Stat 414 - Day 2
Inference for Regression

Last Time:

- Multilevel data is when the structure of the data is characterized by "observational units" at different levels, often from clustering or nesting in the data (e.g., students nested in classrooms)
- Multilevel data needs to be analyzed differently from single level data
- Maximum likelihood estimation is an alternative to least squares estimation. The estimated coefficients are often the same, but estimates of variability do differ.

Example 1: Predicting airfare cont

The **least squares regression model** fits the best fitting line by minimizing the sum of the squared residuals.

```
model1 = lm(price ~ distance, data=airfare); model1
```

Cool trick: The intercept of the regression is 214.99 and the slope of the regression is 0.14.

n = number of observations in study = 12

(a) Explain the df values of 11, 1, and 10. How is residual standard error calculated?

df total = n - 1 = 11, df model = number of slopes in the model = 1, df residual = df total - df model

(b) What is the distinction between R-squared and Adjusted R-squared?

Adjusted R-squared takes into account the degrees of freedom.

R^2 = 1- SSError / SSTotal; R^2adj = 1 - MSError / "MSTotal"

Recall that maximum likelihood estimation produced the same slope and intercept estimates, but differed in how it estimated σ.

```
model1ML <- nlme::gls(price ~ distance, data = airfare, method = "ML")
coef(model1ML); logLik(model1ML); sigma(model1ML)
```

-69.026 (df = 3) 76.19 = sum of squared residuals / 11

Likelihood estimation also includes a mechanism for "penalizing" your fit statistics based on the number of parameters being estimated.

```
AIC(model1ML)  #-2 x log-likelihood + 2p
## [1] 144
BIC(model1ML)  #-2 x log-likelihood + p x ln(n)
## [1] 146
```

(c) Which conveys better "fit," large values or small values?

smaller values of AIC and BIC convey better fit

To deal with the bias in the ML estimation of variance, we can use **restricted maximum likelihood estimation (REML)** which takes the number of parameters into account and produces an unbiased estimator or σ^2.

```
model1REML <- nlme::gls(price ~ distance, data = airfare, method = "REML")
sigma(model1REML)
sqrt(sum(model1$residuals^2)/10)
```

The main thing to keep in mind is that if you want to compare models, use the same estimation procedure (e.g., both use REML or both use ML).

83.46 is how REML estimates "sigma" = sum of squared reisduals / 10

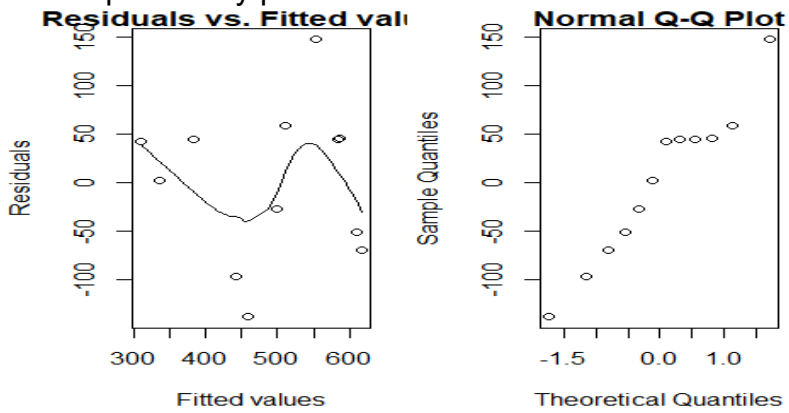
(d) Before we look at p-values and confidence intervals, what are the primary "assumptions" that need to be satisfied for inference in regression models?

Linearity = linear relationship between (average) response and x-variable (form)
Independence of errors
Normal distribution of responses at each x, aka normality of errors
Equal variance of responses at each x

Secondary: x-variables are measured without error

Most important: have the right variables

With more complicated models, an important diagnostic tool is residual plots. The two to start with are a graph of residuals vs. fitted values (aka predicted values) and a histogram and/or normal probability plot of the residuals.



(e) Summarize what you learn from these graphs.

wnat random scatter in residuals vs. fits for linearity, what straight line in QQplot for normlaity

Additional model assumption: "Most importantly, the data you are analyzing should map to the research question you are trying to answer." (Gelman & Hill). Also be on the looking out for "influential observations."

When we think the regression model assumptions are met, we can ask more questions of the model.

```
summary(model1)
## lm(formula = price ~ distance, data = airfare)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 214.9944    69.8120    3.08   0.0116 *
## distance      0.1425     0.0338    4.21   0.0018 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84 on 10 degrees of freedom
## Multiple R-squared:  0.639,  Adjusted R-squared:  0.603
## F-statistic: 17.7 on 1 and 10 DF,  p-value: 0.0018
```

(e) What do each of the standard error values tell you?

SE intercept (69.812) = sample to sample variation of intercepts by chance alone
SE slope (.0338) = sample to sample variation of slopes by chance alone

These convey the accuracy of these statistics from the population values

Detour: Follow this link for the Regression applet (https://www.rossmanchance.com/applets/2021/regshuffle/regshuffle.htm?data=airfare.txt)

- Check the **Show Shuffle Options** box.
- Use the pull-down menu to select the slope as the statistic.
- Press **Shuffle Y-values** a few times. What is the applet doing?
- Set the **Number of shuffles** to 1000 and press Shuffle Y-values.

(f) Describe the distribution of the shuffled statistics (shape, mean, standard deviation).

approximately normal with mean similar to the population slope and standard deviationaround .053

(g) What percentage of the shuffled slopes is at least as extreme as our observed slope?

The observed slope (.142) did not show up in my first 1000 samples. Doing 10,000, got closer to 0.01.

(h) Give 1-2 reasons why the standard error of the slopes does not match the output in the regression table.

The simulation here is about "random assignment" rather than about "random sampling" which is what the traditional regression model assumes...

Model comparison: Let's consider the ANOVA table:

```
#Examine the ANOVA table associated with this model
summary(aov(model1))
##              Df Sum Sq Mean Sq F value Pr(>F)
## distance      1 123463  123463    17.7 0.0018 **
## Residuals    10  69662    6966
```

(i) How is $V(Y)$ calculated from this table? How is the residual standard error found in this table? How is the F-value calculated? How does the p-value compare to the regression table?

SSTotal = 123463 + 69662 = 193125
"MSTotal" = 193125 / 11 = 17556.2

Residual standard error = 6966 = MSError
p-value is in the general ballpark

In general, the F-test compares the $SSError$ for the model without $distance$ to the model with $distance$ and measures the significance of the "drop in SSError" between the two models, accounting for the difference in the degrees of freedom between the two models.

$$F = \big(SSError(reduced) - SSError(full)\big)/1/MSError(full) \sim F(1, n-2)$$

Likelihood Ratio Tests

In a similar manner, we can assess the significance of adding the $distance$ variable to the model by comparing the log-likelihood values for the model with distance and the model without distance.

```
model0ML <- nlme::gls(price ~ 1, data = airfare, method = "ML")
summary(model0ML)
#How much better (in terms of log likelihood) is the model with distance?
teststat = 2*(logLik(model1ML) - logLik(model0ML)); teststat
## 'log Lik.' 12 (df=3)
1 - pchisq(teststat, 1)
## 'log Lik.' 0.00047 (df=3)
```

This test statistic (asymptotically) follows a chi-square(d) distribution with df = different in number of parameters between the two models. It works for "nested models" where one model is derived by setting k coefficients to zero in the other. The validity of a LRT is often "quicker" (don't need as large of a sample size) as the t and F tests.

```
#nice short cut
anova(model0ML, model1ML)
##          Model df AIC BIC logLik   Test L.Ratio p-value
## model0ML     1  2 154 155    -75
## model1ML     2  3 144 146    -69 1 vs 2      12   5e-04
```

(j) Are the p-values the same?

Not exactly, but should be similar

Notes:

Because of the biased nature of the estimate of σ, an alternative proposed in the 1930s is ***restricted (residual) maximum likelihood***, REML. Some properties:

- Maximizes a different likelihood function (special matrix multiplication, some rows constrained), that doesn't depend on the slope coefficients (residuals, removes the "fixed" effects from model).
- Is often an "iterative" estimation process (e.g., estimate slope coefficients, then sigma, then slope coefficients, etc.)
- With simple regression models, the estimated slope coefficients are the same, but the parameter estimate of the variances differ.
- $E(\hat{\sigma}^2_{REML}) = \sigma^2$ (so really unbiased for variance not standard deviation)
- Overall, REMLs are better to use for estimating variances

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)

(a) Is walk a statistically significant predictor of heart disease? How are you deciding?

Look at p-value for slope coefficient of the walk variable(t = 2.162, p-value = .0377)

(b) How do/do not the intercept and slope coefficients change? The p-value for walk? What is an advantage of centering?

centering doesn't change slope coefficient/p-value but makes intercept more meaningful to interpret

(c) How do/do not the intercept and slope coefficients change? The p-value for walk? What is an advantage of centering?

standardizing makes the slope corresponds to a 1SD increase in the x-variable

(d) Carry out a likelihood ratio test of the significance of Walk using restricted maximum likelihood

get the two likelihood values, -109.25 and -107.72, take the difference (make it come out positive) and multiply by two, 2(109.25 - 107.72) = 3.06, df = 1 (one different parameter from intercept only model). The anova command also gives 3.07 with p-value .0798.