```
formula(mod)
```

```
## exp_dens_share ~ coethnic + area + pop
```

```
coeftest(mod)[ ,1:2]
```

```
##             Estimate Std. Error
## (Intercept)   1.3699     0.1430
## coethnic      0.8465     0.1748
## area          0.7306     0.3994
## pop          -1.6815     0.4133
```

15.9 Is there evidence of ethnic favoritism

- $H_0 : \beta_{coethnic} = 0$, $H_1 : \beta_{coethnic} \neq 0$
- Under null compute test statistic
  - $t = \dfrac{b_k - \beta_k}{\widehat{\text{SE}}(b_k)} = \dfrac{0.8465 - 0}{0.1748} = 4.843$
- Is is plausible that t came from a t-distribution with 2009-4 DF?
- Choose critical value (cv) for 5% significance level
  - Two-sided test: `qt(.05/2,2009-4,lower.tail = F) = 1.961`. Consider both tails.
- Reject if t is in rejection region: $|t| >$ cv

15.10 Comments on Testing a Single Parameter

- Since we standardize the estimates, they all have the same distribution
  - If all hypothesis tests use same α they have the same critical value
- We use a t-distribution because we know the small sample properties. For estimators (like MLE) that we only know large sample properties we use a normal
- For these distributions a critical value of $2 \approx (\alpha = 0.05)$
  - Short hand the hypothesis $\beta = 0$. If estimate is twice as large as the standard error, claim evidence β is statistically different than zero

15.11 Relationship Between Hypothesis Tests and Confidence Intervals

- The sampling distribution of b is used to 1) Construct confidence intervals, and 2) Conduct hypothesis test
- The two task are related because they rely on the same distribution
- A confidence interval provides *EVERY* null hypothesis we would *fail to reject* for a *GIVEN* significance level
  - 95% confidence interval, *EVERY* null would *fail to reject* at 5% significance level
  - Likewise a 99% confidence interval every value we would fail to reject at the 1% level

15.12

```
confint(mod,level=.95)
```

```
##                2.5 % 97.5 %
## (Intercept)  1.08946  1.650
## coethnic     0.50378  1.189
## area        -0.05276  1.514
## pop         -2.49195 -0.871
```

```
confint(mod,level=.99)
```

```
##                0.5 % 99.5 %
## (Intercept)  1.0012  1.738
## coethnic     0.3959  1.297
## area        -0.2993  1.760
## pop         -2.7470 -0.616
```

15.13 Summarizing Hypothesis Test with a P-Value

- For a *GIVEN* null hypothesis, a p-value summarizes the *EVERY* significance level we would *reject the null*
- We reject when test statistic is greater than critical value ($|t| > \text{cv}$)
- P-value is the smallest possible α, so the critical value equals the test statistic ($|t| = \text{cv}$)
- Invert the hypothesis test

$$\text{p-value} = 2 \times [1 - \text{CDF-t}(abs(\text{t-value}), \text{DF})]$$

- T-stat for $H_0 : \beta_{coethnic} = 0$, t-stat $= 4.843$
  - `2*(1-pt(abs(4.843),2009-4))=1.377e-06`

15.14 p-values

- Definition: A p-value states the probability that we would observe the test statistic we actually observed or one more extreme and the null is true
  - e.g., probability observe t-stat $= 4.843$ or one more extreme and $\beta_{coethnic} = 0$

  - p-value is NOT the probability that the null is correct (note p-value can be 1!)

- Any α *GREATER* than p-value *reject null*
- Any α *LESS* than p-value *fail to reject null*
- Universal interpretation. Do not need to know specifics of test statistic construction
  - Only need to know H_0 and p-value

15.15 Most Statistical Software Gives You t-stats and p-values for $H_0 : \beta_k = 0$

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.370      0.143    9.58  < 2e-16 ***
## coethnic       0.847      0.175    4.84  1.4e-06 ***
## area           0.731      0.399    1.83    0.068 .
## pop           -1.681      0.413   -4.07  4.9e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

LECTURE 16

Sampling Distribution of Estimates of Estimates

16.1 Many Times We Would Like to Conduct Inference on A Function of the Parameters

- Confidence intervals or hypothesis tests for predicted values given a value of x
  - $\mathrm{E}(y|x) = x'\beta$
  - Estimate $\hat{y} = x'b$
- Hypothesis tests about relation of parameter
  - e.g., $\beta_1 = \beta_2$
  - $\theta = \beta_1 - \beta_2 = 0$
  - Estimate $\hat{\theta} = b_1 - b_2$
- What are the sampling distributions of these estimates?

16.2 Sampling Distribution of Linear Function of the Parameters

- New Parameter $\theta = A'\beta$
- Given A is any column vector $K \times 1$ of known constants
- Estimate $\hat{\theta} = A'b$
- Estimate is unbiased, $\mathrm{Var}(\hat{\theta}) = A' \mathrm{Var}(b) A$
- Distribution: $\hat{\theta} \sim N(A'\beta, A' \mathrm{Var}(b) A)$

16.3 Example, Ethnic Favoritism In Kenya

- We showed that Kenyan Presidents show favoritism to districts that share their ethnicity
- Does democratic elections force presidents to be more equitable?
- Analyze road expenditures in districts in Kenya 1963-2011
  - *road_share* is (%) of national road budget received by district, divided by (%) of national population in district
  - *coethnic* indicator if ≥ 50 percent district i's has same ethnicity as president
  - *democratic* indicator if the president was elected democraticaly

16.4 Ethnic Favoritism In Kenya

$$\begin{aligned} road\_share_i = \beta_1 &+ \beta_2 coethnic_i + \beta_3 democratic_i \\ &+ \beta_4 coethnic_i \times democratic_i + \beta_5 area + \beta_6 pop + \varepsilon_i \end{aligned}$$

- $\beta_2 = 0$ is ethnic favoritism when no democracy
- Does democracy undo ethnic favoritism

  - Is $\beta_2 = -\beta_4$?

- Hypothesis
  - $H_0 : \beta_2 = -\beta_4$
  - $H_1 : \beta_2 \neq -\beta_4$

- Restated: $H_0 : \beta_2 + \beta_4 = 0$
  - This is a linear hypothesis

16.5

```
suppressMessages(library(dplyr))
suppressMessages(library(lmtest))
load(file.path(rdata_loc,'KENYA.RData'))
data = data %>%
  mutate(pop = pop1962/1000000,
         area = area/100000) %>%
  rename(coethnic=president,democratic=multiparty)
mod = lm(exp_dens_share~coethnic+democratic+
           I(coethnic*democratic)+area+pop,data=data)
paste('Number of Observations: ',nobs(mod))
```

```
## [1] "Number of Observations:  2009"
```

```
coeftest(mod)
```

```
##
## t test of coefficients:
##
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.250      0.160    7.82  8.2e-15 ***
## coethnic                     1.462      0.252    5.80  7.7e-09 ***
## democratic                   0.212      0.134    1.58  0.11435
## I(coethnic * democratic)    -1.143      0.338   -3.38  0.00073 ***
## area                         0.737      0.399    1.85  0.06473 .
## pop                         -1.650      0.412   -4.00  6.6e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

16.6 Hypothesis Test of Linear Function

```
A = matrix(c(0,1,0,1,0,0),ncol = 1)
b = mod$coefficients
V = vcov(mod)

theta_hat = t(A) %*%b
theta_se = sqrt(t(A) %*% V %*% A)
t_val = (theta_hat - 0)/theta_se
t_val
```

```
##       [,1]
## [1,] 1.363
```

```
crit_val = qt(.05/2,2009-6,lower.tail = F)
crit_val
```

```
## [1] 1.961
```

```
p_val = 2*(1-pt(abs(t_val),2009-6))
p_val
```

```
##        [,1]
## [1,] 0.1732
```

LECTURE 17

Testing Multiple Hypothesis

17.1 Generalizing the Hypothesis Test

- In testing a single parameter $H_0 : \beta_k = \text{q}$, usually a t-test can be formulated
- We also showed how to formulate a t-test for a linear function of the parameters, $H_0 : A'\beta = \text{q}$
- Now consider any multiple linear hypothesis

$$H_0 : \text{R}\beta = \text{q}$$

- Example: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
  - Test $\text{E}(y|x_1, x_2) = x_2$
  - $H_0 : \beta_0 = 0 \,\&\, \beta_1 = 1 \,\&\, \beta_2 = 0$

17.2 Wald Principle

- Given the NULL: $\text{R}\beta = \text{q}$, re-write as

$$H_0 : \text{R}\beta - \text{q}$$

- Define $m = \text{R}\beta - \text{q}$
- Under the NULL m is $j \times 1$ multivariate normal, $m \sim N(0, \text{Var}(m))$
- Wald Principle: Test statistic

$$W = m'(\text{Var}(m)^{-1})m \sim \chi^2[J]$$

  is chi-squared with $\#r$ degrees of freedom

  - $\text{Var}(m) = \text{R}\,\text{Var}(b)\text{R}'$

17.3 Example 1: Test of a single parameter

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
- $H_0 : \beta_1 = 1$
- Need to choose some R and q so that

$$\begin{aligned} H_0 : \beta_1 &= 1 \\ H_0 : \text{R}\beta - \text{q} &= 0 \\ H_0 : \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} - [1] &= 0 \end{aligned}$$

- $\text{R} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$ and $\text{q} = 1$
- Define $m = \text{R}b - \text{q}$, $\text{Var}(m) = \text{R}\,\text{Var}(b)\text{R}'$

17.4 Example 1: Test of a single parameter

- We know in the test of a single parameter, $H_0 : \beta_1 = 1$

$$(b_1 - 0)/\operatorname{SE}(b_1) \sim N(0,1)$$

- With the Wald Test $m'(\operatorname{Var}(m)^{-1})m$

$$m = \mathrm{R}b - \mathrm{q} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} - [1] = (b_1 - 1)$$

$$\operatorname{Var}(m) = \mathrm{R} \operatorname{Var}(b) \mathrm{R}' = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \operatorname{Var}(b) \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \operatorname{Var}(b_1)$$

- $W = m'(\operatorname{Var}(m)^{-1})m = (b_1 - 1)^2/\operatorname{Var}(b_1) \sim \chi^2[1]$

17.5 Example 2: Test of Multiple Parameters

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
  $\beta_0 = 0$
- $H_0 : \beta_1 = 1$
  $\beta_2 = 0$

$$H_0 : \mathrm{R}\beta - \mathrm{q} = 0_{3\times 1}$$

$$H_0 : \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathrm{R}} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} - \underbrace{\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}}_{\mathrm{q}} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

- $W = (\mathrm{R}b - \mathrm{q})' \left[\mathrm{R} \operatorname{Var}(b) \mathrm{R}'\right]^{-1} (\mathrm{R}b - \mathrm{q}) \sim \chi^2[3]$

17.6 Comment

- Finite Sample Adjustment for OLS
  - Since $\operatorname{Var}(b)$ is estimated $\widehat{\operatorname{Var}(b)}$ need to adjust the test statistic

$$F = W/J \sim F[J, n - K]$$

  - J is the number of restrictions it depends on the hypothesis
    - Number of rows of R
  - K is the dimension of β

17.7 Empirical Example

- Does childhood environment impact wages

$$\ln(wage_i) = \beta_1 + \beta_2 educ_i + \beta_3 Abil_i + \\ \beta_4 MomEd_i + \beta_5 Sib_i + \beta_6 BH_i + \varepsilon_i$$

- Data on same 2,178 interviewed annually, sample size = 17,919
  - Select one row per person when have 7 years of experience

17.8 Koop Tobias Data

```
suppressMessages(library(dplyr))
suppressMessages(library(lmtest))
load(file.path(rdata_loc,'KoopTobias.RData'))
data = data %>%
  filter(PotExper==7)
data[1:10,c('ID','time','logwage','PotExper','Educ','MomEd','Sib','BH')
```

```
##    ID time logwage PotExper Educ MomEd Sib BH
## 1   2    9    2.46        7   15    12   1  0
## 2   4   10    2.48        7   13    12   4  1
## 3   6    7    2.37        7   15    12   2  0
## 4   8    3    2.09        7   13    12   2  1
## 5  12    5    2.81        7   13    13   5  0
## 6  13    4    2.09        7   12    12   4  1
## 7  14    5    2.24        7   12    10   2  1
## 8  15    9    2.37        7   13    10   3  1
## 9  17    5    2.39        7   12     9   2  1
## 10 19    4    3.01        7   12    12   3  1
```

17.9 Estimates

```
# can't control for potential experience because it violates FULL RANK
mod = lm(logwage~Educ+Abil+MomEd+Sib+BH, data = data)
coeftest(mod)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.37192    0.10332   13.28  < 2e-16 ***
## Educ         0.06463    0.00700    9.23  < 2e-16 ***
## Abil         0.10173    0.01665    6.11  1.3e-09 ***
## MomEd        0.00736    0.00460    1.60     0.11
## Sib          0.00893    0.00585    1.53     0.13
## BH          -0.04270    0.03310   -1.29     0.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

17.10 Hypothesis Test

```
#suppressMessages(library("AER"))
#mod$coefficients
#linearHypothesis(mod,'coethnic=-I(coethnic * democratic)')
b = matrix(mod$coefficients,ncol=1)
R = rbind(c(0,0,0,1,0,0),
          c(0,0,0,0,1,0),
          c(0,0,0,0,0,1))
q = matrix(0,ncol=1,nrow=3)
J = dim(R)[1]
V = vcov(mod)
m = R %*% b - q
Fstat = (t(m) %*% solve(R %*% V %*% t(R)) %*% m )/J
Fstat
```

```
##      [,1]
## [1,] 1.87
```

```
#5% test, critical value
crit_val = qf(1 - .05,J,mod$df.residual)
crit_val
```

```
## [1] 2.611
```

```
p_val = 1-pf(Fstat,J,mod$df.residual)
p_val
```

```
##        [,1]
## [1,] 0.1327
```

17.11 Built-In Hypothesis Test

```
suppressMessages(library("AER"))
linearHypothesis(mod,c('MomEd=0',
                       'Sib=0',
                       'BH=0'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## MomEd = 0
## Sib = 0
## BH = 0
##
## Model 1: restricted model
## Model 2: logwage ~ Educ + Abil + MomEd + Sib + BH
##
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1   1474 310
## 2   1471 309  3      1.18 1.87   0.13
```

17.12 Joint Test of Model Significance

```
summary(mod)
```

```
##
## Call:
## lm(formula = logwage ~ Educ + Abil + MomEd + Sib + BH, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1589 -0.2708  0.0213  0.2891  1.7317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.37192    0.10332   13.28  < 2e-16 ***
## Educ         0.06463    0.00700    9.23  < 2e-16 ***
## Abil         0.10173    0.01665    6.11  1.3e-09 ***
## MomEd        0.00736    0.00460    1.60     0.11
## Sib          0.00893    0.00585    1.53     0.13
## BH          -0.04270    0.03310   -1.29     0.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.458 on 1471 degrees of freedom
## Multiple R-squared:  0.173,  Adjusted R-squared:  0.17
## F-statistic: 61.7 on 5 and 1471 DF,  p-value: <2e-16
```

LECTURE 18

Correcting the Standard Errors for Heteroskedasticity

18.1 Getting the Correct Standard Errors

- Given data, $Y = X\beta + \boldsymbol{\varepsilon}_{n\times 1}$
- Need $\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|X)$ to construct standard errors
- In classical LRM we made two assumptions to construct standard errors
  - Homoskedastic: $\mathrm{Var}(\varepsilon_i) = \mathrm{E}(\varepsilon_i^2) = \sigma^2$
  - Non-autocorrelation: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$
    $\implies \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 I_{n\times n}$
- If either of these assumptions are wrong, we have wrong standard errors
  - t-tests, F-tests, confidence intervals are all wrong

18.2 Generalized Linear Regression Model

$$\begin{array}{cc} Y = X\beta + \boldsymbol{\varepsilon} & \text{Linear} \\ \mathrm{E}(\boldsymbol{\varepsilon}|X) = 0 & \text{Mean Independence} \\ \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \Sigma & \text{General Covariance} \end{array}$$

- Implications
  1) OLS is still unbiased
     - Can use b and simply adjust standard errors
     - Need to understand how to correct standard errors
  2) OLS is no longer MVLUE, only LUE
     - There exists (potentially) more efficient estimators
  3) R-squared is still OK

18.3 We Can

- Construct heteroskedasticity-robust standard errors
  - Huber-White Standard errors
  - Robustness to unknown heteroskedasticity
- Test for heteroskedasticity
  - Breusch-Pagan

18.4 Correcting the OLS Standard Errors

- Our approach to heteroskedasticity is to estimate b with OLS and then adjust the standard errors
- Robust standard errors: Robustness to unknown heteroskedasticity
- $\text{Var}(\varepsilon_i) = \sigma_i^2$, $\text{E}(\boldsymbol{\varepsilon\varepsilon}'|X) = \Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$
- We do not need to estimate n new parameters, which is good news

18.5 Huber-White Standard errors

- Heteroskedasticity-robust standard errors after OLS

$$\begin{aligned} \text{Var}(b|X) =& \text{E}((b-\beta)(b-\beta)'|X) \\ =& \text{E}(((X'X)^{-1}X'\varepsilon)((X'X)^{-1}X'\varepsilon)'|X) \\ =& \text{E}((X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X) \\ =& (X'X)^{-1}\,\text{E}(X'\varepsilon\varepsilon'X|X)(X'X)^{-1} \end{aligned}$$

- Estimate $\text{E}(X'\varepsilon\varepsilon'X|X)$ with sample
  - Make degrees of freedom adjustment for small sample

$$\widehat{\text{Var}(b|X)} = \frac{n}{n-K}(X'X)^{-1}X' \begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_n^2 \end{bmatrix} X(X'X)^{-1}$$

18.6 Revisit Koop and Tobias

```
suppressMessages(library(dplyr))
suppressMessages(library(lmtest))
load(file.path(rdata_loc,'KoopTobias.RData'))
data = data %>%
  filter(PotExper==7)
mod = lm(logwage~Educ+Abil+MomEd+Sib+BH, data = data)
coeftest(mod)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.37192    0.10332   13.28  < 2e-16 ***
## Educ         0.06463    0.00700    9.23  < 2e-16 ***
## Abil         0.10173    0.01665    6.11  1.3e-09 ***
## MomEd        0.00736    0.00460    1.60     0.11
## Sib          0.00893    0.00585    1.53     0.13
## BH          -0.04270    0.03310   -1.29     0.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

18.7 Heteroskedastic Robust Covariance Matrix

```
n = nobs(mod)
X = cbind(rep(1,n),data$Educ,data$Abil,data$MomEd,data$Sib,data$BH)
XX = t(X)%*%X
XXi = solve(t(X)%*%X)
S = XXi %*% (t(X)%*%diag(mod$residuals^2)%*%X)%*%XXi
S = (n/mod$df.residual)*S
S
```

```
##             [,1]       [,2]       [,3]       [,4]       [,5]
## [1,]  0.0117515 -7.027e-04  1.004e-03 -1.845e-04 -1.771e-04
## [2,] -0.0007027  6.266e-05 -6.228e-05 -8.193e-06 -3.230e-07
## [3,]  0.0010037 -6.228e-05  2.935e-04 -2.360e-05  8.730e-06
## [4,] -0.0001845 -8.193e-06 -2.360e-05  2.366e-05  6.606e-06
## [5,] -0.0001771 -3.230e-07  8.730e-06  6.606e-06  3.352e-05
## [6,] -0.0001350  2.159e-05  3.258e-05 -2.473e-05 -9.190e-06
##             [,6]
## [1,] -1.350e-04
## [2,]  2.159e-05
## [3,]  3.258e-05
## [4,] -2.473e-05
## [5,] -9.190e-06
## [6,]  1.113e-03
```

18.8 Tell R to Use a Difference Covariance

```
coeftest(mod, vcov = S)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.37192    0.10840   12.66  < 2e-16 ***
## Educ         0.06463    0.00792    8.16  6.9e-16 ***
## Abil         0.10173    0.01713    5.94  3.6e-09 ***
## MomEd        0.00736    0.00486    1.51     0.13
## Sib          0.00893    0.00579    1.54     0.12
## BH          -0.04270    0.03336   -1.28     0.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

18.9 There is a built-in function HC=heteroskedastic consistent

```
library(sandwich)
#vcovHC with "HC0", "HC1", "HC2", "HC3", differenet DF adjustments
coeftest(mod, vcov = vcovHC(mod, "HC1"))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.37192    0.10840   12.66  < 2e-16 ***
## Educ         0.06463    0.00792    8.16  6.9e-16 ***
## Abil         0.10173    0.01713    5.94  3.6e-09 ***
## MomEd        0.00736    0.00486    1.51     0.13
## Sib          0.00893    0.00579    1.54     0.12
## BH          -0.04270    0.03336   -1.28     0.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

18.10 Testing For Heteroskedasticity

- In general, use robust standard errors. More reassuring and not likely to change any conclusions with larger samples
- We would like to test $H_0 : \text{Var}(\varepsilon|x_1, x_2, \cdots, x_K) = \sigma^2$
  - The Null is Data is Homoskedastic, $\text{Var}(\varepsilon)$ does not depend on X
- Since $\text{E}(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \text{E}(\varepsilon^2) - \text{E}(\varepsilon)^2$
- The NULL is, I cannot predict ε^2 with the x's
- $H_0 : \text{E}(\varepsilon^2|x) = \sigma^2$

$$\varepsilon^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_K x_K + \nu$$

18.11 Breusch-Pagan test for heteroskedasticity (BP test)

- Estimate
$$e_i^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_K x_K + \nu$$
- F-test for $\delta_1 = \delta_2 = \cdots = \delta_K = 0$
- Short-cut LM test $n \times R^2_{e^2} \sim \chi_{\#\text{ regressors}}$
- White test, add all second order terms

18.12 Testing For Heteroskedasticity

```
mod_residsq=lm(I(mod$residuals^2)~Educ+Abil+MomEd+Sib+BH, data = data)
summary(mod_residsq)
```

```
##
## Call:
## lm(formula = I(mod$residuals^2) ~ Educ + Abil + MomEd + Sib +
##     BH, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.291 -0.182 -0.123  0.040  4.371
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.030344   0.086351    0.35   0.7253
## Educ         0.018476   0.005850    3.16   0.0016 **
## Abil        -0.018164   0.013917   -1.31   0.1921
## MomEd       -0.005019   0.003847   -1.30   0.1922
## Sib         -0.000959   0.004887   -0.20   0.8445
## BH           0.011024   0.027661    0.40   0.6903
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.383 on 1471 degrees of freedom
## Multiple R-squared:  0.00732,    Adjusted R-squared:  0.00394
## F-statistic: 2.17 on 5 and 1471 DF,  p-value: 0.0553
```

18.13 Testing For Heteroskedasticity

```
bp = nobs(mod_residsq)*summary(mod_residsq)$r.squared
bp
```

```
## [1] 10.8
```

```
1 - pchisq(bp,5)
```

```
## [1] 0.05539
```

```
bptest(mod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod
## BP = 11, df = 5, p-value = 0.06
```

18.14 Joint Hypothesis Test w/ Heteroskedasticity

```
suppressMessages(library("AER"))
linearHypothesis(mod,
                 c('MomEd=0',
                       'Sib=0',
                       'BH=0'),
                 vcov = vcovHC(mod, "HC1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## MomEd = 0
## Sib = 0
## BH = 0
##
## Model 1: restricted model
## Model 2: logwage ~ Educ + Abil + MomEd + Sib + BH
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df    F Pr(>F)
## 1   1474
## 2   1471  3 1.63   0.18
```

Lecture 19

Clustered Standard Errors

19.1 Revisit Koop and Tobias

```
suppressMessages(library(dplyr))
suppressMessages(library(lmtest))
load(file.path(rdata_loc,'KoopTobias.RData'))
data[1:20,]
```

```
##    ID  Abil MomEd FathED BH Sib Educ logwage PotExper time
## 1   1  1.00    12     12  0   1   13    1.82        1    0
## 2   1  1.00    12     12  0   1   18    3.29        3    7
## 3   1  1.00    12     12  0   1   18    3.21        5    9
## 4   1  1.00    12     12  0   1   18    3.06        6   10
## 5   2  1.50    12     12  0   1   15    2.14        4    6
## 6   2  1.50    12     12  0   1   15    2.30        5    7
## 7   2  1.50    12     12  0   1   15    2.40        6    8
## 8   2  1.50    12     12  0   1   15    2.46        7    9
## 9   2  1.50    12     12  0   1   15    2.51        8   10
## 10  2  1.50    12     12  0   1   15    2.50        9   11
## 11  2  1.50    12     12  0   1   15    2.55       10   12
## 12  2  1.50    12     12  0   1   15    2.56       11   13
## 13  2  1.50    12     12  0   1   15    2.60       12   14
## 14  3 -0.36    12     12  1   1   10    1.56        1    2
## 15  4  0.26    12     10  1   4   12    1.85        1    3
## 16  4  0.26    12     10  1   4   12    2.14        2    4
## 17  4  0.26    12     10  1   4   12    1.81        3    5
## 18  4  0.26    12     10  1   4   12    1.15        4    6
## 19  4  0.26    12     10  1   4   13    1.93        4    7
## 20  4  0.26    12     10  1   4   13    2.08        5    8
```

19.2 Filtering to 7 Years of Experience

```
mod = lm(logwage~Educ+Abil+MomEd+Sib+BH,
         data = filter(data,PotExper==7))
coeftest(mod)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.37192    0.10332   13.28  < 2e-16 ***
## Educ         0.06463    0.00700    9.23  < 2e-16 ***
## Abil         0.10173    0.01665    6.11  1.3e-09 ***
## MomEd        0.00736    0.00460    1.60     0.11
## Sib          0.00893    0.00585    1.53     0.13
## BH          -0.04270    0.03310   -1.29     0.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

19.3 Filtering to 6 Years of Experience

```
mod = lm(logwage~Educ+Abil+MomEd+Sib+BH,
         data = filter(data,PotExper==6))
coeftest(mod)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.21265    0.10590   11.45  < 2e-16 ***
## Educ         0.07349    0.00715   10.28  < 2e-16 ***
## Abil         0.08831    0.01692    5.22  2.1e-07 ***
## MomEd        0.00492    0.00475    1.03   0.3011
## Sib          0.01960    0.00609    3.22   0.0013 **
## BH          -0.07135    0.03338   -2.14   0.0327 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

19.4 Filtering to 8 Years of Experience

```
mod = lm(logwage~Educ+Abil+MomEd+Sib+BH,
         data = filter(data,PotExper==8))
coeftest(mod)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.31125    0.10674   12.28  < 2e-16 ***
## Educ         0.07307    0.00722   10.12  < 2e-16 ***
## Abil         0.09412    0.01655    5.69  1.6e-08 ***
## MomEd        0.00735    0.00462    1.59     0.11
## Sib          0.00598    0.00601    1.00     0.32
## BH          -0.04030    0.03365   -1.20     0.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

19.5 Including all 17,919 Observations

```
mod = lm(logwage~Educ+PotExper+Abil+MomEd+Sib+BH,data=data)
coeftest(mod)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.996458   0.033865   29.42   <2e-16 ***
## Educ         0.072183   0.002244   32.16   <2e-16 ***
## PotExper     0.039446   0.000899   43.89   <2e-16 ***
## Abil         0.080142   0.004886   16.40   <2e-16 ***
## MomEd        0.003931   0.001390    2.83   0.0047 **
## Sib          0.004582   0.001790    2.56   0.0105 *
## BH          -0.053655   0.009993   -5.37    8e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

19.6 Heteroskedasticity-robust clustered standard errors

- Does childhood environment impact wages

$$\ln(wage_i) = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 Abil_i + \beta_5 MomEd_i + \beta_6 Sib_i + \beta_7 BH_i + \varepsilon_i$$

- Data on same 2,178 interviewed annually, sample size = 17,919
- This is not a strictly random sample
- When we have repeated observations from similar sample units (people, firms, countries, cities), the disturbances are likely correlated (auto-correlation)
- We need to factor this in when calculating $\Sigma = \mathrm{E}(\boldsymbol{\varepsilon\varepsilon}')$
  - Some non-zero, off-diagonal elements of Σ

19.7 Constructing Clustered Standard Errors

- Suppose the data have G groups, i.e. 100 cities, with multiple observations per city
  - X_g is the block of X belonging to group g
  - e_g is the block of e belonging to group g

$$\mathrm{Var}(b|X) = (X'X)^{-1}\,\mathrm{E}(X'\varepsilon\varepsilon'X|X)(X'X)^{-1}$$

- Use sample to estimate E() with small sample degrees of freedom adjustment

$$\widehat{\mathrm{Var}(b|X)} = \frac{G}{G-1}\frac{n-1}{n-K}(X'X)^{-1}\left(\sum_{g=1}^{G} X_g' e_g e_g' X_g\right)(X'X)^{-1}$$

19.8 Club Sandwich

```
suppressMessages(library(clubSandwich))
#Cluster vcovHC with "CR0", "CR1", "CR1S", ... differenet DF adjustmer
ClusV = vcovCR(mod,data$ID,"CR1")
coeftest(mod,vcov=ClusV)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.99646    0.06932   14.37  < 2e-16 ***
## Educ         0.07218    0.00497   14.52  < 2e-16 ***
## PotExper     0.03945    0.00140   28.19  < 2e-16 ***
## Abil         0.08014    0.01140    7.03  2.2e-12 ***
## MomEd        0.00393    0.00320    1.23    0.219
## Sib          0.00458    0.00406    1.13    0.259
## BH          -0.05366    0.02162   -2.48    0.013 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

19.9 Joint Hypothesis Test

```
suppressMessages(library(AER))
linearHypothesis(mod,c("MomEd=0","Sib=0","BH=0"),vcov=ClusV)
```

```
## Linear hypothesis test
##
## Hypothesis:
## MomEd = 0
## Sib = 0
## BH = 0
##
## Model 1: restricted model
## Model 2: logwage ~ Educ + PotExper + Abil + MomEd + Sib + BH
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df    F Pr(>F)
## 1  17915
## 2  17912  3 2.66  0.047 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lecture 20

Model Specification

20.1 Model Specification

- Model specification describes which variables we choose to include in our regression
- In the end our estimates are just a summary of the data
  - How we specify our model centers around how we would like to summarize the data
  - What aspects of the data are most interesting
- We face a lot of choices
  1) Include things to reduce bias
  2) Include things to learn about the data
- Consider including quadratics, interactions, and categorical data

20.2 Example 1: Simple Linear Function

$$\boxed{y = \beta_0 + \beta_1 x + \varepsilon}$$

- What is the meaning of β_0?
  - β_0 is the value of y when $x =$?
  - It's always important to include a constant in a function. y rarely equals 0 when $x = 0$
- What is the meaning of β_1?
  - $\partial y/\partial x = \beta_1$ (Marginal Effect), change in y given change in x
  - *Simple Linear function assumes constant marginal effect*
- Plotting this function

20.3 Ex 2: Multivariate Linear Function

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon}$$

- What is the meaning of β_0?
- What is the meaning of β_1?
  - β_1 is the effect on x_1 on y holding fixed x_2
- Example

$$\boxed{Sales(quantity) = \beta_0 + \beta_1(Price) + \beta_2(Advertising) + \varepsilon}$$

- β_1 is the effect of price on sales, holding fixed advertising
- This model assumes Price and Advertising have a *constant* marginal effect on sales, i.e., $\partial Sales/\partial Price = \beta_1$

20.4 Ex 3: Quadratic Function

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon}$$

- What is the meaning of β_0?
- What is meaning of β_1?
  - Does it make since to change x_1 holding x_1^2 fixed? No
- *Quadratic function allows marginal effect of x_1 to depend on the level of x_1*
  - $\dfrac{\partial y}{\partial x_1} = \beta_1 + 2\beta_2 x_1$
- What is the meaning of β_2? β_2 describes the *change* in the marginal effect

20.5 Ex 3: Quadratic Function Cont.

- Ex: $Sales = \beta_0 + \beta_1(price) + \beta_2(price^2) + \varepsilon$
- Plot this function
- Show that the marginal effect of price depends on the level of price
- What is the interpretation of β_1?
  - Write new function so β_1 is interpreted as the marginal effect of going from $Price = \$100$ to $Price = \$101$

20.6 Ex 4: Functions with Interactions

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 x_2) + \varepsilon}$$

- What is the meaning of β_0?
- Show that the marginal effect of x_1 depends on x_2
  - $\partial y / \partial x_1 = \beta_1 + \beta_3 x_2$
- *Interactions allow marginal effects to depend on levels of other variables*
- What is β_3
  - The change in the marginal effect of x_1 on y given a one unit change in x_2

20.7 Ex 4: Functions with Interactions Cont.

Ex: $Sales = \beta_0 + \beta_1(Price) + \beta_2(Adv) + \beta_3(price \cdot Adv) + \varepsilon$

- Consider how we might plot this function
- Show that the marginal effect of *price* depends on *Adv*
- Show that the marginal effect of *Adv* depends on *price*

20.8 Ex 5: Functions with Non-Numeric Data

- With numeric data marginal (partial) effects are clear
- Categorical data, no partial effect
  - Effect of moving from some category to another
- Define indicator variable $x_1 = \begin{cases} 1 \\ 0 \end{cases}$

$$\boxed{y = \beta_0 + \beta_1 x_1 + \varepsilon}$$

- What is meaning of β_0? β_0 is $\mathrm{E}(y|x_1 = 0) = \beta_0$
- What is the meaning of β_1
  - $\mathrm{E}(y|x_1 = 1) = \beta_0 + \beta_1$
  - $\beta_1 = \mathrm{E}(y|x_1 = 1) - \mathrm{E}(y|x_1 = 0)$

20.9 Ex 5: Functions with Non-Numeric Data

- How does explanatory variable *monday* relate to dependent variable *sales*?
  - Effect of *monday* relative to $not - monday$
- Define indicator variable

$$x_1 = \begin{cases} 1 & \text{if Monday} \\ 0 & \text{if not Monday} \end{cases}$$

$$\boxed{sales = \beta_0 + \beta_1 x_1 + \varepsilon}$$

- What is meaning of β_0 and β_1?

20.10 Ex 6: Non-numeric Data: Multiple Categories

- Consider multiple indicator variables
- $x_1 = 1$ if Monday, 0 otherwise; $x_2 = 1$ if Tuesday, 0 otherwise; $\cdots$; $x_7 = 1$ if Sunday, 0 otherwise
- An undefined function (what is β_0?):

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon}$$

- A well defined function (now what is β_0?):

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_6 + \beta_6 x_7 + \varepsilon}$$

- Punchline: always need to omit one category, all partial effects are relative to omitted category

20.11 Cornwell and Rupert (1988), Labor Market Data, 595 Individuals, 7 years

- EXP =Work experience, WKS =Weeks worked, ED =Years of education
- OCC =Occupation, 1 if blue collar, IND =1 if manufacturing industry,
- SOUTH =1 if resides in south, SMSA =1 if resides in a city (SMSA),
- MS =1 if married,
- FEM =1 if female,
- UNION =1 if wage set by union contract,
- BLK =1 if individual is black,

20.12 Read Data

```
suppressMessages(library(AER))
suppressMessages(library(clubSandwich))
suppressMessages(library(equatiomatic))
suppressMessages(library(dplyr))
datasrc = "http://www.stern.nyu.edu/~wgreene/Text/Edition7/TableF8-1.cs
data = read.csv(datasrc)
numind = 595
sampyears = 1976:1982
data$year = kronecker(rep(1,numind),sampyears)
data$i = kronecker((1:numind),rep(1,length(sampyears)))
```

20.13 Data

```
data[1:15,]
```

```
##    EXP WKS OCC IND SOUTH SMSA MS FEM UNION ED BLK LWAGE year i
## 1    3  32   0   0     1    0  1   0     0  9   0 5.561 1976 1
## 2    4  43   0   0     1    0  1   0     0  9   0 5.720 1977 1
## 3    5  40   0   0     1    0  1   0     0  9   0 5.996 1978 1
## 4    6  39   0   0     1    0  1   0     0  9   0 5.996 1979 1
## 5    7  42   0   1     1    0  1   0     0  9   0 6.061 1980 1
## 6    8  35   0   1     1    0  1   0     0  9   0 6.174 1981 1
## 7    9  32   0   1     1    0  1   0     0  9   0 6.244 1982 1
## 8   30  34   1   0     0    0  1   0     0 11   0 6.163 1976 2
## 9   31  27   1   0     0    0  1   0     0 11   0 6.215 1977 2
## 10  32  33   1   1     0    0  1   0     1 11   0 6.263 1978 2
## 11  33  30   1   1     0    0  1   0     0 11   0 6.544 1979 2
## 12  34  30   1   1     0    0  1   0     0 11   0 6.697 1980 2
## 13  35  37   1   1     0    0  1   0     0 11   0 6.791 1981 2
## 14  36  30   1   1     0    0  1   0     0 11   0 6.816 1982 2
## 15   6  50   1   1     0    0  1   0     1 12   0 5.652 1976 3
```

20.14 Specification 1

```
mod_wage = lm(LWAGE~EXP + I(EXP^2) + WKS + OCC + IND +
                SOUTH + SMSA + MS + UNION + ED + FEM + year,
              data=data)
```

```
extract_eq(mod_wage,wrap = TRUE,terms_per_line = 4,intercept = "beta")
```

$$\begin{aligned}\text{LWAGE} = &\beta_0 + \beta_1(\text{EXP}) + \beta_2(\text{EXP\^{}2}) + \beta_3(\text{WKS}) + \\ &\beta_4(\text{OCC}) + \beta_5(\text{IND}) + \beta_6(\text{SOUTH}) + \beta_7(\text{SMSA}) + \\ &\beta_8(\text{MS}) + \beta_9(\text{UNION}) + \beta_{10}(\text{ED}) + \beta_{11}(\text{FEM}) + \\ &\beta_{12}(\text{year}) + \epsilon\end{aligned} \tag{1}$$

20.15 Specification 1

```
coeftest(mod_wage,vcov=vcovCR(mod_wage,data$i,"CR1")) %>%
  round(4)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -168.0256     3.9205  -42.86   <2e-16 ***
## EXP            0.0314     0.0042    7.52   <2e-16 ***
## I(EXP^2)      -0.0006     0.0001   -5.90   <2e-16 ***
## WKS            0.0040     0.0016    2.57   0.0102 *
## OCC           -0.1416     0.0268   -5.28   <2e-16 ***
## IND            0.0564     0.0234    2.41   0.0158 *
## SOUTH         -0.0719     0.0263   -2.74   0.0062 **
## SMSA           0.1540     0.0234    6.57   <2e-16 ***
## MS             0.0959     0.0430    2.23   0.0256 *
## UNION          0.0812     0.0233    3.49   0.0005 ***
## ED             0.0550     0.0056    9.91   <2e-16 ***
## FEM           -0.3652     0.0482   -7.57   <2e-16 ***
## year           0.0876     0.0020   44.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

20.16 Specification 2: Treating Year as Categorical

```
mod_wage2 = lm(LWAGE~EXP + I(EXP^2) + WKS + OCC + IND +
                SOUTH + SMSA + MS + UNION + ED + FEM + factor(year),
               data=data)
coeftest(mod_wage2,vcov=vcovCR(mod_wage2,data$i,"CR1")) %>%
  round(4)
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.0840     0.1297   39.19   <2e-16 ***
## EXP                 0.0313     0.0042    7.48   <2e-16 ***
## I(EXP^2)           -0.0006     0.0001   -5.87   <2e-16 ***
## WKS                 0.0039     0.0016    2.50   0.0124 *
## OCC                -0.1412     0.0268   -5.26   <2e-16 ***
## IND                 0.0566     0.0234    2.42   0.0155 *
## SOUTH              -0.0718     0.0263   -2.73   0.0063 **
## SMSA                0.1542     0.0234    6.58   <2e-16 ***
## MS                  0.0963     0.0429    2.24   0.0248 *
## UNION               0.0805     0.0233    3.46   0.0006 ***
## ED                  0.0550     0.0056    9.90   <2e-16 ***
## FEM                -0.3650     0.0482   -7.57   <2e-16 ***
## factor(year)1977    0.0746     0.0060   12.44   <2e-16 ***
## factor(year)1978    0.1961     0.0099   19.86   <2e-16 ***
## factor(year)1979    0.2836     0.0101   27.96   <2e-16 ***
## factor(year)1980    0.3626     0.0098   36.89   <2e-16 ***
## factor(year)1981    0.4369     0.0113   38.65   <2e-16 ***
## factor(year)1982    0.5208     0.0121   43.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

20.17 Specification 3: Discontinuity in Education

```
mod_wage_ed = lm(LWAGE~EXP + OCC + IND +
                   SOUTH + SMSA + MS + UNION + ED +
                   I(ED>=13) + I((ED>=13)*(ED-13)) + FEM + factor(year)
                 data=data)
coeftest(mod_wage_ed,vcov=vcovCR(mod_wage_ed,data$i,"CR1")) %>%
  round(4)
```

```
##
## t test of coefficients:
##
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.5008     0.1215   45.29   <2e-16 ***
## EXP                            0.0065     0.0012    5.19   <2e-16 ***
## OCC                           -0.1420     0.0271   -5.24   <2e-16 ***
## IND                            0.0628     0.0235    2.68   0.0075 **
## SOUTH                         -0.0723     0.0276   -2.62   0.0088 **
## SMSA                           0.1537     0.0239    6.43   <2e-16 ***
## MS                             0.1120     0.0441    2.54   0.0112 *
## UNION                          0.0785     0.0237    3.31   0.0009 ***
## ED                             0.0503     0.0090    5.59   <2e-16 ***
## I(ED >= 13)TRUE               -0.0240     0.0455   -0.53   0.5983
## I((ED >= 13) * (ED - 13))      0.0176     0.0164    1.08   0.2824
## FEM                           -0.3598     0.0504   -7.14   <2e-16 ***
## factor(year)1977               0.0832     0.0059   14.11   <2e-16 ***
## factor(year)1978               0.2094     0.0095   21.95   <2e-16 ***
## factor(year)1979               0.3009     0.0101   29.84   <2e-16 ***
## factor(year)1980               0.3813     0.0099   38.65   <2e-16 ***
## factor(year)1981               0.4559     0.0116   39.47   <2e-16 ***
## factor(year)1982               0.5388     0.0125   43.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

20.18 Specification 4: Fully Interacted by Gender

```
mod_wage_gender = lm(LWAGE~(EXP + OCC + IND +
                            SOUTH + SMSA + MS + UNION + ED + year)*FE
                     data=data)
coeftest(mod_wage_gender,vcov=vcovCR(mod_wage_gender,data$i,"CR1")) %>%
  round(4)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -173.0524     4.2671  -40.56   <2e-16 ***
## EXP            0.0072     0.0013    5.47   <2e-16 ***
## OCC           -0.1232     0.0295   -4.18   <2e-16 ***
## IND            0.0530     0.0246    2.16   0.0310 *
## SOUTH         -0.0711     0.0301   -2.36   0.0182 *
## SMSA           0.1551     0.0253    6.14   <2e-16 ***
## MS             0.1105     0.0463    2.39   0.0169 *
## UNION          0.0705     0.0248    2.84   0.0046 **
## ED             0.0547     0.0061    8.99   <2e-16 ***
## year           0.0903     0.0022   41.66   <2e-16 ***
## FEM           -1.4913    12.7401   -0.12   0.9068
## EXP:FEM       -0.0046     0.0038   -1.22   0.2243
## OCC:FEM       -0.1798     0.0617   -2.92   0.0036 **
## IND:FEM        0.1874     0.0736    2.54   0.0110 *
## SOUTH:FEM      0.0061     0.0621    0.10   0.9214
## SMSA:FEM      -0.0063     0.0648   -0.10   0.9221
## MS:FEM         0.0670     0.0761    0.88   0.3789
## UNION:FEM      0.0270     0.0737    0.37   0.7142
## ED:FEM         0.0054     0.0138    0.39   0.6950
## year:FEM       0.0006     0.0065    0.09   0.9267
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lecture 21

Fixed Effects

21.1 Causal Effects with LRM

- Relying on the linear regression model for causal effects is difficult
  - Our causal interpretation relies solely on if the assumption of mean independence
- We cannot be certain if mean independence holds
  - But we can hopefully be certain that it *plausibly* holds

21.2 Issue with OLS when Assumptions Fail

- Consider two cities: (1) With high unemployment, low average education, and high crime rates, (2) Low unemployment, high education, and low crime rates
- LRM: $crmrte_{it} = \beta_1 + \beta_2 unem_{it} + \beta_3 educ_{it} + \varepsilon_{it}$
- OLS *identifies* an effect of unemployment on crime that is not attributed to *educ*
  - But estimated effect *could be* attributed to unobserved differences *that are not unemployment*
- OLS can only hold fixed things included in the model, but does not hold fixed things *not* included

21.3 Causal Effect with Regression

- To make regression work we need to include all relevant variables in the model
  - Including them is the only way to hold them fixed
- Solution #1: Enumerate all relevant variables and include them in the model
- Solution #2: Estimate model using only *within* subject variation
  - Holds fixed all observed and *unobserved* differences between subjects
  - Fixed effect estimator

21.4 Fixed Effect Model

- With repeated observations on the same sampling units (e.g., person, firm, state, neighborhood, county, industry, etc.)
- Include a a full set of indicator variables for the groups (fixed effects)
  - Captures all unmeasured and unnamed effects for each individual (super controls)

$$crmrte_{it} = \beta_1 + \beta_2 unem_{it} + \alpha_i + \delta_t + \varepsilon_{it}$$

- Treats "city" as a category and estimates the effect of unemployment holding city characteristics fixed
  - α_i, all the things that determine *crmrte* in city i except *unemp*
- I added time fixed effects too

21.5 Crime Rate Data

```
suppressMessages(library(dplyr))
load(file.path(rdata_loc,'crime2.RData'))

data %>%
  select(pop,crimes,unem,officers,year,area) %>%
  slice_head(n=10)
```

```
##       pop crimes unem officers year  area
## 1  229528  17136  8.2      326   82  44.6
## 2  246815  17306  3.7      321   87  44.6
## 3  814054  75654  8.1     1621   82 375.0
## 4  933177  83960  5.4     1803   87 375.0
## 5  374974  31352  9.0      633   82  49.8
## 6  406297  31364  5.9      685   87  49.8
## 7  176496  15698 12.6      245   82  74.0
## 8  201723  16953  5.7      259   87  74.0
## 9  288446  31202 12.6      504   82  97.3
## 10 331728  34355  7.4      563   87  97.3
```

```
num_cities = nrow(data)/2
data$city = rep(1:num_cities,each=2)
```

21.6 Biased: Effect of Unemployment Rate on Crime (crimes per 1000 people)

```
mod1 = lm(crmrte~unem,data=filter(data,year==87))
summary(mod1)
```

```
##
## Call:
## lm(formula = crmrte ~ unem, data = filter(data, year == 87))
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -57.5  -27.0  -10.6   18.0   79.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   128.38      20.76    6.18  1.8e-07 ***
## unem           -4.16       3.42   -1.22     0.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.6 on 44 degrees of freedom
## Multiple R-squared:  0.0326, Adjusted R-squared:  0.0106
## F-statistic: 1.48 on 1 and 44 DF,  p-value: 0.23
```

21.7 FE Estimator Has Indicator for Each Group

```
mod2 = lm(crmrte~unem + factor(city) + factor(year),data=data)
mod2
```

```
##
## Call:
## lm(formula = crmrte ~ unem + factor(city) + factor(year), data = dat
##
## Coefficients:
##    (Intercept)            unem   factor(city)2   factor(city)3
##        51.4892          2.2180         17.2917          4.6885
##  factor(city)4   factor(city)5   factor(city)6   factor(city)7
##         7.0067         24.4979         43.5656          2.5462
##  factor(city)8   factor(city)9  factor(city)10  factor(city)11
##        12.0272        -10.2511         14.6022         -2.0704
## factor(city)12  factor(city)13  factor(city)14  factor(city)15
##        28.4537         94.9551          8.4767         32.4837
## factor(city)16  factor(city)17  factor(city)18  factor(city)19
##        69.4379         72.7238         18.4954         77.4585
## factor(city)20  factor(city)21  factor(city)22  factor(city)23
##        67.7533         -8.2399         -6.8375          8.1835
## factor(city)24  factor(city)25  factor(city)26  factor(city)27
##        37.0960         40.6148         12.6237         32.4070
## factor(city)28  factor(city)29  factor(city)30  factor(city)31
##        36.0383         -8.3650        -15.7475         27.9807
## factor(city)32  factor(city)33  factor(city)34  factor(city)35
##         0.0859          4.5692         13.8534         33.8508
## factor(city)36  factor(city)37  factor(city)38  factor(city)39
##        36.9232         71.4389         -4.2835         32.0111
## factor(city)40  factor(city)41  factor(city)42  factor(city)43
##         8.7139         67.3832         70.7641         44.3814
## factor(city)44  factor(city)45  factor(city)46  factor(year)87
##         1.7220        -17.7531         -3.2771         15.4022
```

21.8 fixest Package in R

```
suppressMessages(library(fixest))
mod3 = feols(crmrte~unem |city + year,data=data)
summary(mod3)
```

```
## OLS estimation, Dep. Var.: crmrte
## Observations: 92
## Fixed-effects: city: 46,  year: 2
## Standard-errors: Clustered (city)
##      Estimate Std. Error t value  Pr(>|t|)
## unem    2.218     0.8154    2.72 0.0092419 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 9.80504     Adj. R2: 0.774292
##                 Within R2: 0.1267
```

21.9 Important Identification Result with FE

- Consider the fixed effects model

$$y_{it} = x_{it}'\beta + \alpha_i + \varepsilon_{it}$$

- Estimates of β are identified from "within" variation
  - e.g., how unemployment relates to crime in a city and then averages over cities
  - Compare "between" estimator relationship is identified comparing different cities with different crime and unemployment
  - OLS (w/o FE) identifies "total" variation ("between" plus "within")

21.10 Identification of FE Model

- Data organization for FE model

$$Y = X\beta + D\alpha + \boldsymbol{\varepsilon}$$

$$D = \begin{bmatrix} \mathbb{1}_1 & 0 & 0 & \cdots & 0 \\ 0 & \mathbb{1}_2 & 0 & \cdots & 0 \\ 0 & 0 & \mathbb{1}_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbb{1}_n \end{bmatrix}$$

- D is a matrix of indicator variables
  - If the first group has 5 observations then $\mathbb{1}_1$ is a column vector of 1's

21.11 Identification of FE Model

$$Y = X\beta + D\alpha + \boldsymbol{\varepsilon}$$

- Partial out D

$$\begin{aligned} \mathbf{M}_D Y =& \mathbf{M}_D X\beta + \mathbf{M}_D D\alpha + \boldsymbol{\varepsilon} \\ Y^* =& X^*\beta + \boldsymbol{\varepsilon} \\ (y_{it} - \overline{y}_i) =& (x_{it} - \overline{x}_i)'\beta + \varepsilon_{it} \end{aligned}$$

- The best predictor of y only knowing the group is the sample average of the group

21.12 Corollaries of FE Idendtification

- *Corollary #1*: Any groups without multiple observations (i.e., have only 1 observation) will effectively not be used in estimation of β
  - $y_{it} - \overline{y}_i = 0$ and $(x_{it} - \overline{x}_i) = 0$, so $e = 0 - 0'b = 0$ for all values of b does not contribute to SSE
- *Corollary #2*: Any variables in x_{it} that are fixed (the same value) over time cannot be include, $(x_{it} - \overline{x}_i) = 0$, violates full rank

21.13 Estimating Fixed Effects Models

- Approach #1: Make the matrix D and apply least squares

$$Y = X\beta + D\alpha + \boldsymbol{\varepsilon}$$

  - Requires large matrix inversion. Manageable with moderate number of fixed effects

- Approach #2: Partial the fixed effects out
  - Subtract the group mean from all variables

$$(y_{it} - \overline{y}_i) = (x_{it} - \overline{x}_i)'\beta + \varepsilon_{it}$$

  - More common with large number of fixed effects

- Use clustered standard errors. With multiple fixed effects, cluster variable with largest number of groups

Lecture 22

Linear Probability Model

22.1 Linear Probability Model

- What if y is a discrete variable?
  - e.g., Belongs to union ($y = 1$), does not belong to a union ($y = 0$)
  - $\mathrm{E}(y) = \Pr(\text{Belongs to Union})$
- Linear regression model with discrete y
  - $\mathrm{E}(y|x) = \Pr(y = 1|x) = \beta_0 + \beta_1 x$
    - $100 \times \beta_1$ is percentage point (pp) change in $\Pr(y = 1)$ from one unit increase in x
- Example: $\widehat{UNION} = 0.962 - 0.047ED$
  - Increasing ED by one unit decreases the probability of belonging to a union by 4.7 pp.

Lecture 23

Statistics Concepts for MLE

23.1 Review

- Definitions (and types) of random variables
- Probabilities and probability density functions
- Parametric distributions
- Multivariate random variables

23.2 Discrete Random Variables (univariate)

- *A discrete random variable* is a variable that takes a countable number of values with certain probabilities
- X takes J possible values $x_1, x_2, \ldots, x_J$
- $\Pr(X = x_j) = \pi_j$
- π_j is a probability, lies in 0 to 1, i.e., $\pi_j \in [0, 1]$
- The probabilities must sum to 1, i.e., $\sum_{j=1}^{J} \pi_j = 1$
- Examples: coin toss, die roll, employment status, purchase decision, firm entry, college major, occupational choice, healthcare status, school enrollment

23.3 Continuous Random Variables (univariate)

- *A continuous random variable* is a variable that takes values from a continuous range
- The outcomes of a continuous random variable are defined by its *probability density function* (PDF), $f(x)$, such that

$$\Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

- Properties of $f(x)$
  - Defined over the whole number line, i.e. $\int_{-\infty}^{\infty} f(x)dx = 1$
  - Non-negative, i.e., $f(x) \geq 0$ for all x
  - Is not a "probability" (could be greater than 1)

23.4 Continuous Random Variables

- Examples: In nature, tons. In datasets, technically none (all variables need to be rounded to store)
- Often we assume a random variable is continuous. Why?

1) It is convenient
   - While technically not continuous could take a large number of values

- Represent the likelihood of particular value with simple function/small number of parameters

2) It is reasonable

- Even if a variable is not technically continuous, it very closely resembles a continuous variable

23.5 Continuous Random Variables

- It is useful to define a random variables Cumulative Distribution Function (CDF)

$$F(z) = \Pr(X \leq z) = \int_{-\infty}^{z} f(x)dx$$

- $F(z) \in [0, 1]$ for all values of z
- Can be used to construct outcome probabilities

$$\begin{aligned}\Pr(a \leq X \leq b) &= \int_{a}^{b} f(x)dx \\ &= \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx \\ &= F(b) - F(a)\end{aligned}$$

23.6 Central to Maximum Likelihood Estimation

- Working with probabilities (PDFs, CDFs) is central to MLE
- We will mostly focus on *parametric distributions*
  - i.e., π_j or $f(x)$ can be described by a parametric function (function described by a finite set of parameters)
  - as opposed to non-parametric distributions
- For example if $X \sim N(\mu, \sigma^2)$ then

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

23.7 Example: Poisson (discrete)

- If X follows a Poisson distribution then we write $X \sim \text{Pois}(\lambda)$
- $X \in \mathbb{N}_0$, that is takes any natural number starting at zero, i.e., $0, 1, 2, 3, \ldots$
- Even though X includes infinity it is still discrete because it is countable
- $\Pr(X = k) = \dfrac{\lambda^k \exp(-\lambda)}{k!}$
  - If $\lambda = 1$ then $\Pr(X = 2) \approx 0.1839$

23.8 Multivariate Random Variables

- Multivariate Discrete
  - Joint Probability; $\Pr(A, B)$
  - Conditional probability; $\Pr(A|B)$
- Multivariate Continuous
  - Joint Density; $f(x_1, x_2)$
  - Conditional Distribution; $f(x_1|x_2)$

23.9 Bi-Variate Normal

- Two random variables, X_1 and X_2
  - $X_1 \sim N(\mu_1, \sigma_1^2)$
  - $X_2 \sim N(\mu_2, \sigma_2^2)$
  - $\text{Cov}(X_1, X_2) = \sigma_{12}$
- Define $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}'$ a vector of random variables

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right)$$

- $f(x_1|x_2)$; conditional distribution of x_1 knowing x_2
- $x_1|x_2 \sim N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2) \, , \, \sigma_1^2 - \frac{(\sigma_{12})^2}{\sigma_2^2} \right)$

23.10 Multi-Variate Normal

- X is $K \times 1$ vector of multivariate normal
  - $X \sim N(\mu, \Sigma)$
- μ is $K \times 1$ (vector of means)
- Σ is $K \times K$ (variance-covariance matrix)
- Joint density

$$f(x) = |2\pi\Sigma|^{-1/2} \exp\left(-1/2(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

  - $|\cdot|$ is the determinant

23.11 Conditional Density of Multi-Variate Normal

- X is $K \times 1$ vector of multivariate normal, $X \sim N(\mu, \Sigma)$
- Partition X into X_1 and X_2
  - $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with sizes $\begin{bmatrix} q \times 1 \\ (K-q) \times 1 \end{bmatrix}$
- Partition μ and Σ
  - $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$
- $f(X_1|X_2)$; conditional distribution of X_1 knowing X_2
- $X_1|X_2 \sim N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)\,,\, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$

Lecture 24

Introduction to Maximum Likelihood Estimation

24.1 Overview

- Introduction / Motivation
- Review Some Concepts
- Formalize the Maximum Likelihood Estimator
- Simple Maximum Likelihood Examples

24.2 Distributions

- Typically, with distributions we say something like ...
  - If $X \sim N(\mu, \sigma^2)$, what are the values of X we are most likely to observe
- Maximum Likelihood Estimation (MLE) is concerned with inverting this relationship
  - Observing X AND *knowing* that X is normally distributed, what is most likely value for μ, σ^2
  - We perform estimation by
    - Using objects that describe the distribution of X given μ, σ^2; synonymously provides the distribution (likelihood) of μ,σ^2 given X
    - Finding the maximum of this likelihood

24.3 Modeling and Estimation

- Econometric Model: A set of assumptions (restrictions) on the joint distribution of the variables (observed and unobserved)
- Estimation: Choosing a joint a distribution among many (possibly infinite) that could have generated the data
- Maximum likelihood estimation
  - Make full set of assumptions on the distribution of the data
    - Fully parametric: Distributional assumptions on the unobserved (errors) of the model
  - Find the value of the parameters that most likely generated the data

24.4 Review Properties of Estimators

- Unbiasedness: In finite sample $\mathrm{E}(\hat{\theta}) = \theta$
  - Outside least squares not central issue
  - More focus on asymptotic unbiasedness (consistency)

- Consistency: Bias and variance go to zero as $n \to \infty$
  - Most important property
- Asymptotic normality
- Asymptotic efficiency

24.5 An Important Class of Estimators

- *Extremum Estimators*: An estimator obtained as the optimizer (min or max) of a criterion functions, e.g.,
$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, q(\theta|\text{data})$$
- Under general conditions *Extremum Estimators* are *consistent*
- *M-estimators* are an important extremum estimator
  - Criterion function is a sample average
  - Under general conditions M-estimators are asymptotically normal and efficient
- Least squares and maximum likelihood are *M-estimators*

24.6 Principle of Maximum Likelihood

- Econometric model describes how data is generated from a set of underlying parameters
- The likelihood of the parameters given the observed data will be a continuous function
- Use computational methods to find the parameters that maximize this likelihood
- MLE is *Consistent* (extremum estimator) *Efficient and Asymptotically normal* (M-estimator)
- Draw back (assumes a correctly specified model)
  - None of these things if model is incorrect
  - Efficiency at the cost of robustness

24.7 Constructing the Likelihood

- Observed data with sample size n
- Because the observed data contains lots of data points, the likelihood is a *joint probability/-density*
- For continuous random variables
  - Joint Density; $f(x_1, x_2) = f(x_1|x_2)f(x_2)$
- Independence; $f(x_1|x_2) = f(x_1)$
  - $f(x_1, x_2) = f(x_1)f(x_2)$
- Thus joint distribution n independent observations

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) =& f(x_1) f(x_2) \cdots f(x_n) \\ =& \prod_{i=1}^{n} f(x_i) \end{aligned}$$

24.8 Example: Time to Quit First Job

- Data: Years employed in first job (6 observations)
  - $\begin{bmatrix} 4.1 & 2.4 & 0.12 & 0.9 & 0.31 & 0.1 \end{bmatrix}$
- Assume exponential function (used with rate data)
  - $f(x_i|\lambda) = \lambda e^{-\lambda x_i}$
- The Likelihood (of λ given the data)

$$\begin{aligned} L(\lambda) =& f(x_1, x_2, x_3, x_4, x_5, x_6|\lambda) \\ =& f(x_1|\lambda) f(x_2|\lambda) f(x_3|\lambda) f(x_4|\lambda) f(x_5|\lambda) f(x_6|\lambda) \\ =& \left(\lambda e^{-\lambda 4.1}\right) \left(\lambda e^{-\lambda 2.4}\right) \left(\lambda e^{-\lambda 0.12}\right) \times \cdots \\ & \left(\lambda e^{-\lambda 0.9}\right) \left(\lambda e^{-\lambda 0.31}\right) \left(\lambda e^{-\lambda 0.1}\right) \end{aligned}$$

24.9 Maximizing the Likelihood

- $L(\lambda)$ is a continuous function of λ

$$\begin{aligned} L(\lambda) =& \left(\lambda e^{-\lambda 4.1}\right) \left(\lambda e^{-\lambda 2.4}\right) \left(\lambda e^{-\lambda 0.12}\right) \times \cdots \\ & \left(\lambda e^{-\lambda 0.9}\right) \left(\lambda e^{-\lambda 0.31}\right) \left(\lambda e^{-\lambda 0.1}\right) \end{aligned}$$

- The maximum of any monotonic transformation of $L(\lambda)$ is also the maximum of $L(\lambda)$
- Log-likelihood function $LL(\lambda) = \ln L(\lambda)$

$$\begin{aligned} LL(\lambda) =& \ln \left[\left(\lambda e^{-\lambda 4.1}\right) \left(\lambda e^{-\lambda 2.4}\right) \left(\lambda e^{-\lambda 0.12}\right) \times \cdots \right. \\ & \left. \left(\lambda e^{-\lambda 0.9}\right) \left(\lambda e^{-\lambda 0.31}\right) \left(\lambda e^{-\lambda 0.1}\right) \right] \end{aligned}$$

- This expression simplifies drastically where most of the products get replaced by sums because $\ln(a \times b) = \ln(a) + \ln(b)$

24.10 Why Maximize the Log-Likelihood Instead of Likelihood

1) Feasibility:
   - Likelihood is the product of many probabilities. Multiplying many small numbers, the computer will eventually round down to zero
   - Taking the log of numbers less than one leads to large negative numbers

- Thus, instead of working with product of small numbers we work with sum of large numbers

2) Simplicity:

- Derivatives of sums are much easier and avoids extensive application of the product rule

24.11 Maximizing the Log-Likelihood

- In general. Define the data X_1, X_2, …, X_n
  - Define all of the parameters θ

$$\begin{aligned} LL(\theta) = \ln L(\theta) &= \ln \left[f(X_1, X_2, \ldots, X_n | \theta) \right] \\ &= \ln \left[\prod_{i=1}^{n} f(X_i | \theta) \right] \\ &= \sum_{i=1}^{n} \ln \left(f(X_i | \theta) \right) \end{aligned}$$

- Maximize the log-likelihood

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \ln \left(f(X_i | \theta) \right)$$

24.12 Maximizing the Log-Likelihood In Practice

- Write a function that takes as input a vector of parameters and returns the log-likelihood value (a scalar)
- Send this function to built in **minimizer** `optim`
  - Min. of negative of function is equal to Max.
  - RE-WRITE your function to return the negative of log-likelihood
- Specialized algorithms to find maximum
- Important. ALWAYS use `sum(log(like))`
  - NEVER use `log(prod(like))`

24.13 Ex: Time to Quit First Job

```
x = c(4.1,2.4,0.12,0.9,0.31,0.1)

#likelihood calculation
calclike = function(lam) lam*exp(-lam*x)

#evaluating the likelihood at at Lambda = 1
prod(calclike(1))
```

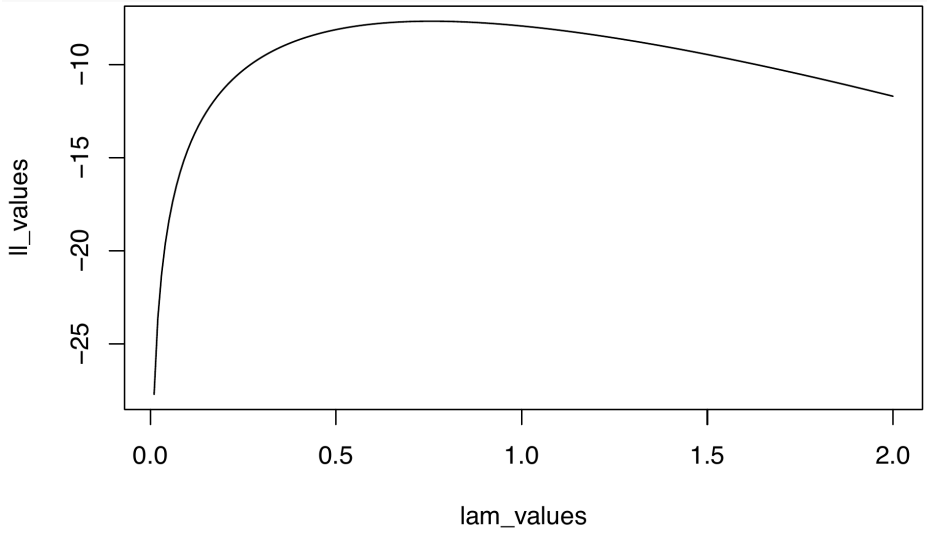
```
## [1] 0.0003598
```

```
#evaluating the log-likelihood at at Lambda = 1
sum(log(calclike(1)))
```

```
## [1] -7.93
```

24.14 Ex: Time to Quit First Job

```
llfun = function(lam) sum(log(calclike(lam)))
lam_values = seq(.01,2,by=.01)
ll_values = sapply(lam_values,llfun)
plot(lam_values,ll_values,'l')
```



24.15 Ex: Time to Quit First Job

```
fun2min = function(lam) -sum(log(calclike(lam)))
#This is the function that finds the minimum
#Need to supply initial guess (1) to the function
mle_res = optim(1,fun2min)
lam_hat = mle_res$par
sprintf('MLE of LAMBDA %0.5f',lam_hat)
```

[1] "MLE of LAMBDA 0.75664"

```
sprintf('Expected no. of years in first job %0.7f',1/lam_hat)
```

[1] "Expected no. of years in first job 1.3216314"

```
sprintf('or %0.7f',mean(x))
```

[1] "or 1.3216667"

Lecture 25

Methods and Practice of Numerical Optimization

25.1 Introduction

- Recall: The ordinary least squares estimator was derived by minimizing the sum of squared errors
  - We did not need specialized algorithms to find the minimum of this function
  - It was a simple *quadratic function* and we used calculus and algebra to write a 'closed-form' solution
- Maximum likelihood estimation requires optimizing (maximizing) much more complicated functions
  - We need specialized algorithms to solve these problems
  - This is the field of *numerical optimization*

25.2 What are the Model Parameters

- In OLS, we had two sets of parameters β and σ^2
  - Estimated separately with different equation
- In MLE, we estimate everything together
  - $\theta = \{\beta, \sigma^2\}$
- In general we say θ is $K \times 1$ vector of parameters
  - θ can become quite complicated
- Example, suppose we estimated 3 linear regression models
  - $\theta = \{\beta_1, \sigma_1^2, \beta_2, \sigma_2^2, \beta_3, \sigma_3^2, \sigma_{12}, \sigma_{13}, \sigma_{23}\}$

25.3 Numerical Optimization

- Our Situation: Maximizing $LL(\theta)$ directly is difficult
  - But, evaluating $LL(\theta)$ for a given θ is possible
- All numerical optimization entails iterative algorithms
  - We begin with initial guess $\theta^{(0)}$
  - Generate a sequence $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$
- Goal is to algorithmically choose new points so:

$$LL(\theta^{(m+1)}) > LL(\theta^{(m)})$$

- Need *stopping criteria*; definition of convergence
- A *fast algorithm* requires fewest evaluations of $LL(\theta)$

25.4 Approaches to Numerical Optimization

- Two broad approaches to choosing search direction. Going from $\theta^{(m)}$ to $\theta^{(m+1)}$

1) Gradient (derivative) based methods. Fast, more precise for smooth functions
   - If θ is a scalar (bisection)
   - If θ is multidimensional (Newton based methods)
2) Non-gradient based methods. More robust
   - Nelder-mead

25.5 Numerical Optimization Comments

1) Local versus Global optimization
   - Need to re-run at different starting values to ensure global maximum
2) Sometimes elements of θ are constrained by economic or statistical models
   - Constrained optimization *can be* very slow
   - Alternative re-parameterize and use unconstrained optimizer
3) Researcher coding time to implement
   - Faster optimizers require additional problem specific coding. An unfortunate trade-off between researcher time and computer time

25.6 Gradient Based Methods

- The derivative of the Log-Likelihood is called the gradient
- If θ is $K \times 1$ then there are K partial derivatives

$$\mathrm{g}(\theta) = \begin{bmatrix} \frac{\partial LL(\theta))}{\partial \theta_1} \\ \frac{\partial LL(\theta))}{\partial \theta_2} \\ \vdots \\ \frac{\partial LL(\theta))}{\partial \theta_K} \end{bmatrix}_{K \times 1}$$

- The gradient plays an important role. Because if $\hat{\theta}$ maximizes $LL(\theta)$, then

$$\mathrm{g}(\hat{\theta}) = \mathbf{0}_{K \times 1}$$

25.7 How to Calculate the Gradient?

- We can either derive and code up the gradient (first derivative) ourselves or calculate it numerically
- Given $f(z)$, wish to calculate $f'(z_0)$
  - Numerical derivative $f'(z_0) \approx \frac{f(z_1) - f(z_0)}{z_1 - z_0}$
- Downside of numerical derivatives is computing time is proportional to the number of parameters
  - If model has K parameters, calculating the gradient requires K calls to the log-likelihood EACH iteration
  - If algorithm requires 1000 iteration then need $1000 \times K$ evaluation of log-likelihood

25.8 Attempting to Get Most Accurate Numerical Derivative Possible

- Numerical derivative $\frac{f(z_1) - f(z_0)}{z_1 - z_0}$
- Issues:
  - If z_1 is far from z_0 then derivative is inaccurate
  - If z_1 is too close to z_0 then computer rounding makes derivative is *VERY* inaccurate
- Best Practice
  - $\triangle z = \sqrt{\text{machine.precision}} * \text{sign}(z_0) * \max(|z_0|, 1)$
  - $z_1 = z_0 + \triangle z$
  - $\frac{f(z_1) - f(z_0)}{z_1 - z_0}$. Divide by actual difference not $\triangle z$

25.9 Generic Numerical Derivative Function

```
numderiv = function(fun,x){
  signx = if (x==0) 1 else sign(x)
  h = sqrt(.Machine$double.eps)*(signx*max(abs(x),1))
  x1 = x + h
  g = (fun(x1) - fun(x))/(x1 - x)
  return(g)
}
```

25.10 Bisection Algorithm (univariate optimization)

- Choose θ_1 and θ_2 that bracket the mode. i.e., $g(\theta_1) > 0$ and $g(\theta_2) < 0$
- Iterate
  - Calculate $\overline{\theta} = (\theta_1 + \theta_2)/2$

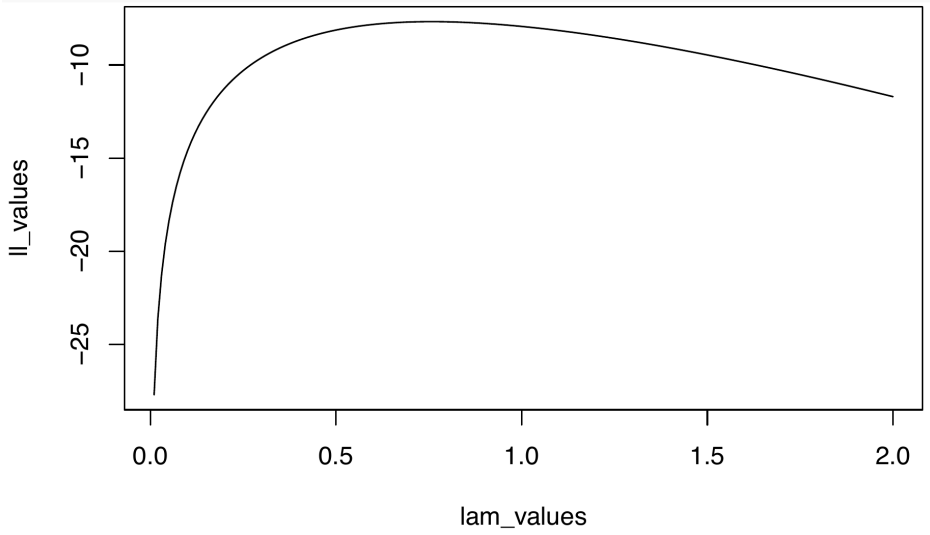
- Calculate $g(\overline{\theta})$
- IF $g(\overline{\theta}) \approx 0$ QUIT, OTHERWISE continue
- IF $g(\overline{\theta}) > 0$ then replace $\theta_1 = \overline{\theta}$, OTHERWISE $\theta_2 = \overline{\theta}$

25.11 Example: Time to Quit First Job

- Data: Years employed in first job (6 observations)
  - $\begin{bmatrix} 4.1 & 2.4 & 0.12 & 0.9 & 0.31 & 0.1 \end{bmatrix}$
- Assume exponential function (used with rate data)
  - $f(x_i|\lambda) = \lambda e^{-\lambda x_i}$
- CDF Probability quit before x years $F(x) = 1 - e^{-\lambda x}$

25.12 Ex: Time to Quit First Job

```
x = c(4.1,2.4,0.12,0.9,0.31,0.1)

#log likelihood function
llfun = function(lam) sum(log(lam*exp(-lam*x)))

lam_values = seq(.01,2,by=.01)
ll_values = sapply(lam_values,llfun)
plot(lam_values,ll_values,'l')
```



25.13

```
# Starting values
lam = c(.1,2)
# check gradient at starting values
c(numderiv(llfun,lam[1]),numderiv(llfun,lam[2]))
```

```
## [1] 52.07 -4.93
```

```
for (numIter in 1:1e6){
  lam_hat = mean(lam)
  gr_hat = numderiv(llfun,lam_hat)
  if (abs(gr_hat)<1e-6) break
  if (gr_hat>0){
    lam[1] = lam_hat
  } else {
    lam[2] = lam_hat
  }
}
sprintf('MLE of LAMBDA %0.5f',lam_hat)
```

```
## [1] "MLE of LAMBDA 0.75662"
```

```
sprintf('LL value %0.12f',llfun(lam_hat))
```

```
## [1] "LL value -7.673361398512"
```

```
sprintf('Number of Iterations %d',numIter)
```

```
## [1] "Number of Iterations 24"
```

25.14 Determining Convergence

- The decision to stop the algorithm and declare convergence is up to the researcher
- Some choices (m is the iteration count)
  - Gradient close to zero $\mathrm{g}(\theta^{(m)}) \approx 0$
  - Small change in the parameters $\theta^{(m)} \approx \theta^{(m+1)}$
  - Small change in log-likelihood $LL(\theta^{(m)}) \approx LL(\theta^{(m+1)})$
- Convergence is any of the above or some combination
- Example: $\max(abs(\theta^{(m)} - \theta^{(m+1)})) < 1e-6$

25.15 Newton-Rhapson

- If θ is multi-dimensional cannot use bisection
- Since $LL(\theta)$ is a very complicated function, construct a simpler function that approximates LL
  - i.e., $Q(\theta) = A + B\theta$
- Make $Q(\theta)$ a good approximation near $\theta^{(0)}$
  - Should have the same slope, so $B = \mathrm{g}(\theta^{(0)})$

- Should be equal at $\theta^{(0)}$, so $A = LL(\theta^{(0)}) - B\theta^{(0)}$

$$\begin{aligned} Q(\theta) =& A + B\theta \\ =& LL(\theta^{(0)}) - B\theta^{(0)} + B\theta \\ =& LL(\theta^{(0)}) + B(\theta - \theta^{(0)}) \\ =& LL(\theta^{(0)}) + \mathrm{g}(\theta^{(0)})(\theta - \theta^{(0)}) \end{aligned}$$

25.16 Taylor Series Approximation to LL

- Given parameter values $\theta^{(0)}$, a *first order Taylor series approximation* tells us the *direction* to move θ to improve the log-likelihood

$$Q(\theta) = LL(\theta^{(0)}) + \mathrm{g}(\theta^{(0)})(\theta - \theta^{(0)})$$

  - If $\mathrm{g}(\theta^{(0)})$ is *positive* then increase θ
  - If $\mathrm{g}(\theta^{(0)})$ is *negative* then decrease θ
- Does not tell us how far

25.17 Taylor Series Approximation to LL

- *Second order Taylor series approximation*

$$Q(\theta) = LL(\theta^{(0)}) + \mathrm{g}(\theta^{(0)})'(\theta - \theta^{(0)}) + \frac{1}{2}(\theta - \theta^{(0)})'\mathrm{H}(\theta^{(0)})(\theta - \theta^{(0)})$$

- H is called the Hessian (matrix of second derivatives)
- This function resembles $LL(\theta)$, but $Q(\theta)$ can be maximized in closed form. It's a quadratic function
- Strategy
  - Treat $Q(\theta)$ as surrogate function. Maximize $Q(\theta)$ instead of $LL(\theta)$
  - This gets you a $\theta^{(1)}$ that is closer to the maximum of $LL(\theta)$ than $\theta^{(0)}$
  - Create a new surrogate function $Q(\theta)$ at $\theta^{(1)}$ and repeat until you reach $\hat{\theta}$ that maximizes $LL(\theta)$

25.18 Maximizing the Quadratic Approximation

$$Q(\theta) = LL(\theta^{(0)}) + \mathrm{g}(\theta^{(0)})'(\theta - \theta^{(0)}) + \frac{1}{2}(\theta - \theta^{(0)})'\mathrm{H}(\theta^{(0)})(\theta - \theta^{(0)})$$

- Anything that involves $\theta^{(0)}$ is a constant (the equation is mostly constants)

$$\begin{aligned} \partial Q(\theta)/\partial\theta =& \mathrm{g}(\theta^{(0)}) + \mathrm{H}(\theta^{(0)})(\theta - \theta^{(0)}) \\ 0 =& \mathrm{g}(\theta^{(0)}) + \mathrm{H}(\theta^{(0)})(\theta^{(1)} - \theta^{(0)}) \\ -\mathrm{H}(\theta^{(0)})(\theta^{(1)} - \theta^{(0)}) =& \mathrm{g}(\theta^{(0)}) \\ (\theta^{(1)} - \theta^{(0)}) =& -\mathrm{H}(\theta^{(0)})^{-1}\mathrm{g}(\theta^{(0)}) \\ \theta^{(1)} =& \theta^{(0)} - \mathrm{H}(\theta^{(0)})^{-1}\mathrm{g}(\theta^{(0)}) \end{aligned}$$

- This approach is based on Newton's method
- The step-size is $-\mathrm{H}(\theta^{(0)})^{-1}\mathrm{g}(\theta^{(0)})$

25.19 Newton-Rhapson

- Algorithm
  - Begin with $\theta^{(m)}$
  - Calculate $\mathrm{g}(\theta^{(m)})$ and $\mathrm{H}(\theta^{(m)})$
  - Update $\theta^{(m+1)} = \theta^{(m)} - \mathrm{H}(\theta^{(m)})^{-1}\mathrm{g}(\theta^{(m)})$
  - Repeat until $\mathrm{g}(\theta^{(m+1)}) \approx 0$
- Issues
  - May never converge
  - Using math to derive formulas for $\mathrm{g}(\theta)$ and $\mathrm{H}(\theta)$ may take a lot of researcher time
  - Using computer to derive $\mathrm{H}(\theta)$ may take a very long time

25.20 Quasi-Newton with Numerical Derivative

- Use numerical derivative for $\mathrm{g}(\theta)$
- Calculating $\mathrm{H}(\theta)$ is not necessary. A variety of ways to approximate it (quasi-newton methods)
  - The most common is BFGS
- In practice additional parameter λ to modify search direction
  - $\theta^{(m+1)} = \theta^{(m)} - \lambda\mathrm{H}(\theta^{(m)})^{-1}\mathrm{g}(\theta^{(m)})$
  - Choose λ to ensure $LL(\theta^{(m+1)}) > LL(\theta^{(m)})$

25.21 Numerical Optimization In Practice

- In general will use built-in optimizers, but need to be familiar how optimizers work so can successfully debug and make adjustments to the algorithm
- `optim` is general unconstrained optimizer

```
optim(par, fn, gr = NULL, ...,
      method = c("Nelder-Mead", "BFGS", "CG", "L-BFGS-B", "SANN",
                 "Brent"),
      lower = -Inf, upper = Inf,
      control = list(), hessian = FALSE)
```

  - 'gr' argument to supply the actual gradient (otherwise uses numerical derivatives)
  - 'method' choose optimizer
  - 'control' set max. iterations, convergence criteria, etc.

25.22 Optim

```
mle_res = optim(1,
                function(lam) -llfun(lam),
                method='BFGS',
                control = list(maxit=1e6,abstol=1e-6))

lam_hat = mle_res$par

sprintf('MLE of LAMBDA %0.5f',lam_hat)
```

```
## [1] "MLE of LAMBDA 0.75662"
```

```
sprintf('LL value %0.12f',-mle_res$value)
```

```
## [1] "LL value -7.673361398516"
```

```
sprintf('Number of Iterations %d',mle_res$counts[['function']])
```

```
## [1] "Number of Iterations 22"
```

Lecture 26

Likelihood Ratio Test

26.1 Comment

- It is unusual to discuss hypothesis testing *BEFORE* learning the sampling properties of MLE
- Hypothesis testing is about if the data is consistent with a particular *restriction* on the parameters
  - i.e. $H_0 : \theta = 0$
- If $LL(\theta = 0)$ is really far from to $LL(\hat{\theta}_{MLE})$, then this tells us $\theta = 0$ is very *unlikely* (reject the null)
- Key Takeaway: The shape of the log-likelihood function reveals the range of plausible values. How precise our estimate is
  - This fact will be important when we discuss the sampling properties.

26.2 Revisit Example: Time to Quit First Job

- Data: Years employed in first job (6 observations)
  - $\begin{bmatrix} 4.1 & 2.4 & 0.12 & 0.9 & 0.31 & 0.1 \end{bmatrix}$
- Assume exponential function $f(x_i|\lambda) = \lambda e^{-\lambda x_i}$
- $\widehat{\lambda}_{MLE} = 0.7566204$
- In what sense is this a precise estimate?
  - i.e., It is possible that $\lambda = 0.5$ and observe 0.7566204 due to randomness
- One of the earliest forms of hypothesis testing: likelihood ratio test
  - If $LL(0.5)$ is a lot smaller than $LL(\widehat{\lambda}_{MLE})$ then this offers lots of evidence against $\lambda = 0.5$

26.3 Likelihood Ratio Test (on MLE parameter θ)

- Consider the hypothesis that $H_0 : \theta = \overline{\theta}$
- Then the test statistic

$$\begin{aligned} LR =& -2 \ln \left(\frac{L(\overline{\theta})}{L(\widehat{\theta}_{MLE})} \right) \\ =& -2 \left[\ln L(\overline{\theta}) - \ln L(\widehat{\theta}_{MLE}) \right] \\ =& -2 \left[LL(\overline{\theta}) - LL(\widehat{\theta}_{MLE}) \right] \\ =& 2 \left[LL(\widehat{\theta}_{MLE}) - LL(\overline{\theta}) \right] \sim \chi^2 [1] \end{aligned}$$

- Thus if the likelihood is very flat, lots of $\overline{\theta}$ will have similar LL value to $LL(\widehat{\theta}_{MLE})$ (c.f. LL very peaked)
  - Can't reject these many null. Our estimate is not very precise

26.4

```
x = c(4.1,2.4,0.12,0.9,0.31,0.1)

#log likelihood function
llfun = function(lam) sum(log(lam*exp(-lam*x)))
mle_res = optim(1,function(lam) -llfun(lam),method='BFGS',
                control = list(maxit=1e6,abstol=1e-6))
lam_hat = mle_res$par
sprintf('MLE of LAMBDA %0.5f',lam_hat)
```

```
## [1] "MLE of LAMBDA 0.75662"
```

```
sprintf('LL value at MLE %0.6f',llfun(lam_hat))
```

```
## [1] "LL value at MLE -7.673361"
```

```
sprintf('LL value at Hypothesis %0.6f',llfun(0.5))
```

```
## [1] "LL value at Hypothesis -8.123883"
```

```
lr_test = 2*(llfun(lam_hat)-llfun(0.5))
sprintf('LR test statistic %0.4f',lr_test)
```

```
## [1] "LR test statistic 0.9010"
```

```
sprintf('p-value for LR test %0.4f',1-pchisq(lr_test,1))
```

```
## [1] "p-value for LR test 0.3425"
```

Lecture 27

Sampling Properties of the Maximum Likelihood Estimator

27.1 Sampling Properties of MLE

- MLE is an extremum estimator: *It is consistent*
- MLE is also and M-estimator: *It is efficient and normally distributed*
- Both of these properties are about *large samples*
- Todo: How to calculate the variance-covariance matrix of parameter estimates?
  - Given an estimate of the variance-covariance matrix we can conduct inference
  - Standard errors, confidence interval, hypothesis testing, Wald test, etc.
- All inference is based on *normal* distribution because of large sample application

27.2 Recall Likelihood Ratio Test

- If the likelihood is very FLAT, lots of alternative θ values will have similar LL value to $LL(\widehat{\theta}_{MLE})$
  - Can't reject these many null. Our estimate is not very precise
- Alternatively, If the likelihood is very PEAKED, few alternative θ values will be as compelling as $\widehat{\theta}_{MLE}$
  - Our estimate is very precise
- We measure the precision of our estimate with the Standard Error
- *This indicates the standard error of the MLE estimate will be related to the shape of the log-likelihood function*

27.3 Standard Errors (precision) of the Maximum Likelihood Estimator

- If the log-likelihood is very peaked at the maximum
  - The estimate is very precise (small standard errors) and we are confident that we are close to the solution
- If the log-likelihood is very flat, then estimator is very imprecise (large standard errors)
  - Can't rule out many other plausible values
- Thus *standard errors will be determined by peakedness or flatness of the log-likelihood function*
  - What describes the peakedness or flatness of a function?

27.4 Standard Errors (precision) of the Maximum Likelihood Estimator

- The standard errors will be determined by peakedness or flatness of the log-likelihood function
- The second derivative of a function describes the rate of change of a function slope
- A very large (in absolute value) second derivative describes a 'highly peaked' function
- *The standard error of MLE is INVERSELY related to the second derivative of the log-likelihood function*

27.5 Features of the Log-Likelihood

- The Log-Likelihood of θ ($K \times 1$ vector of parameters):

$$\begin{aligned} LL(\theta) &= \sum_{i=1}^{n} \ln\left(f(y_i|\theta)\right) \\ &= \ln\left(f(y_1|\theta)\right) + \ln\left(f(y_2|\theta)\right) + \cdots + \ln\left(f(y_n|\theta)\right) \end{aligned}$$

- Derivatives of the Log-Likelihood
  - Gradient /Score ($K \times 1$ vector)

$$\mathrm{g}(\theta) = \frac{\partial \ln\left(f(y_1|\theta)\right)}{\partial \theta} + \cdots + \frac{\partial \ln\left(f(y_n|\theta)\right)}{\partial \theta}$$

  - Hessian ($K \times K$ matrix)

$$\mathrm{H}(\theta) = \frac{\partial^2 \ln\left(f(y_1|\theta)\right)}{\partial \theta \partial \theta'} + \cdots + \frac{\partial^2 \ln\left(f(y_n|\theta)\right)}{\partial \theta \partial \theta'}$$

27.6 Properties of MLE (θ_0 stands for true (unknown value of the parameters)

1) Consistency: plim $\hat{\theta} = \theta_0$
   - Possibly biased in small sample, e.g. variance divided by n instead of $n - K$
2) Asymptotic Normality: $\hat{\theta} \sim N\left[\theta_0, \{I(\theta_0)\}^{-1}\right]$
   - $I(\theta_0)$ is information matrix
   - $I(\theta_0) = -\,\mathrm{E}_0\left[\partial^2 \ln L / \partial \theta_0 \partial \theta_0'\right]$
3) Efficiency: Achieves Cramer-Rao Lower bound, in class of consistent asymptotically normal distributions
   - Efficiency at cost of robustness

27.7 Estimation of MLE

- Point Estimate
  - $\hat{\theta} = \text{argmax}_\theta \, LL(\theta)$
  - Root finding: $\text{g}(\hat{\theta}) = 0$
- Variance Covariance of Estimates
  - Theoretical $\text{Var}(\hat{\theta}) = \{I(\theta_0)\}^{-1} = \{-\text{E}_0 \left[\text{H}(\theta_0)\right]\}^{-1}$
  - Estimate: $\widehat{\text{Var}(\theta)} = -\left\{\sum_{i=1}^{n} \text{H}_i(\hat{\theta})\right\}^{-1}$

27.8 Constructing Standard Errors. Ask `optim` to give you the Hessian

```
x = c(4.1,2.4,0.12,0.9,0.31,0.1)

#log likelihood function
llfun = function(lam) sum(log(lam*exp(-lam*x)))
mle_res = optim(1,function(lam) -llfun(lam),method='BFGS',
                hessian = TRUE,
                control = list(maxit=1e6,abstol=1e-6))
lam_hat = mle_res$par
#NOTE: we don't need a negative because llfun is already negated
lam_vcov = solve(mle_res$hessian)

lam_se = sqrt(diag(lam_vcov))

sprintf('MLE of LAMBDA %0.5f',lam_hat)
```

```
## [1] "MLE of LAMBDA 0.75662"
```

```
sprintf('SE of MLE of LAMBDA %0.5f',lam_se)
```

```
## [1] "SE of MLE of LAMBDA 0.30889"
```

27.9 Hypothesis Testing

- We have an estimate of the parameters and the variance covariance matrix. Everything is normally distributed
- We can test the null hypothesis for a single parameter using z-stat
  - $H_0 : \theta_k = \theta_0$
  - z-stat: $Z = \dfrac{\hat{\theta}_k - \theta_0}{\text{SE}(\hat{\theta}_k)} \sim N(0,1)$
  - DO NOT use t-distribution like we do with OLS
- Critical value for 1% two-sided test: `qnorm((1-.01/2))`
- P-value for test: `2*(1-pnorm(abs(Z)))`

27.10 Hypothesis Testing: Wald Test

- $H_0 : R\theta = q$
  - R is a $J \times \dim(\theta)$ restriction matrix
  - q is $\dim(\theta) \times 1$ hypothesized values
- Define $m = R\hat{\theta} - q$
- $\text{Var}(m) = R\,\text{Var}(\hat{\theta})R'$
- $m'\,\text{Var}(m)^{-1}m \sim \chi[J]$

Lecture 28

Bootstrapping and Resampling Methods

28.1 Getting Standard Errors

- For a consistent estimator, $\hat{\theta} \to \theta_0$, but need to know variance-covariance matrix to conduct inference
  - Variance covariance matrix comes from sampling properties of the estimator
- As our estimators become more complex (and problem specific), deriving the theoretical variance-covariance matrix becomes challenging
- Would like to conduct inference for a broad set of estimators

28.2 Getting Standard Errors

- In many cases ...
  1) ... we can only estimate the model and don't know how to calculate the variance-covariance matrix
  2) ... there are too many second derivatives making the hessian infeasible to calculate or invert
  3) ... we might want standard errors that are functions of the parameters
  4) ... we might want more robust inference than offered by theoretical forumlas
- In these cases we can use re-sampling methods to get the standard errors

28.3 Re-sampling Methods

- Two classes of re-sampling methods (with weird names from the 1950's)
  - Bootstrap and jackknife
  - Only requirement to get standard errors is the model can be estimated
- Principle: use the observed data to create pseudo-data set that mimics sampling variability present in the data
  - Re-estimate the model on many pseudo-datasets
  - If pseudo-datasets mimic sampling variability, then distribution of estimates will reflect sampling distribution

28.4 Comments on Re-sampling Methods

- We can really conduct inference on anything
- Standard errors based on re-sampling methods tend to be larger than theoretical formulas, but potentially more credible
- We won't cover Jackknife, it is less common

  - Delete 1-jackknife is my favorite resampling method but has bad small sample properties

- Lots (and lots) of bootstrap strategies
- We will only cover Non-parametric bootstrapping
- Re-sampling methods can also be used to estimate the bias in consistent estimators (not OVB)

28.5 Bootstrapping Standard Errors

- What is an estimators sampling distribution?
  - $\hat{\theta}$ is an estimate (a random variable)
  - If our estimator consistent AND normally distributed, then with sufficient sample size $\hat{\theta} \sim N(\theta, ?)$
- Suppose we have 1,000 samples and for each sample we compute $\hat{\theta}_r$, for $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_{1000}$
  - We can calculate the variance of our sampling distribution directly
  - $\widehat{\text{Var}(\hat{\theta})} = \frac{1}{1000-1} \sum_{r=1}^{1000} (\hat{\theta}_r - \overline{\hat{\theta}})^2$
    - Where $\overline{\hat{\theta}} = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\theta}_r$

28.6 Bootstrapping Standard Errors

- Don't have 1,000 different samples, we have 1. Generate new estimation samples (pseudo-data) to mimic the sampling distribution
- Non-parametric bootstrap (most general approach)
  - Idea: If our one sample is a random draw from the population, then a sample of our sample is also a random draw from the population

1) Sample *with* replacement 1,000 times
2) Estimate the model on the 1,000 samples
3) Calculate variance-covariance of sample estimates
   - Or any other inference object

28.7 Example: OLS Estimator

```
suppressMessages(library(lmtest))
load(file.path(rdata_loc,'nbasal.RData'))
mod = lm(points~age+exper+expersq+coll+center+forward,data=data)
coeftest(mod)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.5206     6.9420    5.12  6.0e-07 ***
## age          -1.0438     0.2933   -3.56  0.00044 ***
## exper         2.2829     0.4043    5.65  4.2e-08 ***
## expersq      -0.0716     0.0235   -3.05  0.00252 **
## coll         -1.3376     0.4485   -2.98  0.00313 **
## center       -2.2966     0.9673   -2.37  0.01831 *
## forward      -0.8219     0.7352   -1.12  0.26463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

28.8 Non-Parametric Bootstrap

```
num_boot = 1000
n = nrow(data)
b_vals = matrix(0,nrow=num_boot,ncol=length(mod$coefficients))
for (s in 1:num_boot){
  boot_idx = sample(1:n,replace=TRUE)
  boot_data = data[boot_idx,]
  boot_mod = lm(points~age+exper+expersq+coll+center+forward,
                data=boot_data)
  b_vals[s,] = boot_mod$coefficients
}
vcov_bs = cov(b_vals)
sqrt(diag(vcov_bs))
```

```
## [1] 7.80008 0.33286 0.42290 0.02765 0.43491 1.20894 0.65936
```

28.9 Non-Parametric Bootstrap

```
coeftest(mod,vcov=vcov_bs)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.5206     7.8001    4.55  8.1e-06 ***
## age          -1.0438     0.3329   -3.14   0.0019 **
## exper         2.2829     0.4229    5.40  1.5e-07 ***
## expersq      -0.0716     0.0276   -2.59   0.0102 *
## coll         -1.3376     0.4349   -3.08   0.0023 **
## center       -2.2966     1.2089   -1.90   0.0586 .
## forward      -0.8219     0.6594   -1.25   0.2137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

28.10 Using 10,000 Samples

```
coeftest(mod,vcov=vcov_bs)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.5206     7.6747    4.63  5.8e-06 ***
## age          -1.0438     0.3277   -3.19   0.0016 **
## exper         2.2829     0.4218    5.41  1.4e-07 ***
## expersq      -0.0716     0.0279   -2.57   0.0108 *
## coll         -1.3376     0.4380   -3.05   0.0025 **
## center       -2.2966     1.2158   -1.89   0.0600 .
## forward      -0.8219     0.6607   -1.24   0.2146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

28.11 Confidence Intervals

```
conf_bs = apply(b_vals,2,function(x) quantile(x,c(0.025,0.975)))
cbind(confint(mod),t(conf_bs))
```

```
##               2.5 %   97.5 %   2.5%   97.5%
## (Intercept) 21.8515 49.18982 20.644 50.7721
## age         -1.6214 -0.46621 -1.700 -0.4127
## exper        1.4868  3.07892  1.547  3.2022
## expersq     -0.1178 -0.02537 -0.134 -0.0266
## coll        -2.2208 -0.45445 -2.249 -0.5462
## center      -4.2014 -0.39186 -4.652  0.1442
## forward     -2.2696  0.62579 -2.123  0.4665
```

Lecture 29

Structural Equation Modeling (also sometimes referred to as factor models)

29.1 Two Ways to Use Data to Make Decisions

1) What is the effect of x on y

   "If I new how x impacted y I could make great decisions"

   – Use data to estimate parameters

     * e.g., OLS, $y = \beta_0 + \beta_1 x$

2) *I KNOW* x has a big impact on y but *I DON'T KNOW* x. x is difficult to observe/measure

   "If I only new what x was I could make great decisions"

- Structural Equation Modeling: Use the observed data to *infer IMPORTANT unobserved data*

29.2 Sometimes We Use Observed Data to Proxy Things We Don't Observe

- Players with lots of points and rebounds are superstar basketball player
- Students that get lots of questions correct on an exam know the material well
- Days with long wait times are crowded days at Disnyeland
- *We are seeking a more scientific way to quantify these *latent* variables (superstar basketball player, knowledgable students, crowded days*

29.3 Latent Variable Estimation

- Latent variables are variables that are extremely intuitive, important, but difficult to measure
  - Someone's IQ, How much someone likes travel, How good someone is at basketball
- The latent variable effects observed variables
  - e.x. People that like to travel spend a lot of money on vacations
- Structural equation modeling uses disparate measures to estimate latent variables
  - Feed in a lot of data: days traveled, travel location, money spent on travel, etc. to estimate preferences for travel

29.4 Example: Use NBA data to identify player quality

- Let θ_i denote the quality of player i
- Observed y_{i1} ($points_i$) and y_{i2} ($assists_i$)
- How to use y_{i1} and y_{i2} to figure out θ_i?
- Makeshift approaches

  - Points=ability: $\theta_i = y_{i1}$
  - Arbitrary weighted average: $\theta_i = (3/4)y_{i1} + (1/5)y_{i2}$
- Structural equation modeling uses maximum likelihood to optimally create $\mathrm{E}(\theta_i | y_{i1}, y_{i2})$

29.5 Model

- Assume $\theta_i \sim N(\gamma, \delta)$
- y_{i1}, y_{i2} are noisy measures of θ_i

$$\begin{aligned} y_{i1} &= \beta_1 + \lambda_1 \theta_i + \varepsilon_{i1} \\ y_{i2} &= \beta_2 + \lambda_2 \theta_i + \varepsilon_{i2} \end{aligned}$$

- λ is the loading (how important θ_i is for y_{ij}
- ε_{i1} and ε_{i2} is the noise
  - The component of points that is unrelated to our measure of player quality
  - Assume $\mathrm{Var}(\varepsilon_{ij}) \sim N(0, \sigma_j^2)$
- *Important: y is dissimilar data so all parameters are specific to y_{i1} and y_{i2}*

29.6 Identification Issue #1

- Location of θ_i (γ) is not identifiable
- Expected value of points

$$\begin{aligned} \mathrm{E}(y_{i1}) &= \mathrm{E}(\beta_1 + \lambda_1 \theta_i + \varepsilon_{i1}) \\ &= \beta_1 + \lambda_1 \mathrm{E}(\theta_i) + \mathrm{E}(\varepsilon_{i1}) \\ &= \beta_1 + \lambda_1 \gamma \end{aligned}$$

- Observe two things ($\mathrm{E}(y_{i1})$ and $\mathrm{E}(y_{i2})$) but need identify three parameters β_1, β_2 and γ
  - Need to normalize one (set to zero)
- Any is ok, but we will set $\gamma = 0$
  - Good interpretation: positive value of θ denote above the mean and negative value below

29.7 Identification Issue #2

- Scale of θ_i (δ^2) is not identifiable
- Covariance of points and assists

$$\mathrm{Cov}(y_{i1}, y_{i2}) = \ldots = \lambda_1 \lambda_2 \delta^2$$

- Need to normalize (set equal to 1) λ_1, λ_2, OR δ^2
- Any is ok, but we will set $\delta^2 = 1$
  - Good interpretation: θ_i will be a standard normal (i.e., 2 will be a large value)

29.8 Identification Issue #3

- Normalizing $\delta^2 = 1$ need to identify λ_1 and λ_2
  - But only observed $\text{Cov}(y_{i1}, y_{i2}) = \lambda_1 * \lambda_2 * 1$
  - One equation and TWO unknowns!
- Could assume $\lambda_2 = \lambda_1$ (this defeats the purpose of handling dissimilar data)
- We need another measure y_{i3} ($rebounds_i$)

$$y_{i3} = \beta_3 + \lambda_3\theta_i + \varepsilon_{i3}$$

- Now three equations and three unknowns
  - $\text{Cov}(y_{i1}, y_{i2}) = \lambda_1\lambda_2$, $\text{Cov}(y_{i1}, y_{i3}) = \lambda_1\lambda_3$, $\text{Cov}(y_{i2}, y_{i3}) = \lambda_2\lambda_3$

29.9 Generalizing the Framework

- Observe J (at least 3) important measures y_{i1}, y_{i2}, ..., y_{iJ} for each i
  - Define data for i $Y_i = \begin{bmatrix} y_{i1} & y_{i2} & \dots & y_{iJ} \end{bmatrix}'$
- Model parameters
  - $\beta = \begin{bmatrix} \beta_1 & \beta_2 & \dots & \beta_J \end{bmatrix}'$
  - $\Lambda = \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_J \end{bmatrix}'$
  - $\Sigma = diag(\begin{bmatrix} \sigma_1^2 & \sigma_2^2 & \dots & \sigma_J^2 \end{bmatrix})$
- Unobserved
  - $\theta_i \sim N(\gamma, \delta^2)$
  - $\varepsilon_i = \begin{bmatrix} \varepsilon_{i1} & \varepsilon_{i2} & \dots & \varepsilon_{iJ} \end{bmatrix}'$

29.10 Joint Distribution of the Observed Data

- $Y_i = \beta + \Lambda\theta_i + \varepsilon_i$ ($J \times 1$ multivariate normal)
- To construct likelihood need mean and covariance

$$\begin{aligned} \text{E}(Y_i) &= \text{E}(\beta + \Lambda\theta_i + \varepsilon_i) \\ &= \beta + \Lambda\,\text{E}(\theta_i) + \text{E}(\varepsilon_i) \\ &= \beta + \Lambda\gamma \\ \text{Var}(Y_i) &= \text{Var}(\beta + \Lambda\theta_i + \varepsilon_i) \\ &= \text{Var}(\Lambda\theta_i) + \text{Var}(\varepsilon_i) \\ &= \Lambda\,\text{Var}(\theta_i)\Lambda' + \text{Var}(\varepsilon_i) \\ &= \Lambda\delta^2\Lambda' + \Sigma \end{aligned}$$

- You need to impose the constraints that you choose for your model

29.11 Maximum Likelihood Estimation

- Need the PDF of multivariate Normal

$$f(Y_i|M,V) = |2\pi V|^{-1/2} \exp\left(-1/2(Y_i - M)'V^{-1}(Y_i - M)\right)$$

- $LL = \sum_{i=1}^{n} \ln(f(Y_i))$
- The model parameters: $\beta_1, \beta_2, \beta_3, \lambda_1, \lambda_2, \lambda_3, \sigma_1^2, \sigma_2^2, \sigma_3^2$
- From assumption ($\gamma = 0$ and $\delta^2 = 1$):
  - $\mathrm{E}(Y_i) = M = \beta$
  - $\mathrm{Var}(Y_i) = V = \Lambda\Lambda' + \Sigma$
- Notice $\sigma^2 > 0$ so we need to consider this constraint in estimation

29.12 Read in Data

```
load(file.path(rdata_loc,'nbasal.RData'))
data[1:10,c('wage','exper','points','assists','rebounds','draft')]
```

```
##    wage exper points assists rebounds draft
## 1  1002     4   15.5     4.5      3.9    19
## 2  2030     5   13.3     8.8      2.5    28
## 3   650     1    5.5     0.2      3.3    19
## 4  2030     5    7.3     1.5      5.1     1
## 5   755     3   10.8     2.6      4.3    24
## 6  2014     9   11.3     1.5      4.9     4
## 7  1065     1   15.1     1.4      7.2    40
## 8   420     3    6.6     0.7      4.2    47
## 9   150     1    3.1     2.0      0.7    NA
## 10 3050    12   26.0     2.3      6.5     3
```

29.13 Set-up Maximum Likelihood

```
# set up the data
Y = as.matrix(data[,c('points','assists','rebounds')])

J = ncol(Y)

# Starting values (just assumes lambda positive number)
# this is vector of parameter: beta, lambda, sigma2
# notice I am taking log of the variance
sv = c(apply(Y,2,mean),
       c(3,1,1),
       log(apply(Y,2,var)))
```

29.14 Log-Likelihood Function

```
#Use the mvtnorm library to calculate the pdf
library(mvtnorm)

llfn = function(parms){
  # split the parameters
  #first J are betas
  BETA = matrix(parms[1:J],ncol=1)

  #second J are lambdas
  LAM = matrix(parms[(J+1):(2*J)],ncol=1)

  #last J are log of variance, need exponentiation
  SIGMA = diag(exp(parms[(2*J+1):(3*J)]))

  M = BETA
  V = LAM %*% t(LAM) + SIGMA
  ll = sum(log(dmvnorm(Y,M,V)))
}
```

29.15 Estimation

```
ml = optim(sv,function(x) -llfn(x),
           method='BFGS',hessian = TRUE,
           control = list('maxit'=1e6,abstol=1e-6))

ml_ll = llfn(ml$par)
sprintf('%0.4f',ml_ll)
```

```
## [1] "-2007.9451"
```

```
ml_est = matrix(ml$par,nrow=J)
ml_est[,3] = exp(ml_est[,3])
ml_est
```

```
##        [,1]  [,2]    [,3]
## [1,] 10.206 5.898 0.02999
## [2,]  2.408 1.129 3.09339
## [3,]  4.397 1.631 5.68049
```

```
ml_vcov = solve(ml$hessian)
ml_se = matrix(sqrt(diag(ml_vcov)),nrow=J)
```

29.16 Posterior Distribution ($\mathrm{E}(\theta_i|Y_i)$)

- Given ML estimates $\hat{\beta}$, $\hat{\Lambda}$, $\hat{\gamma}$, $\hat{\delta}^2$, $\hat{\Sigma}$

$$\begin{bmatrix} Y_i \\ \theta_i \end{bmatrix} \sim N\left(\begin{bmatrix} \beta + \Lambda\gamma \\ \gamma \end{bmatrix}, \begin{bmatrix} \Lambda\delta^2\Lambda' + \Sigma & \Lambda\delta^2 \\ \delta^2\Lambda' & \delta^2 \end{bmatrix}\right)$$

$$\mathrm{E}(\theta_i|Y_i) = \gamma + (\delta^2\Lambda')(\Lambda\delta^2\Lambda' + \Sigma)^{-1}(Y_i - (\beta + \Lambda\gamma))$$

- Note: if θ_i was observed we could just estimate

$$\hat{\theta}_i = \hat{\alpha}_0 + \hat{\alpha}_1 y_{i1} + \hat{\alpha}_2 y_{i2} + \hat{\alpha}_3 y_{i3}$$

  - This would be identical to the formula above (but we didn't observe θ and solved the problem anyway!)

29.17 Discussion

- Structural Equation Modeling can be extended well beyond our simple example, but is outside of the scope of this class
  - Y could stem from lots of data types
  - θ could be mulitdimensional (e.g., offensive skill and defensive skill)
- Structural Equation Modeling is equivalent to confirmatory factor analysis
  - Factor analysis is a data reduction technique
  - Economists should be better at this than statisticians because we should have better hypothesis about the theory of the larger underlying mechanisms