Math 455 Class 8 February 6

New homework assigned (#4). Actually uses R. Make your pdf file with R Markdown.

Today: the interpretation of regression coefficients: What do the $\hat{\beta}_i$ mean?

- What we would ideally like: let's say that X_1 = amount of our blood-pressure lowering drug. We want $\hat{\beta}_1 = -2$ to mean: if we give the patient 1 unit of our drug then this causes the patient's blood pressure to go down by 2 points (mm Hg in the usual scale)

- This is not true: regression analysis by itself does not imply any causal connection between the explanatory variables $X_1, \dots, X_p$ and the response Y.

- Standard example Y = shark attacks in Florida

  X = Ice cream sales in Florida.

  If we measure and run the regression, we can get a statistically significant result.

  This does NOT imply that sales of ice cream cause sharks to attack people..

- So how do we convince the FDA to approve our drug? Answer: regression <u>in combination</u> with other analyses. (Hill Criteria)

  The Hill Criteria are commonsense checks on whether X could cause Y, e.g. Temporality: X must happen before Y otherwise there is no way for X to cause Y.

  (More about this in Ch. 5.)

- If the causal interpretation is not immediately justified, how do we describe the result of our regression?

Answer : some standard phrasings that do not imply a causal connection:

"associated with"; e.g.

"Taking 1 unit of our drug was associated with having blood pressure 2 pts lower."

- Other pitfalls in writing about regression coefficients.

  - the regression coefficient is only defined in the context of the regression that produced it.

    if we run a new regression with more (or fewer) variables, the estimated coeff $\hat{\beta}_1$ of X_1 may change size, sign, or significance.

  - Thus, when writing about the regression coefficient we have to say, "holding $\underbrace{X_2, .., X_p}$ equal" or some such thing.

    other factors/explanatory variables.

Sometimes: "controlling for covariates".

Example: ~~The~~ We report $\hat{\beta}_1 = -2$ for our blood-pressure drug. The FDA is skeptical. "How do we know that the group getting the drug is not younger on average, and thus has lower blood pressure because the elderly tend to have higher blood pressure?"

Our answer is, "we included $X_2 = \text{Age}$ in the regression, and holding age constant, patients who took 1 unit of our drug tended to have blood pressure 2 units lower."

(This is covered under Hill Criteria, Strength (of Association) - does the association still hold after controlling for known covariates.)

- Another issue: try to produce a sentence that has an easy interpretation: use appropriate units. Don't say: Taking 100 units of our drug was associated with having a blood pressure 200 points lower.

200 points lower = dead.

Hard to interpret.

If your coefficient $\hat{\beta}_1$ is 0.0000007231,

change units!

For now: linear changes of variable, e.g. km in place of mm, will not change the statistical quality of your regression. (More about this in Ch. 7.)

Regression summary: if we have a data set with n rows and a regression with p undetermined parameters $\beta_0, \beta_1, \dots, \beta_{p-1}$

the F-stat has $p-1$ numerator and $n-p$ denominator d.o.f.

The F-stat can be recovered from n, p, and R^2. (Formula: Wackerly 11.44 & HW)

Discussion of pitfalls of R:

```
lmod ← lm( Y ~ X + X^2 )
```

the X^2 does not show up in the output.

this is a pitfall of Wilkinson-Rogers notation.

We need to write `lm( Y ~ X + I(X^2) )`

I = identity operator, protects X^2 from being interpreted ~~as~~ in Wilkinson-Rogers notation.

```
lmod ← lm( Y ~ -1 + X )
     ← lm( Y ~ 0 + X )
```

regression with 0 intercept:

$\hat{\beta}_0$ pre-specified to be 0.