Math 455 Class 15 February 23

Midterm 1: Monday March 2 in class

no books, notes, or calculators.

Note: Friday Feb 27th is a holiday.

No office hours Thursday Feb 26

or Tuesday March 3.

Class Wednesday March 4 will be a pre-recorded lecture where I discuss solutions to Midterm 1.

If you are still having tech issues with Gradescope, please see me!

Solutions to problems in the Faraway text should be produced using RMarkdown.

Why? Solutions to these problems involve presenting tables and graphs, plus the code that produced them. This is what RMd is for!

Proto-questions for Midterm.

Q: What's the difference between $E[\varepsilon|X] = 0$ and $Cov(\varepsilon, X) = 0$?

A: First note that $E[\varepsilon|X] = 0$ implies $Cov(\varepsilon, X) = 0$.

Tools: Defn of Covariance

Law Iterated Expectation. (Wackerly Ch. 5)

$$E[E[\varepsilon|X]] = E[\varepsilon]$$

$$Cov(\varepsilon, X) = E[\varepsilon X] - E[\varepsilon]E[X]$$

If $E[\varepsilon|X] = 0$, then $E[\varepsilon] = 0$.

So $Cov(\varepsilon, X) = E[\varepsilon X]$.

By Law of It. Exp. $E[\varepsilon X] = E[E[\varepsilon X|X]]$

In the context of linear regression X is not a RV.

So $E[\varepsilon X] = E[X\, E[\varepsilon | X]]$

$= E[X \cdot 0] = 0.$

Thus $Cov(\varepsilon, X) = 0 - 0 = 0.$

In other words $E[\varepsilon | X] = 0 \Rightarrow Cov(\varepsilon, X) = 0.$

This is similar to the result guaranteed by OLS:

$Cov(e, X) = 0.$

Note that e = residuals $\neq \varepsilon$ = errors.

e can be regarded as an estimate of the realized value of ε.

Hierarchy of Conditions:

(1) Independence of ε and X, plus $E[\varepsilon] = 0$

implies:

(2) $E[\varepsilon | X] = 0.$

implies

(3) $Cov(\varepsilon, X) = 0.$

Why think about $E[\varepsilon|X]$?

Q: Why think about conditional expectation?

A: Linear Regression is a conditional model.

It is a model for the distribution of $Y|X$.

It makes no assumption about the distribution of X itself: X is not even random.

Note that $Cov(\varepsilon, X) = 0$ <u>does not</u> <u>Imply</u> that $E[\varepsilon|X] = 0$.

Example: let $X = -1, 0, 1$ each w prob $1/3$.

Let $Y = X^2 - \frac{2}{3}$.

We claim $Cov(Y, X) = 0$.

But $E[Y|X] \neq 0$, for example

$E[Y|X=0] = -\frac{2}{3}$.

Note that an example for $E[\varepsilon|X] = 0$

Does not imply indep. can be found in Wackerly. Ch. 3 (?).

More on proto-questions for midterm.

We discussed 3 senses in which linear regression is optimal.

(1) OLS estimators coincide with MLE..

(2) Gauss-Markov. Theorem.

(3) Linear regression is orthogonal projection.

Proto-question for (3):

orthogonal projection from what vector space to what other vector space?

orthogonal projection is a linear map, and so corresponds to a matrix. What is that matrix?

More about (1): this requires the assumption that the errors are Normal.

PDF of Normal Dist.

$$f(\mu, \sigma; x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

What is the likelihood function?

It is the same function as the PDF.

BUT: We regard it as a function of the parameters, not x.

This means: the likelihood function is NOT a probability density function of the parameters.

Because a PDF needs to integrate to 1.

OK: What good is a likelihood function?

A: It is used in the method of maximum likelihood.

That is, if we wish to produce estimators for our parameters, given a bunch of data, we can write down the likelihood function (which has our data in it) and ask, what values of μ, σ. ~~maximi~~ (or whatever

parameters are in question), maximize the function?

These values $\hat{\mu}, \hat{\sigma}$ are called the maximum likelihood estimators of μ, σ.

Q: In principle, with our Linear Regression

A: Model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

and our data:

X	$x_1 \dots x_n$
Y	$Y_1 \dots Y_n$

Q: We could write down a likelihood function and try to find $\hat{\beta}_0, \hat{\beta}_1$ maximizing the function.

A: (Note that our data are treated as known constants.)

> This method yields exactly the least-squares estimators.

Q: Why is this true?

A: Because linear regression assumes:

- independence of errors ε_i
- normality of errors: $\varepsilon_i \sim N(0, \sigma^2)$.

Q: How do we write down the likelihood function?

A: Use independence, which implies PDF is a product.

$\varepsilon_i \sim N(0, \sigma)$ $\quad \varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$

PDF of ε_i: $\frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right)$

Joint PDF of $\varepsilon_1, \ldots, \varepsilon_n$: is a product.

of $\frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right)$ (from $i=1$ to n)

How do we get from maximizing

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right)$$

to least squares?

2 standard reductions:

Suppose $f(\beta_0, \beta_1)$ is <u>any</u> function we want to maximize.

The $\hat{\beta}_0, \hat{\beta}_1$ that maximize f also work for $2f(\beta_0, \beta_1)$

In fact, if c is any positive number (not depending on β_0, β_1) then the $\hat{\beta}_0, \hat{\beta}_1$ work also for $cf(\beta_0, \beta_1)$.

Also: if $\hat{\beta}_0, \hat{\beta}_1$ are the values maximizing $\log f(\beta_0, \beta_1)$ then $\hat{\beta}_0, \hat{\beta}_1$ also work for $f(\beta_0, \beta_1)$.

Going back to our likelihood fcn.

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right)$$

We can ignore the constants $\frac{1}{\sqrt{2\pi}}, \sigma$.

in front.

Then take log: get new fcn

$$\sum_{i=1}^{n} -\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}$$

Again $\frac{1}{2\sigma^2}$ is a positive factor,

as in $c\,f(\beta_0, \beta_1)$ above.

So now we have

$$\sum_{i=1}^{n} -(Y_i - \beta_0 - \beta_1 X_i)^2$$

maximizing this means minimizing the sum of squared residuals, because of the negative sign.

Conclusion: with the standard assumptions of linear regression the OLS estimators are the same as the MLE.

(Also true for many indep. vars.)

Proto-midterm-questions on R:

Q: Wackerly p. 53 computes a 95% CI using "predict" but there is no "95%" in the input to predict. How do we know that this is a 95% CI?

A: Ask an LLM where in the R manual this is discussed: it will give you a reference to predict.lm, predict.lm has a parameter "level", which has default value 0.95.

Q: How do you multiply 2 matrices in R?

A: A %*% B NOT A * B

A * B gives elementwise mult. so

$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ $B = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$ has $A * B = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$

But matrix product $AB = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}$.

More practical questions: Suppose we have a linear model $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

We want to compare it to another linear model

$$Y = \alpha_0 + \alpha_1 (X_1 + X_2) + \varepsilon$$

How do you estimate the second model.?

How do you compare the models?

(With R commands.)

What about a 3rd linear model.

$$Y = \gamma_0 + \gamma_1 X_1 + 2X_2 + \varepsilon$$

How can we estimate this? and compare?