Stat 414 - Day 3
Inference for Regression

Last Time:

- Transformations can be useful for linearizing data and for normalizing data
- Polynomial models can be useful for modeling nonlinear data

---

When the basic model assumptions are considered met, we can continue to examine confidence intervals and p-values.

Example 1: Pace of Life and Heart Disease

On a recent international trip, we noticed that we were walking down the street much faster than other people. This reminded me of a study by Levine (1990) on "pace of life" in different countries. One way he measured pace of life was "average walking speed of randomly chosen pedestrians." He then explored a possible association with incidence of heart disease (age-adjusted death rate due to ischemic heart disease). We will look at data for 36 U.S. cities, for walking speeds over a distance of 60 feet (measured during business hours on a clear summer day along a main downtown street, no units)

```
iscamsummary(PaceData$Heart)
```

Is walk a statistically significant predictor of heart disease?

```
model1 = lm(Heart ~ Walk, data = PaceData)
plot(Heart ~ Walk, data = PaceData)
abline(model1, col = "red")
#Note, lines doesn't work here because the data aren't sorted, so giving it the int
ercept and slope of the line to add

#Ask more questions of the model
summary(model1)
```

(a) Identify and interpret the residual standard error (and r).

(b) How does the residual standard error compare to the standard deviation of Heart?

Examine the ANOVA table

```
anova(model1)
```

(c) Confirm the calculation of the residual standard error from this table.

(d) How is the F value for Walk calculated? What are the degrees of freedom for this F-value? Why? What does this F-value (and p-value) tell you? (What is the null hypothesis?)

(e) How does the F value for Walk compare to the t-statistic for Year in the first table? How do the p-values compare? What null hypothesis does this p-value test?

(f) In the Regression Table output, what is the standard error of the slope coefficient for Walk? How do we interpret this value? Demo: Regression applet PaceData

(g) What if we had centered the Walk variable first?

```
#centering = subtract the variable mean from each value
model2 = lm(Heart ~ I(Walk - mean(Walk)), data = PaceData)
anova(model2)
```

(h) What if we had standardized the walk variable first?

```
#The scale command allows you to center and/or standardize (also divide by the SD,
think z-score)
zWalk = scale(PaceData$Walk, mean(PaceData$Walk), sd(PaceData$Walk))
#I used a capital W!
model3 = lm(Heart ~ zWalk, data = PaceData )
anova(model3)

plot(PaceData$Heart ~ zWalk)
abline(model3)
```

For each region in the United States, 3 large cities, 3 medium-size cities, and 3 smaller cities were selected. Let's see whether the regional differences are statistically significant.

```
boxplot(Heart ~ Region, data = PaceData)
summary(aov(Heart ~ Region, data = PaceData))
```

The above ANOVA is equivalent to a "linear model" ...

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

(i) Why is this a linear model? But what should we worry about?

Now fit the multiple regression model.

```
#Note the shortcut
anova(model5 <- lm(Heart ~ Walk + Region, data = PaceData))
```

(j) What do you conclude?

(k) Does the coefficient of Walk change between the two models? What does that tell you?

(l) What do you learn from the following model?

```
model6 <- lm(Heart ~ Walk + Region + Talk + Bank + Watch, data= PaceData)
summary(model6)
```

Is talk a statistically significanct predictor of heart disease?

The most common method to check for linear associations among the explanatory variables is variance inflation factors. The car package quickly gives us VIF values.

```
#install.packages("car")
car::vif(model6)
```

(m) What do you learn? What should you do?

Multicollinearity often arises with quadreatic models.

```
KYDerby23 = read.table("https://www.rossmanchance.com/KYDerby23.txt", header=TRUE)
model2 = lm(speed~Year + I(Year^2), data = KYDerby23)
car::vif(model2)
```

Centering or Standardizing can reduce multicollinearity between "product terms":

```
centeredyear = KYDerby23$Year - mean(KYDerby23$Year)
centeredyear.squared =centeredyear*centeredyear
model2b = lm(speed ~ centeredyear + I(centeredyear^2))
car::vif(model2b)
```

(n) Did we improve multicollinearity?

Not sure why these $year$ and $year^2$ are no longer collinear?

```
plot(centeredyear ~ I(centeredyear^2))
```

All we have done is move the curve in the quadratic relationship to be in the middle of our x-space. This is fine, even beneficial, having strongly related x-variables just causes problems with they "line up."

(o) How do we interpret the intercept of this model with both variables centered?

Notes

- The above focuses on "statistical significance" but it is also important to look at "standardized effect sizes" to help assess "practical significance." (They are unitless so don't depend on the scaling of the variables.) See the Quiz for some calculations.
- Centering a variable (by subtracting the mean from each value) can help with
  - making the intercept more interpretable (when x is at the mean rather than when x is at zero, which may not be a value in dataset)
  - making comparings of slopes more meaningful (a one SD change)
  - reducing multicolinearity in ''product'' terms

For next time Review interpreting coefficients with categorical variables.

Quick Review: Analysis of Variance (ANOVA)

- Traditionally thought of as method for comparing population means but more generally is an accounting system for partitioning of variability of response into model (explained) and error (unexplained)
- To compare group means, assume equal variances (and conditional normality and independence)
  - Pooled variance: $MSE = s_p^2 = \frac{(n_1-1)s_1^2+\cdots+(n_I-1)s_I^2}{(n_1-1)+\cdots+(n_I-1)}$
  - $s_p = s_\epsilon$ = residual standard error = root mean square error
- *SSTotal* = $\Sigma(y_i - \bar{y})^2$ = *SSModel* + *SSError* with df = $n - 1$
- R^2 = 1 – SSError / SSTotal ≈ 1 - s_ϵ^2/s_y^2 = 1 - $\frac{MSError}{MSTotal} = R^2_{adj}$
  - Does knowing "x" tell me a lot about "y"?
- **ICC** = $(MSGroups - MSError)/(MSGroups + (k-1)MSError)$
  - Does knowing response in one region tell me much about next response from that region? ("reliability")
- *F* = MSGroups / MSError (Between group variation / Within group variation)
  - $F = \frac{R^2}{1-R^2} \times \left(\frac{n-I}{I-1}\right)$
  - Values larger than 4 are generally significant
  - When have only two groups, equivalent to a *pooled* two-sample *t*-test