Math 455 Class 10 February 12

Homework due time changed in Gradescope to 11:59pm
(as per your request)

The regression summary output gives you the absolute F-stat, which compares the null model $Y = \beta_0 + \varepsilon$ to the full model.

Anova can be used to compute a relative F-stat as in Wackerly 11.14. (Faraway also)

F-stat compares 2 NESTED models.

Answer to last HW problem is in Wackerly 11.14.

Faraway p. 33

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

- $\hat{\beta}$ is a vector of RVs.
- $\hat{\beta}$ is an unbiased estimator of the true (but unknown) β.
- The components of $\hat{\beta}$ have a (known) variance-

covariance matrix.

~~Rec~~ • In the 1-explanatory-variable case, this corresponds to Wackerly, 11.5 11.6

• For the many-explanatory-variables case, see Wackerly 11.11, 11.12

Remember that the explanatory variables are not taken to be random: they might be chosen by the experimenter.

(Experimental Design)

A desirable property of an experimental design: our estimate $\hat{\beta}_1$ does not depend on $\hat{\beta}_2$ and vice versa.

What this means mathematically is we want.

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = 0$$

(and $\mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) = 0$ if $i \neq j$.)

(or if not 0, at least small).

Part of experimental design might be choosing an X.

Maybe we can choose a good X and get these 0-covariance properties?

For a little more, see Faraway 2.11

For a lot more, courses in Experimental Design.

Q: How can we be tested on this fuzzy stuff?

A: See multiple choice questions on sample midterm.

Also, for more questions, ask an LLM.

Explanation vs. Prediction.

Prediction: we want to use our regression results to predict a NEW Y_{n+1} not in our data set from an X_{n+1} not in our data set.

Data:		New data points
X:	$X_1, X_2 \quad \cdots \quad X_n$	$X_{n+1} \quad X_{n+2}$
Y:	$Y_1 \; Y_2 \quad \cdots \quad Y_n$	$Y_{n+1} \quad Y_{n+2}$

Example in ~~the~~ text, Ch. 4.

Y = rental price of homes

X = explanatory variables # bedrooms,
bathrooms
lot size
sq. feet

We have a data set with rental prices
and these ~~d~~ data # beds, etc for homes.

We get regression coeff from R.

Now a new house comes on the market.

We know the values of # beds etc.

- We do not know the Y_{n+1}
- We want to predict Y_{n+1}
- With our original data set (where we knew all the X and $Y_1, \dots, Y_n$ at ~~on~~ the same time) we are "explaining" the values $Y_1, \dots, Y_n$.

Within "prediction" (Ch 4)

2 possible tasks:

- a single new house comes on the market and we know the values of # bedrooms, etc. want to predict rental price.

- several houses with the same characteristics are coming on the market over the next year Say they all have 3 bedrooms and 2 bathrooms, etc. We want to predict the <u>average</u> rental price of all these homes.

Important principle: when we estimate something, the estimate should come with an error.

An estimator is an RV and has a std. dev.

Our "point estimate" in these 2 cases is the same, but the error is <u>less</u> in the second case. This is because averaging reduces the error.