Math 455 Class 8 February 7

New homework is posted to Gradescope.

How to do this homework:

use the universal textbook problem-solving method:

imitate the examples in the textbook.

in this book there are often italicized bits of advice / rules of thumb that are very useful.

Example Ch 5, p. 59 on the interpretation of regression coefficients. (Sentence interprets $\hat{\beta}_1$)

There is some art to this. First examples that are bad:

<u>BAD</u>: Eating 100 avocados per day caused patients to have blood pressure 1000 points lower.

Why this is bad: 3 reasons.

- implies a causal connection. Remember ice-cream sales and shark attacks.

  Regression does not prove a causal connection.

  We can use regression to give evidence for a causal connection (Hill Criteria, 5.7)

- Hard to imagine the change in explanatory variable, as it is impossible (?) to eat 100 avocados per day, every day, for the rest of your life.
- The change in dependent variable (response) is also beyond all comprehension.

Fixing these three things:

Better: Patients who ate 1 avocado per day tended to have blood pressure 10 points lower.

Notice that we have changed the numbers and "tended to have" does not imply a causal connection.

More subtle problems: see, e.g. "changes of scale", p. 103

If your regression coefficient is $\hat{\beta}_1 = 0.0000009134$

Hard to read and interpret!

Maybe you should change units: cm vs. km?

Q: What if there is no way to do this?

A: Fix it somehow! There is always a way, e.g. drop the variable.

Running a regression in R

Example p. 36

lmod ← lm (Species ~ Area + Elevation, gala)

↑

linear model object : contains results. we get data out of this, data like regression coeff, p-values, using R functions, like summary.

lm is the "linear model" function.

"Species ~ Area + Elevation" is Wilkinson-Rogers notation for specifying the model.

in this case Y = Species X_1 = Area
X_2 = elevation.

the model specified is

$$Y = \beta_0 + \beta_1 \cancel{\text{Species}} X_1 + \beta_2 X_2 + \varepsilon.$$

There are lots of bells / whistles / conventions in this notation.

Some of these are hard to understand, e.g.

" Species ~ . "

this . means "all variables in the data set OTHER THAN Species"

Other things that can be done: remove the intercept, change the regression so one ~~one~~ or more coefficient is specified in advance, etc.

More about the homework:

Textbook (Faraway) does NOT spend a lot of time explaining the theory. More info in Wackerly, et.al, Ch. 11.

In particular a regression has an F-statistic. this is used for the test of H_0: all β_i are 0 (except β_0) vs. H_a: at least one of $\beta_1, \ldots, \beta_p$ is not 0.

There is also a <u>relative</u> F-test, which can be used to compare 2 nested models.

Nested means there is a small model contained in a large one.

Model 1: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$

nested: Model 2 $\supset$ Model 1.

Models that are <u>not</u> nested:

Model 3: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$

Model 4: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \varepsilon$.

Model 3 has X_3, model 4 does NOT.

Even though model 4 is larger, it does NOT contain model 3.

Improving your regression: add/drop variables transform variables.

"transform" might mean: replace Y with $\log(Y)$ or $\sqrt{Y}$ or ...., replace X with X^2, or add X^2 to the variable set etc.

Notice that this may: • change coefficients

• change interpretation of coefficients.

• change significance of coefficients.

This allows us to consider

$$Y = C e^{\beta_1 X_1 + \beta_2 X_2 + \varepsilon} \quad \Big\} \text{ exponential.}$$

$$\underbrace{\log Y = \log C + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}_{\text{linear!}}$$

See problem 2 on HW.