Fall, 2023 *Thursday, Sept. 28*

Stat 414 - Day 4
Categorical variables

Last Time:

- Statistical significance of a coefficient ($H_0 : \beta_1 = 0$)
    - t-test: $t = \hat{\beta}_1 / SE(\hat{\beta}_1)$ with $n - p - 1$ df and $SE(\hat{\beta}_1)$ measuring the sample-to-sample variation in the slope random variable (here p is number of slopes)
    - F-test (will match t when only one coefficient. Otherwise tests all slope coefficients at once ("overall model utility test")
- Practical significance
    - R^2 is proportion of variation in response explained by the model (is the model useful?)
    - Residual standard error is the square root of Mean Square Error (typical or average prediction error)
    - "raw" coefficient (how large is the impact?), can also standardize to make comparable to other coefficients (variables with different scales)
- Starting to see a "theme" in whether or not we account for degrees of freedom and how that relates to bias in an estimator

---

Example 1: Pace of Life and Heart Disease, cont

Regressing Heart on Region:

(a) What are the degrees of freedom for the F-test and why? What null hypothesis is being tested by this F-statistic?

How does this relate to the "prediction equation"?

```r
model4 <- lm(Heart ~ Region, data = PaceData )
summary(model4)
```

(b) Write out the prediction equation for this model.

The above predication equation is using "indicator coding" where "dummy variables" are created behind the scenes and put into the model. The missing coefficient becomes the reference group.

(d) Interpret the coefficient for Northeast in the above model. (*Hint*: Keep in mind that slopes are about differences)

(e) What does the t-test for the Northeast coefficient tell you?

(f) Interpret the intercept in the above model

Another way to parameterize the model with categorical variables is “effect coding.” An “effect” is how much higher/lower a treatment/group is from the overall mean.

```
model4b = lm(Heart ~ Region, data = PaceData , contrasts=list(Region="contr.sum"))
summary(model4b)
```

(g) Interpret the intercept in the above model

(h) Interpret the Northeast coefficient in the above model

(i) How have the *F*-statistic and p-values changed and why?

Notes

- The above focuses on “statistical significance” but it is also important to look at “standardized effect sizes” to help assess “practical significance.” (Standardized effect sizes are unitless so don’t depend on the scaling of the variables.)
- For next time: interactions!