Stat 414 - Day 6
Partial *F* tests

Last Time:

- Scatterplot matrix is good for examining each variable's individual relationship with response variable and to check for linear associations among the explanatory variables
  - The Variance Inflation Factor (VIF) for a variable is related to the strength of the linear dependence of that variable and all the predictors in the model (aka multicollinearity). VIF Values larger than 5 or 10 are usually flagged. Can create strange behavior in a model (e.g., coefficients of the wrong sign).
- Variables are often pre-processed before added to a model, e.g., centering, standardizing, min-max scaling. These can aid with interpretation.
  - Centering (and standardizing) can help reduce certain kinds of multicollinearity (e.g., polynomial model).

Example 1: Reconsider the Pace of Life data.

We did not observe much multicollinearity in this dataset, but you still might consider collapsing the 4 "pace" variables into one "index."

```
paceindex = with(PaceData, Walk + Talk + Bank + Watch)
cor(PaceData$Heart, paceindex)
```

Let's compare the following 3 models:

```
summary(model1 <- lm(Heart ~ paceindex, data = PaceData))
summary(model2 <- lm(Heart ~ Region, data = PaceData))
summary(model3 <- lm(Heart ~ paceindex + Region, data = PaceData))
```

(a) What does the F-statistic in model 3 test?

(b) From this output, can you answer the question: Is the paceindex statistically significant after adjusting for Region? If not, what do you need to do?

(c) From the above output, can you answer the question: Is the Region statistically significant after adjusting for paceindex? If not, what do you need to do?

To assess the significance of a set of terms, you want to carry out a "partial F-test" by looking at the "drop in SSError" from adding the terms to an existing model.

(d) What is the SSError when only paceindex is in the model?

```
anova(model1)
sum(model1$residuals^2)
```

(e) What is the SSError in the model that includes both paceindex and Region? How much have we reduced the unexplained variation when adding Region to the first model? How many terms did we have to use to get that reduction?

```
anova(model3)
```

To decide whether this drop in SSError is statistically significant, we need to compare the value to something. We will use the MSError from the "full model" because we consider it the most accurate estimate of σ^2. Dividing these terms by their degrees of freedom gives us an F-statistic, where the df for the numerator is the difference in the df for the two models.

$$F = \frac{(drop\ in\ SSError/difference\ in\ model\ df)}{MSError(full)}$$

(f) Calculate this value from the above output.

A short-cut in R that we will make a lot of use of this quarter for comparing "nested" models:

```
anova(model1, model3)
```

(g) What do you learn from this output? What do I mean by "nested models"?

(h) What do you learn from the following? Is this equivalent to an earlier result? Explain.

```
anova(model2, model3)
```

But could we have just run `anova(model3)` to assess the added contribution of paceindex after adjusting for Region?

(i) For this output, what hypotheses are tested with the p-value for paceindex?

While the t-statistics are always adjusted for all other variables in the model, the F-statistics are "sequential" instead. One approach is to change the order in which you enter the variables in the model.

```
anova(model4 <- lm(Heart ~ Region + paceindex, data = PaceData))
```

Another approach is to examine "Type 2" sums of squares.

```
#Here I want to use "big A" anova
#library(car)
car::Anova(model3, type = 2)
```

Notes

In fact, the residual standard error is used all over the place, e.g.,

- Smaller values indicate a better fitting model
- 95% prediction interval ≈ $\hat{y} \pm 2\hat{\sigma}$
- $se(\hat{\beta}_i) = \hat{\sigma}/\sqrt{(n-1)Var(X)}$

So it's important that we estimate σ "correctly." In particular, the "basic linear model" assumes $Var(\epsilon) = \sigma^2$. If you don't believe you have sufficient heterogeneity, one approach is to make different distributional assumptions.

For next time Adjusted associations and Interactions