Overview of Regression & ANOVA

Different Names for the Same Thing

Ryne Estabrook

Department of Psychology
University of Virginia

July 24, 2009

Goals of Today's Talk

- Today, I'll be giving an overview of regression and ANOVA.
- Topics we'll go over:
  - Some background knowledge.
  - Basics of regression & ANOVA as GLM simplifications.
  - Model selection, nonlinearity, diagnostics.
- Oh, and a few things about this talk:
  - Interrupt with questions as you have them.
  - This isn't inherently a programming talk, but I'd be happy to answer whatever questions you may have about your statistical program(s) of choice.
- More talks to come!

Whad'Ya know?

Not much, you?

- There are a few mathematical & statistical concepts that you need to know to understand regression, ANOVA & statistics in general.
- I'll start with a quick review of these topics:
  - Measurement.
  - Statistics as a concept.
  - Descriptive statistics.
  - Distributions.
  - Degrees of freedom.
- Then we'll move on to the topics of the day.

Measurement

- Measurement is “the assignment of numbers to things according to a rule (Stevens, 1939)."
- The numbers we assign have some property that exists in our data.
  - We identify Nominal, Ordinal, Interval and Ratio scales, which correspond to specific mathematical properties.
  - These properties can be summarized as Equality, Order, Addition, & Multiplication.
- We can then use the mathematics to take advantage of those numerical properties.
  - Its not that we use numbers just because we like numbers, but the relationships in data can be described by the relationships between numbers.

Statistics

- Statistics is the branch of mathematics that deals with data.
  - There's already a lot of math out there, that describes any relationship you can put in appropriate numerical terms.
  - Think of math as a language you can co-op, rather than reinventing the wheel for every new construct or topic.
- There are two general classes of statistics:
  - Descriptive Statistics, which I'll review now, and
  - Inferential Statistics, which includes regression and ANOVA.

Descriptive Statistics

- Descriptive statistics are used to describe data.
  - Wow, that's really deep.
  - You may also hear the term *summary statistics*.
  - The point of these statistics is to take more data than you can hold in your head at once, and reduce it in such a way that you understand as much of the data as possible with as few numbers as possible.
- The most common descriptive statistics are measures of central tendency:
  - Mode: any scale of measurement, not often analyzed.
  - Median: Ordinal-Ratio scales, used in non-parametric statistics.
  - (Arithmetic) Mean: Interval & Ratio scales, used in parametric statistics, has units, tied to distributions.

Comparing Measures of Central Tendency

Aspect	Mode	Median	Mean
Scales	Nominal, Ordinal, Interval & Ratio	Ordinal, Interval & Ratio	Interval, & Ratio
Observations Used	Varies	1-2	All
Formula	No	“Sorta"	Yes
Advanced Statistical Properties	No	Some	Lots
Sensitive To Extreme Observations	No	No	Yes

Descriptive Statistics

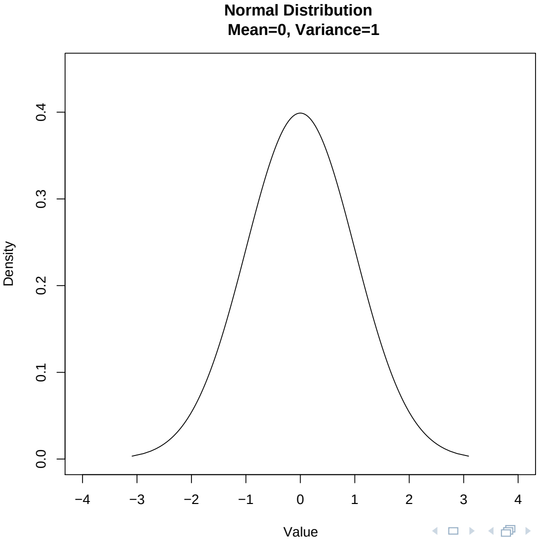
No one is average

- The other class of descriptive stats you need to know are measures of spread.
  - I won't focus on ranges or quantiles, the measures of spread tied to the median and non-parametric statistics.
- If you use the mean for central tendency, you'll likely use variance & covariance as measures of spread.
  - Variance is the sum of squared deviations from the mean for any variable.
  - Standard deviation is the (positive) square root of variance.
  - Covariance is the sum of the products of the deviations from the mean for two variables.
    - The covariance of any variable with itself is its variance.
  - Correlation is covariance in standard units.

Normal Distribution

- The Normal Distribution is defined by mean and variance, and symmetric around the mean.
  - The mean is the most value of X with the highest probability.
- The Normal Distribution is scaled by the standard deviation.
  - If you multiply X by 2, both the mean and standard deviation double, and the variance quadruples.
  - 68.2% falls between $\mu \pm \sigma$, 95.4% between $\mu \pm 2\sigma$
- The Normal Distribution is defined from $-$ inf to inf.
  - Most computer approximations of the normal distribution place a cutoff beyond which the probability of X is zero.

Normal Distribution, Probability Density Function



The χ^2 Distribution

- The χ^2 distribution is the square of the normal distribution.
- The version of the normal distribution used in the χ^2 is the standard normal, with μ=0 and σ^2=1.

$$X \tilde{} N(0,1)$$

$$\chi_1^2 = X^2$$

- This version of the χ^2 distribution has 1 degree of freedom.
- The χ^2 distribution with *k* degrees of freedom is the sum of *k* independent squared standard normal distributions.
- The *F* distribution with (d_1, d_2) degrees of freedom is the ratio of two χ^2 distributions, divided by their *df*.

Degrees of Freedom

- Degrees of Freedom (*df*) is an important & confusing statistical topic.
- Mathy Definition:
  - Dimensionality of a random vector.
  - Refers to how many components of a vector need to be known before a vector is determined.
- Less-Mathy Definition:
  - The number of (remaining) unique pieces of information in a set or system.
  - *df* are the units by which we measure information.
  - Most often, we'll talk about the dimensionality of data in terms of number of subjects.

Other Stuff

- I may hint at some matrix operations.
  - If you're unfamiliar or rusty, a matrix is a rectangular table of numbers, on which you can do various types of math (including special types of addition and multiplication).
  - A vector is a matrix with either one column (column vector) or one row (row vector), and a scalar is a matrix with one row and one column (just a number).
- I also assume some knowledge of probability and distributions.
  - The distributions I just referenced are the basic ones.
  - Advanced distributions and probability knowledge is very useful.
- Any other questions before we move on?

General Linear Model

- The general linear model is a very general model that describes linear relationships between variables.
- It is typically expressed in matrix terms:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$$

- where:
  - **Y** is an *n* x *k* matrix of dependent variables,
  - **X** is an *n* x *p* matrix of independent variables,
  - β is an *p* x *k* matrix of estimated terms that define the relationships between **X** & **Y**, and
  - **E** is an *n* x *p* matrix of residuals or errors.

General Linear Model

Univariate

- This model can be used for a large variety of data types and models, especially multivariate models.
  - Multivariate means multiple dependent variables, and doesn't relate to number of independent variables.
- If you want to do univariate analyses, it gets simpler.

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$$

- where:
  - **Y** is a vector of length *n* of dependent variables,
  - **X** is an *n* x *p* matrix of independent variables,
  - β is a vector of length *p* estimated terms that define the relationships between **X** & **Y**, and
  - **E** is a vector of length *n* of residuals or errors.

General Linear Model

Univariate, no matrices

- We still have that pesky **X** matrix.
- If we have *p* independent variables, we can split that into *p* lists or vectors, and get this:

$$\mathbf{Y} = \mathbf{X_1}\beta_\mathbf{1} + \mathbf{X_2}\beta_\mathbf{2} + \ldots + \mathbf{X_p}\beta_\mathbf{p} + \mathbf{E}$$

- where:
  - **Y** is a vector of length *n* of dependent variables,
  - $\mathbf{X_1}$ - $\mathbf{X_p}$ are vectors of length *n* of independent variables,
  - $\beta_\mathbf{1}$ - $\beta_\mathbf{p}$ are scalar estimated terms that define the relationships between **X** & **Y**, and
  - **E** is a vector of length *n* of residuals or errors.

General Linear Model

- That should look pretty familiar, and can be used for a great many things, depending on how **X** and **Y** look.
- Stuff we need to know to estimate this model:
  - The characteristics of **X** and **Y**, including scale of measurement and distributional characteristics.
  - Some treatment for residuals or the conditional distribution of **Y** given **X**.
  - A method for estimation, or a way to pick β.
- Today, we're talking about two different applications of the GLM, defined by the characteristics above: ANOVA & regression.

Regression

Little bit of history

- So some of the first people to use this method were interested in “regression to the mean."
  - Galton (Darwin’s cousin) was interested in why the offspring of tall people were shorter than their parents.
  - Legendre and Gauss began the use of the method of least squares for estimation problems.
- Regression analysis is fundamentally about prediction:
  - If a horse can run X mph, how fast will his offspring run?
  - For any set of values on some predictors or independent variables, what is my predicted value of a dependent varaiable?
  - When X goes up 1 unit, how does Y move?
- And of course, it’s now recognized as a subset of the GLM.

Regression Lines

- Being GLM, we're going to fit some lines.
- More accurately, the formula we're using for prediction consists of linear combinations.
- The equation for simple regression looks like so:

$$
\begin{aligned}
Y &= b_0 + b_1 X + e \\
\hat{Y} &= b_0 + b_1 X
\end{aligned}
$$

- Y & X are observed variables, b_0 & b_1 define the predicted relationships.
  - e is our error or residual term, and $\hat{Y}$ is the predicted value of Y (Y-e).
- For those who've forgotten, the equation for a line is:

$$Y = mX + b$$

Regression Lines

This is a line

picture of a line

Regression Lines

“Multivariate Lines"

- We can have however many predictors we want, which means we don’t have lines anymore.
- With two predictors, we have a regression plane.
  - We have three variables (2 IV, 1 DV), we have to make three dimensional plots (plot in 3-space).
- With three or more predictors, we have a regression hyperplane.
  - We have four or more variables (3+ IV, 1 DV), we have to make high-dimensional plots.
  - These plots (in 4-space), are hard to draw. In 4-space, you can try to animate to use time, but that’s more flashy than useful.

Fitting Models

Something's missing

- The “goal" of regression is to yield some estimates for our parameters.
  - What's a parameter? Its a value that is fixed for a given sample or experiment, and is the “best" estimate of that value in the population being applied to.
  - Y and X aren't parameters, because they each vary.
  - The relationship between Y and X (b_1) and the intercept of Y (b_0) are, because they're fixed for the sample (we just don't know them yet).
- To get parameter estimates, we need some criterion by which to pick estimates (a method of estimation).
- This criterion is known as an estimator or objective function. In general, we're looking to have the lowest difference between Y and $\hat{Y}$, however we define that.

Fitting Models

Something's missing

- So we want to make the smallest errors possible, $e = Y - \hat{Y}$.
- How do we do that?
  - We can't take a raw difference, as $\hat{Y} = \inf$ gives the lowest possible error.
  - Smallest absolute error underlies non-parametric stats.
  - Smallest squared error, or $(Y - \hat{Y})^2$ underlies parametric stats.
- This least-squares criterion is the most common way to find parameters and solve regression problems.

(Ordinary) Least Squares Estimation: OLS

It's SLO backwards

- Why use least squares?
  - It forces all errors to be positive, making minimization meaningful.
  - By the Gauss-Markov theorum, OLS estimation yields the best linear unbiased estimates of parameters, errors are homoskestastic and have an expectation of zero.
  - It is equivalent to maximum likelihood estimation, which has additional assumptions.
- Maximum likelihood?
  - Assumes DV is normally distributed with unknown mean and variance.
  - Just as the mean is the value at which variance is minimized, maximum likelihood estimate minimizes squared error (residual variance).

More OLS

You're SLO backwards!

- It's the distributional assumptions that make OLS and ML really flexible.
  - If errors are normally distributed, then regression becomes a statement of *mean structure*.
  - Normality assumptions, like those in the Central Limit Theorum, get at standard errors.
  - CLT assumptions and random sampling allow us to generalize our regression parameters as parameter estimates that apply to the populations we study.
  - Under OLS, regression becomes a function of the covariance matrix of the independent and dependent variables.

Expressing Regression using Covariance

You're SLO backwards?

- Here, I'll distinguish between raw units regression parameters (b_i) and standardized parameters (β_i).
- Under simple regression, the regression parameters can be expressed as:

$$\begin{aligned} b_1 &= r_{XY} * \frac{\sigma_Y}{\sigma_X} = \frac{Cov(X,Y)}{Var(X)} \\ b_0 &= \bar{Y} - b_1 * \bar{X} \end{aligned}$$

- Under multiple regression, the regression parameters can be expressed much more complexly.

Putting it all together

- We're trying to predict a dependent variable, and we express that variable's mean structure as a function of a set of independent or predictor variables.
- Alternately, the DV is normally distributed, with parameters:

$$
\begin{aligned}
Y &\tilde{} N(b_0 + b_1 X_1 \ldots b_j X_j, \sigma_e^2), OR \\
Y &= b_0 + b_1 X_1 \ldots b_j X_j + e, e\tilde{}N(0, \sigma_e^2)
\end{aligned}
$$

- We estimate the *b* or β parameters by minimizing the variance of the residuals.

Putting it all together

- Let's interpret this regression equation:

$$\begin{aligned} Y &= 19.47 + 1.04 * X_1 - 0.43 * X_2 + e \\ Var(e) &= 3.43 \end{aligned}$$

Putting it all together

- Let's interpret this regression equation:

$$
\begin{aligned}
Y &= 19.47 + 1.04 * X_1 - 0.43 * X_2 + e \\
Var(e) &= 3.43
\end{aligned}
$$

- When X_1 and X_2 are zero, the predicted value of Y is 19.47.

Putting it all together

- Let's interpret this regression equation:

$$\begin{aligned} Y &= 19.47 + 1.04 * X_1 - 0.43 * X_2 + e \\ Var(e) &= 3.43 \end{aligned}$$

- When X_1 and X_2 are zero, the predicted value of Y is 19.47.
- Y goes up 1.04 units for every unit X_1 increases.

Putting it all together

- Let's interpret this regression equation:

$$\begin{aligned} Y &= 19.47 + 1.04 * X_1 - 0.43 * X_2 + e \\ Var(e) &= 3.43 \end{aligned}$$

- When X_1 and X_2 are zero, the predicted value of Y is 19.47.
- Y goes up 1.04 units for every unit X_1 increases.
- Y goes down 0.43 units for every unit X_2 increases.

Putting it all together

- Let's interpret this regression equation:

$$
\begin{aligned}
Y &= 19.47 + 1.04 * X_1 - 0.43 * X_2 + e \\
Var(e) &= 3.43
\end{aligned}
$$

- When X_1 and X_2 are zero, the predicted value of Y is 19.47.
- Y goes up 1.04 units for every unit X_1 increases.
- Y goes down 0.43 units for every unit X_2 increases.
- The residual variance of Y (variance around regression line) is 3.43.

Question Slide

- This is a question slide.
- Stuff we've learned:
  - The basics of regression analysis, both simple and multiple.
  - How we estimate regression parameters.
  - Very basic interpretation.
- Stuff we haven't learned:
  - Advanced interpretation.
  - Model diagnostics.
  - Other flavors of the GLM.
- So, ask some questions already.

Analysis of Variance

- ANOVA was developed by R.A. Fisher in the late 1910s, published in 1921 and 1925.
- His goal and application was agribusiness, specifically Guinness brewing, which was one of the driving forces behind finite statistics.
  - The driving force behind infinite statistics was 16^{th}-17^{th} mathematicians making money counseling gamblers. Quant is really about gambling and beer.
- One of the benefits (if not the principle benifit) of ANOVA is that it is easy to do by hand, and is built for small samples.
- It became an often used technique in experimental psychology for the same reasons.
  - ANOVA then “expanded" to deal with more complex issues.
- Let’s see how it works, analyzing variance from a GLM perspective.

Approaching the GLM from Variance

We're sneaking up on it!

- Lets return to that GLM formula, considering a single dependent variable (y) and a single predictor (x).
- We can parse the variance of something into two components: the part shared or caused by something else, and the unique part.
- We typically use a coefficient (b) to describe the variance in Y that is shared with X.

$$
\begin{aligned}
y &= b * x + e \\
Var(y) &= Var(b * x) + Var(e)
\end{aligned}
$$

- If you squint, you can see that this is regression, with some assumptions and transformations.

Approaching the GLM from Variance

- If we assume that we have groups and some DV (Y), then we can approach this like so:

$$Var(Y) = Var(GroupMeans) + Var(WithinGroups)$$

- This formula is a part of the GLM when we define “group means" and “within groups" a very specific way:
  - Var(Group Means) is the variance of each person’s group mean from the mean for all people.
  - Var(Within Groups) is the variance of each person’s score from their respective group mean.
  - I’ll clear up the math very soon. We have to be careful about adding & subtracting variances.
- Let’s introduce this with an example.

Example: Height

- Let's say I wanted to know the heights of some group of people.
  - I know that the mean is my best guess for the height of any one of them, if my criteria is “lowest squared deviation."
- So we know that the mean and variance aren't perfect in this case, because we know there are robust height differences between the sexes.
  - Put another way, we know that the mean isn't really a great way to describe this distribution. Each group should have a mean.
- How can we include that information?

Example: Height

- How can we include it?
  - *t*-tests evaluate the difference in group means.
  - Regression would predict height from an intercept and slope (e.g., β_0=mean female height, β_1=sex difference).
- But we can also analyze this via variance, hence the term analysis of variance (ANOVA).
- Some of the variance in Y is not best described as variance in Y, but variance in the group means.

$$Var(Y) = Var(GroupMeans) + Var(AroundGroupMeans)$$

A Note About Variances.

E[b]

- Variances are a real pain to add and subtract, especially when we're parsing them into parts.
  - We would have to invoke the kind of formulas used when creating pooled variances.
  - Luckily, there is an easier way.
- Another word for variance is mean squared error, often *MS* in ANOVA terminology.
  - $Y_i - \bar{Y}$ is error, then we square it and take the mean.
  - The standard deviation is root mean squared error.

$$Var(Y) = MS_y = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{df_y}$$

A Note About Variances.

E$^{\#}$?

$$Var(Y) = MS_y = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{df_y}$$

- Another way to talk about the “squared-error" component is as sums of squares, or *SS*.
  - Sums of squares (around the mean) are usually subscripted by the variable and/or group they come from.
- We can now simplify the formula for variance even further.
  - In a minute, this will make ANOVA pretty easy.

$$Var(Y) = MS_y = \frac{SS_y}{df_y}$$

ANOVA

$B^{\phi 13}$

- If we throw the variance formulas into our ANOVA.

$$
\begin{aligned}
Var(Y) &= Var(GroupMeans) + Var(WithinGroups) \\
\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{df_y} &= \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_{group})^2 + \sum_{i=1}^{n}(\bar{Y}_{group} - \bar{Y})^2}{df_y}
\end{aligned}
$$

- If we cancel out the degrees of freedom from both sides and rename the sums of squares, we get:

$$SS_y = SS_{betweengroups} + SS_{withingroups}$$

- Now it's starting to look like ANOVA.

ANOVA

Height Example

	Sex		
	Female	Male	Total
Mean	65.07	68.50	66.79
Var (MS)	16.12	15.46	18.60
n	50	50	100

- Ok, let's think about that height example again.
- We need to calculate three things:
  - The sums of squares for the total sample (SS_y)
  - The sums of squares between or across the groups ($SS_{between}$, or $\sum(\bar{Y}_{group} - \bar{Y})^2$)
  - The sums of squares within the groups or from the respective group means (SS_{within} or $\sum(Y_i - \bar{Y}_{group})^2$)

ANOVA

Sums of Squares of Y

- The sums of squares of Y are pretty easy to calculate.
  - If MS = $\frac{SS}{df}$, then SS = MS*df
- We already have the mean squares of Y; they're the variance!

$$\begin{aligned} SS_y &= MS_y * df_y \\ &= Var(Y) * (n-1) \\ &= 18.60 * (100-1) = 1841.18 \end{aligned}$$

ANOVA

$SS_{withingroup}$

- The sums of squares within each group are pretty easy to calculate as well.
- We already have the mean squares of Y for each group; they're the within group variances.
- In the same way that $SS_y = MS_y * df_y$, SS_{within} is calculated for each group, summing across all groups (from j=1 to k).

$$
\begin{aligned}
SS_{within} &= \sum_{j=1}^{k} MS_j * df_j \\
&= Var(Y|F) * (n_F - 1) + Var(Y|M) * (n_M - 1) \\
&= 16.12 * (50 - 1) + 15.46 * (50 - 1) = 1547.18
\end{aligned}
$$

ANOVA

$SS_{betweengroups}$

- So the $SS_{betweengroups}$ is the deviation of the group means from the *grand mean* for each person.
- That means the deviation of groups from the grand mean must be weighted by sample size.

$$
\begin{aligned}
SS_b &= \sum_{i=1}^{n}(\bar{Y}_{ij} - \bar{Y})^2 = \sum_{j=1}^{k} n_j(\bar{Y}_j - \bar{Y})^2 \\
&= (n_F) * (\bar{Y}_F - \bar{Y})^2 + (n_M) * (\bar{Y}_M - \bar{Y})^2 \\
&= 50 * (65.07 - 66.79)^2 + 50 * (68.50 - 66.79)^2 = 294.00
\end{aligned}
$$

ANOVA

Ta-Da!

- So we have all of the sums of squares! We've done ANOVA!

$$
\begin{aligned}
SS_y &= SS_{between} + SS_{within} \\
1841.18 &= 294.00 + 1547.18
\end{aligned}
$$

- Are there sex differences in height?
- Answer: 294.

Question Slide

- ANOVA is just a permutation of the GLM.
- Instead of talking about predictor variables, we're talking about splitting a sample into unordered (nominal) groups.
- Stuff we know:
  - We can get all of the components we need for ANOVA from sample statistics.
  - ANOVA just parses variance into stuff due to group differences and other variance.
- Stuff we don't know:
  - How to turn 294 into an answer.
- Questions?

Back to GLM

- The point I've been hammering on is that, as special cases of the GLM with specific types of data, regression & ANOVA are the same thing.
  - ANOVA is restricted to categorical (& unrelated) predictors, but otherwise is OLS regression.
  - As implied before, we can include categorical predictors in regression, using a coding scheme.
  - When R runs an ANOVA, it turns your equation into a regression, runs that, then translates it back.
- From here on out, I'll talk about the two models as a single model!
  - We'll work through the height example from both regression and ANOVA frameworks.
  - Next on the docket: model fit & inference.

Assessing fit

Decisions, decisions

- The decision framework typically employed in basic statistics is null hypothesis testing (NHT).
  - We formulate two mutually exclusive and exhaustive hypothesis about the relationship between predictor(s) and dependent variables.
  - The null hypothesis is often one of no relationship.
- How does it work?
  - We get some estimate of the IV-DV relationship.
  - We use the parameters to estimate the probability of a relationship this strong or stronger occurring if none actually existed.
  - We decide to retain or reject the null hypothesis of no relationship, based on a criterion.
  - This criterion is an error rate (α), saying that any outcome less likely than α leads to rejection of the null hypothesis.

Assessing fit

Height Data Redux

Height: ANOVA Version	Female	Male	Total
Mean	65.07	68.50	66.79
Var (MS)	16.12	15.46	18.60
n	50	50	100

Cov	Height	Sex
Height	18.60	
Sex	0.86	0.25
Means	66.79	0.50

$r_{Height,Sex} = 0.400$

- Here's the data again, with some additional information.
- So let's run the analyses, both from ANOVA and regression.

Assessing fit

Height Data Redux

ANOVA

$$\begin{aligned} SS_y &= SS_{between} + SS_{within} \\ 1841.18 &= 294.00 + 1547.18 \end{aligned}$$

Regression

$$\begin{aligned} Y &= b_0 + b_1 * Sex, \\ &\quad Sex = 0 \; for \; Females, 1 \; for \; Males \\ Y &= 65.07 + 3.43 * Sex \end{aligned}$$

- Ok, we've run ANOVA and regression.
- Let's run some tests and evaluate these models.

F-test

- The first test we'll run will be the F-test, which can be applied to either regression or ANOVA.
- What if there is no effect of the grouping variable?
  - In that case, we would expect the variance or MS between groups to be equal to the variance or MS within groups.
  - Alternatively, we would expect the ratio of the between group MS to the within group MS to be one.

$$\mathit{Test\ Statistic} = \frac{MS_b}{MS_w} = \frac{\frac{SS_b}{df_b}}{\frac{SS_w}{df_w}}$$

- Wait, isn't SS the sum of independent squared deviations from the normal distribution?
- That sounds a whole lot like the ratio of two χ^2 distributions. That's the F-distribution!

F-test

$$F - Statistic = \frac{MS_b}{MS_w} = \frac{\frac{SS_b}{df_b}}{\frac{SS_w}{df_w}}$$

- If there is no effect of grouping, we'd expect MS_b and MS_w to be equally valid estimates of the sample variance.
- If there is an effect, then MS_b will be bigger.
- The same logic applies to regression, too.

ANOVA Table

May be familiar

Effect	SS	df	MS	F	p
Between	294.00	1	294.00	18.62	$<.001$
Within	1547.18	98	15.79		
Total	1841.18	99			

- We've split *df* into df_b (1 because we're estimating one more mean with two groups than with one), and df_w (100 people-2 estimated means=98).
- We get an F ratio of 18.62, which on 1 and 98 degrees of freedom has a probability of .0000381369.
- If we set a criterion of .05, then we would reject the null hypothesis of no relationship.

Regression Table

May be familiar

Effect	Est	SE	t	p	β
Intercept	65.07				
Slope	3.43				

R^2 = ???, Residual Variance = 15.79, Residual df = 98.

- Here are the incomplete regression results.
- We do have a residual variance (15.79), which can become a residual sums of squares (RSS=15.79*98=1547.18).
- Hey, that is exactly the SS_w from the ANOVA! We had SS_y before we started, and we split up degrees of freedom based on parameters instead of means.

$$F - Statistic = \frac{\frac{SS_y - RSS}{df_{total} - df_e}}{\frac{RSS}{df_e}} = \frac{\frac{1841.18 - 1547.18}{99 - 98}}{\frac{1547.18}{98}} = \frac{\frac{294}{1}}{\frac{1547.18}{98}} = 18.62$$

Nesting

Its how you make a house a home

- The F-test was secretly our first test of *nested models*.
- Two models are nested when one is a special case of the other, or when all of the parameters in the littler one are in the big one.
- When this occurs, we can attribute all improvements in fit to the different parameters in the larger model.
- In the last model, we were actually comparing two models:
  - Null Model: Height = b_0 + e
  - Alt. Model: Height = b_0 + b_1 * Sex + e
- The null model is the alternative, with b_1 set to zero.
- Nested models can only be compared on the *exact* same data.

Nesting

Redoing the F-test

- Instead of a simple test, we can think of the F-test for regression as a way to compare two nested models.
- If the smaller model is N and the larger is A, then the F statistic becomes:

$$F - Statistic = \frac{\frac{RSS_N - RSS_A}{df_N - df_A}}{\frac{RSS_A}{df_A}} = \frac{\frac{1841.18 - 1547.18}{99 - 98}}{\frac{1547.18}{98}} = \frac{\frac{294}{1}}{\frac{1547.18}{98}} = 18.62$$

- The answer didn't change, because the residual variance of the null model (called an intercept-only model) is the variance in Y.

Other Nested Model Tests

The Likelihood Ratio Test

- One common test is the likelihood ratio test (LR).
  - Any model estimated with maximum likelihood gets a likelihood value.
  - The ratio of these two likelihoods provides another nested model test.
  - -2 times the natural log of this ratio is χ^2 distributed, with *df* equal to the difference in number of parameters.
  - Because of the properties of the logarithm, this is expressed more simply as -2LL$_A$- -2LL$_N$.
- Why use this over the F-test?
  - F-test only works for residual variances.
  - LR test can work for any model using ML.

$$-2LL = -2 * \log L(\beta|y, X) = n\left[\log\left(\frac{2\pi RSS}{n}\right) + 1\right]$$

Other Nested Model Tests

More Height Example

Model	-2LL
Intercept-Only	575.04
Sex	557.64

- χ^2_1 = 17.40, p=0.000030.
- What do you know, the same answer (within 6 digits of rounding error)!
  - All of your significance tests should come out the same, because they're all testing the same thing!
  - If they don't, check your math and assumptions.

Non-Nested Model Tests

More Height Example

- We've been discussing global model comparison of nested models.
- What if your models aren't nested? Here are two ML-based tools.
  - Akaike's Information Criterion (AIC), defined as -2LL + 2*number of parameters.
  - Bayesian Information Criterion (BIC), defined as -2LL + log(n)*number of parameters.
- Simply compare the values: lowest AIC or BIC wins.
- Compare as many models at once as you like.
- Never use AIC or BIC to compare nested models.

Parameter Model Tests

Standard Errors

- We've been discussing global model comparison.
- We also can test individual parameters using their standard errors.
  - Standard errors are the standard deviations of a sampling distribution.
  - We can then express any regression effect as so many standard errors (SDs) away from zero.
  - This is just a *t*-test (remember, the *t* distribution is just the normal distribution with a sampling correction).
- Caveat: judging the significance of effects based on their values can be affected by multicolinearity.

Regression Table

May be familiar...again.

Effect	Est	SE	t	p	β
Intercept	65.07	0.56	116.19	$<.001$	
Slope	3.43	0.80	4.29	$<.001$	.400

R^2 = .160, Residual Variance = 15.79, Residual df = 98.

- Here are the complete regression results.
- We can run a t-test on the slope coefficient, with *df* equal to the residual *df*.
- We get the same answer, because the only parameter we're really testing is the sex effect.

Actually assessing fit

It's about time!

- All of the tests we've run so far are simple accept-reject decisions.
  - They don't tell us how big an effect is, only if it's larger than we would expect by chance.
  - "Chance" will move with sample size.
- Null hypothesis testing is at best, incomplete, and at worst, evil.
  - Cohen's (1994) "The earth is round, $p < .05$" is a classic critique of NHT.
  - I think NHT should be a necessary but not sufficient condition for accepting a new hypothesis.
- We need tools that actually assess fit, typically referred to as effect size.

Confidence intervals

OK, not actually “fit" yet.

- Instead of simply stating a p-value, we can express regression effects in terms of confidence intervals.
  - Our parameter estimate or point estimate will be surrounded by a CI.
  - We’ll typically use a 95% CI, to match NHT conventions.
- CIs are often built using the standard error.
  - For a 95% CI, find the 97.5^{th} and 2.5^{th} percentiles of the reference (t) distribution.
  - Go that many standard errors above and below the parameter estimate.

Confidence Intervals

May be familiar...again again.

					CI		
Effect	Est	SE	t	p	Lower	Upper	
Intercept	65.07	0.56	116.19	$<.001$	63.96	66.18	
Slope	3.43	0.79	4.33	$<.001$	1.86	5.00	.400

R^2 = .160, Residual Variance = 15.79, Residual df = 98.

- We're 95% sure the intercept is between 63.96 and 66.18 inches.
- We're 95% sure the sex difference in heights is between 1.86 and 5.00 inches.
- We can talk about raw units a little more clearly here. Despite a ridiculously low p-value, we can still only nail down the sex differences in height to a 3.14 inch range.

Fit

Effect size

- Two other measures of effect size I want to discuss.
- Coefficient of determination: R^2
  - This is the proportion of variance explained in a regression, compared to the baseline model.
  - ANOVA calls this η^2.
- Standardzied regression weights:
  - Typically marked using β.
  - This is the regression weight if all variables are standardized.

Fit

Should be familiar.

					CI		
Effect	Est	SE	t	p	Lower	Upper	
Intercept	65.07	0.56	116.19	<.001	63.96	66.18	
Slope	3.43	0.79	4.33	<.001	1.86	5.00	.400

R^2 = .160, Residual Variance = 15.79, Residual df = 98.

- Sex accounts for 16% of the variance in height.
- Moving one standard deviation on the sex variable (.5 sexes, which makes no sense), corrosponds to a .4 SD change in height.
  - Alternatively, sex differences in height are .8 standard deviations.

Other Fun Regression Problems

Laundry List

- Correlated predictors (multicollinearity) is a problem.
  - As the correlations between predictors gets more extreme, it gets harder to tell them apart. If your predictors are correlated, it may be hard to assess fit in either variable.
  - ANOVA assumes independence of predictors.
- If any IV can be completely accounted for (is linearly dependent) on some combination of other IVs, regression breaks.
  - Example: if you try and predict weight from both height in inches and height in meters, you can never assign a regression weight to either of them without knowing the other. There's no unique solution.

Other Fun Regression Problems

Laundry List

- You can't have more predictors (and intercepts) than observations.
  - There's no point in explaining n pieces of information with n or more parameters.
  - When the number of observations and number of predictors (including intercept) are the same, model fit is perfect.
- There's still other assumptions we haven't talked about:
  - Violations of homoscedasticity and normality of residuals, error in predictors and proper specification are all important to varying degrees.
  - We'll get to them in the *Model Diagnostics* section.

Questions

- ANOVA and regression are equivalent provided the data match up.
- We need to rely on distributional tests to decide whether the effect of a predictor is greater than we would expect from chance.
  - There are more than a few of them.
- We need to make sure we answer “how big is the effect," not just “is it there?"
- Regression and ANOVA have some other assumptions we haven’t dealt with yet.

Nonlinearity & Interactions

- Not everything is a nice linear relationship.
  - Length, area and volume all have perfect & nonlinear relationships.
  - I'm sure there are examples in your data too.
  - So all we have to do is include nonlinear components.
- So what's nonlinearity doing in the general *linear* model?
  - The model defines linear relationships, but may specify linear relationships between nonlinear transformations.
  - We'll also discuss polynomial terms & interactions.

Nonlinearity & Interactions

- There are four general approaches to analyzing non-linear relationships:
  - Monotonic nonlinear transformations.
  - Polynomial terms/polynomial regression.
  - Nonlinear regression.
  - Nonparametric regression.
- So what should you use?
  - The first two are certainly the most common.
  - Nonlinear regression and nonparametric regression are topics for another day.

Nonlinearity & Interactions

Transformations

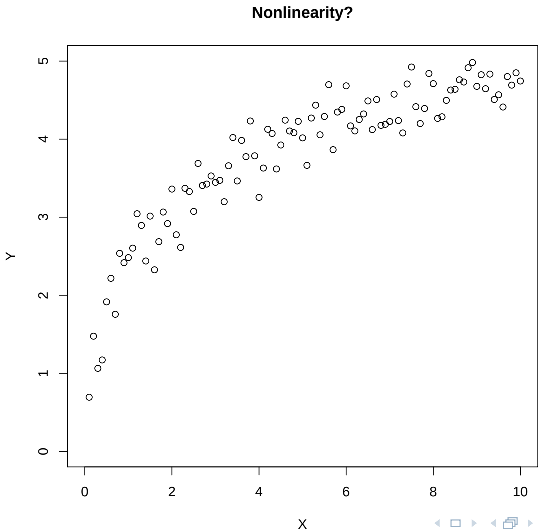
- Why bother?
  - Incorrect model specification is an assumption violation. Think of it like leaving out a predictor.
  - Transformed variables can show you relationships you wouldn't have otherwise seen.
  - It's really not that hard!
- So how do we do it?
  - Add new variables to the model that are functions of existing variables.

Nonlinearity & Interactions

Linear Transformations

- What's a linear transformation?
  - It's a transformation you make either by adding a constant to a variable, multiplying it by something, or both.
  - It's that equation for a line again (X_{new}=m*X_{old}+b).
  - It will have zero effect on model fit.
- Why do it?
  - It can aid interpretation, especially for polynomials and interactions.
  - Centering and standardization are two forms of linear transformations.

What do we do with this?



What do we do with this?

- First come up with some transformations to test.
  - I'll test a linear model, as well as a model with the IV squared and the log of the IV.
- So how do we do it?
  - We'll fit models with each of these variables, compare them, and pick the best fitting.

First we pick a model

Model	-2LL	df	AIC
Intercept	283.11	2	287.11
x	157.29	3	163.29
log(x)	44.10	3	50.10
x + x^2	98.31	4	106.31
x + log(x)	43.76	4	51.76
x + x^2 + log(x)	43.76	5	53.76

- The natural log of x is the clear winner here.
- We should compare everything with the intercept model, and anything with x to the simple x model, but none are close to the critical value (95^{th} percentile for χ^2_1=3.84).
- *df* includes a term for residual variance, which is consistent with GLM/SEM specification of regression. It just adds 1 to every df calculation.

Then we analyze it

					CI		
Effect	Est	SE	*t*	p	Lower	Upper	
Intercept	2.42	0.05	45.20	$<.001$	2.31	2.53	
log(x)	1.03	0.03	31.17	$<.001$	0.96	1.09	.95

R^2 = 0.91, Residual Variance = 0.30, Residual df = 98.

- This model fits *really* well.
- You will never get a fit like this without simulated data.
- How do we interpret this?

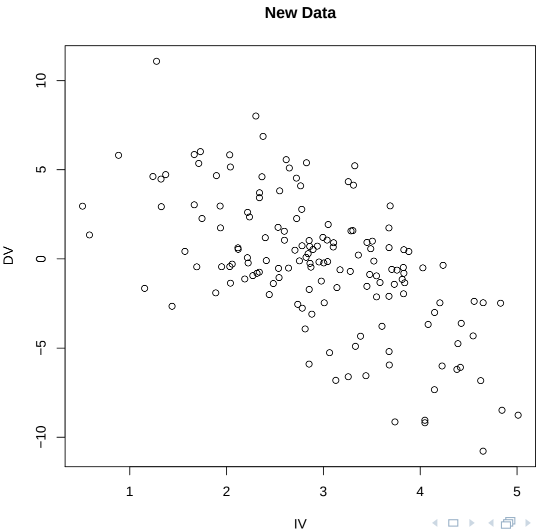
Picking Transformations

- How can one go about picking how to transform data?
  - Sight. Plot the data, see what you see.
  - Theory.
  - Common non-linear functions: logarithms, powers, trigonometric functions.
- Polynomial terms are a flexible way for including curvature in effects.
  - Add additional transformations of the linear effect, starting with squaring (X^2).
  - One version of power transformations, or power polynomials.

Polynomial Terms

- Power polynomials, or polynomial regression, involves adding power transformations of a variable in a specific order.
- One begins with a simple linear model
  - $\hat{Y} = b_0 + b_1 X$
- Then, add additional variables to the model, with the new variables being X raised to the next power.
  - $\hat{Y} = b_0 + b_1 X + b_2 X^2$
  - $\hat{Y} = b_0 + b_1 X + b_2 X^2 + b_3 X^3$
  - $\hat{Y} = b_0 + b_1 X + b_2 X^2 + \ldots + b_j X^j$
- Use nested model comparisons and residual checks (coming soon) to select a model.

Polynomial Terms



First we pick a model

Model	-2LL	df	δ-2LL	δdf
Intercept	822.14	2		
x	750.15	3	71.99	1
x + x^2	744.73	4	5.42	1
x + x^2 + x^3	744.71	5	0.02	1

- Adding X improves fit over an intercept only model,
  - χ^2_1 =71.99, p$<$.001.
- Adding X^2 improves fit over a linear model,
  - χ^2_1 = 5.42, p=.020.
- Adding X^3 improves fit over a quadratic model,
  - χ^2_1 = 0.02, p=.889.

Let's look at that model

					CI		
Effect	Est	SE	t	p	Lower	Upper	β
Intercept	3.34	1.81	1.86	0.07	-0.24	6.92	
X	0.40	1.28	0.31	0.76	-2.12	2.92	-0.10
X^2	-0.50	0.22	-2.33	0.02	-0.93	-0.08	-0.73

R^2 = .403, Residual Variance=2.926, Residual df=147.

- How should we interpret this?

Other Notes about Polynomials

- You should always include all lower order terms.
  - If you have a quadratic term, you should include the linear as well, regardless of its predictive ability.
  - Higher order terms are only reflective of whatever curvature they display if all lower terms are present.
- Polynomial terms tend to be multicollinear.
  - In the last analysis, X and X^2 had a .98 correlation.
  - Odd-even terms (say X and X^2) can be forced to zero covariance by centering.
  - This won't affect odd-odd or even-even relationships.

Let's look at that model

I've centered X

					CI		
Effect	Est	SE	t	p	Lower	Upper	β
Intercept	0.15	0.30	0.50	0.62	-0.45	0,75	
X	-2.56	0.26	-9.86	$<.001$	-3.08	-2.05	-0.63
X^2	-0.50	0.22	-2.33	0.02	-0.93	-0.08	-0.15

R^2 = .403, Residual Variance=2.926, Residual df=147.

- First, I subtracted the mean from X. Then I squared it to create X^2.
- Notice what did change (Intercept, X and βs).
- Notice what didn't (model fit and X^2).

Interactions

- The last topic for this section is interaction.
- What is interaction?
  - It's typically shorthand for multiplicative interaction, although there are other types.
  - It's also known as moderation (Baron & Kenny, 1986).
  - It's a way to have the effect of one IV depend on other IVs.
  - People tend to have difficulty interpreting them.
- Let's look at one, shall we?

Interactions

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 + e$$

- What's going on here?

Interactions

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 + e$$

- What's going on here?
  - We've added a new variable, which is the product of X_1 and X_2.
  - This variable is very often difficult to interpret, in no small part because the units are ridiculous (inch-pounds, anyone?).
  - And because this variable is not only correlated with each of the variables that it consists of, but you literally can't affect X_1 without affecting $X_1 X_2$, neither X_1 nor X_2 can be interpreted independently.
  - What to do now?

Interactions

$$Y = b_0 + b_1X_1 + (b_2 + b_3X_1)X_2 + e$$

- Let's throw some parentheses in.
  - The effect of X_2 depends on X_1.

Interactions

$$Y = b_0 + b_2X_2 + (b_1 + b_3X_2)X_1 + e$$

- Let's throw some parentheses in.
  - The effect of X_1 depends on X_2.

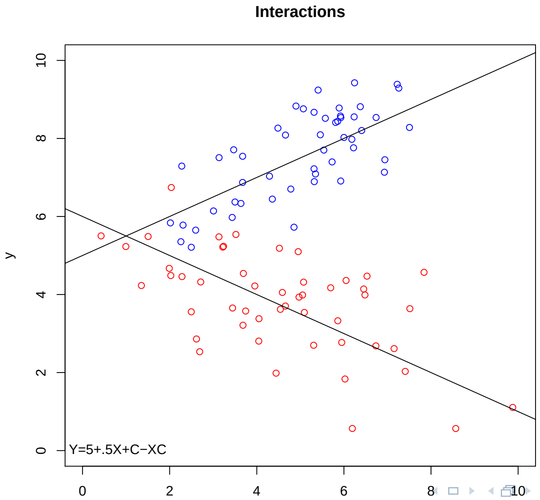
Interactions

$$
\begin{aligned}
Y &= b_0 + b_1 X_1 + (b_2 + b_3 X_1) X_2 + e \\
Y &= b_0 + b_2 X_2 + (b_1 + b_3 X_2) X_1 + e
\end{aligned}
$$

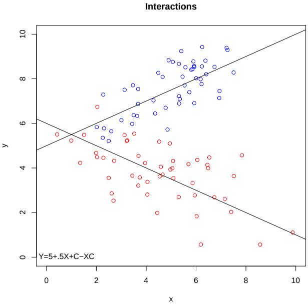
- Now we can interpret the interaction easier.
  - The effects of either X variable depend on the other one.
  - You wouldn't worry about describing b_1 or b_2 by themselves, because that's only part of the effect!
- Let's look at some graphs to see if this sinks in.

Interactions

X is Interval, C is Categorical



Interactions



- This is *the* classic example.
- What's the effect of X on Y?
  - The effect is 0.5-c.
  - The "intercept" is 5+c.
  - Many people break this into two regression equations (one for c=0, another for c=1), but I think that makes generalizing to higher order interactions difficult.
- We could also interpret the effects of c (1-X, with an "intercept" of 5+.5X).

Question Slide

- Just because we're using the GLM doesn't mean that we can't model nonlinear effects.
  - We can add nonlinear transformations, either one-off or polynomial versions.
  - We can use multiplicative interactions to create dependencies between the effects.
- Just be careful about interpretation!
- Questions?

Assumptions and Detecting Violations

- Regression and ANOVA have some specific assumptions.
- If you meet them, fantastic!
- If not, that’s a problem at some level.
- Now we’ll talk about:
  - What they are,
  - How to detect these problems, and
  - What to do if there’s a problem.

What are the assumptions of regression?

Common Ones.

- Residuals are normally distributed.
  - This is related to the OLS & ML estimators.
  - If we're minimizing squared residuals, we're minimizing variance. If residuals aren't normal, this doesn't work.
  - OLS regression is fairly robust to deviations to normality provided the other assumptions are met.
  - Q-Q plots & histograms are your best tools.
- Residuals are homoscedastic.
  - We only have one residual/error term, and it needs to be equally appropriate over the range of X.
  - Plotting residuals against X is likely the best method.
  - Be careful when the distribution of X isn't symmetric.

What are the assumptions of regression?

Uncommon Ones.

- No error in the predictors.
  - All of the “error" is assumed to be in the dependent measure.
  - This can lead to biased or suppressed estimates of the XY effect, especially standardized estimates.
  - Very commonly violated, and rarely talked about (it’s pretty robust).
- Independent variables are correctly specified.
  - You have to correctly specify the variables (i.e., include all of the important ones).
  - All relationships must be linear.
- Residuals are independent.
  - No clustered and longitudinal data.

So let's start looking at residuals!

- The short answer to all of these assumptions is to plot your data from every angle and see what you find.
  - We typically look at bivariate plots, involving one independent variable and either a residual or a dependent variable.
  - 3D graphs are typically difficult to read, particularly for diagnosis.
  - Higher order dimensionality is a little tougher.
- What do you look for?
  - Extreme observations.
  - Non-linearity of effects.
  - Unusual distributions.
- You just have to practice.

Extreme Observations

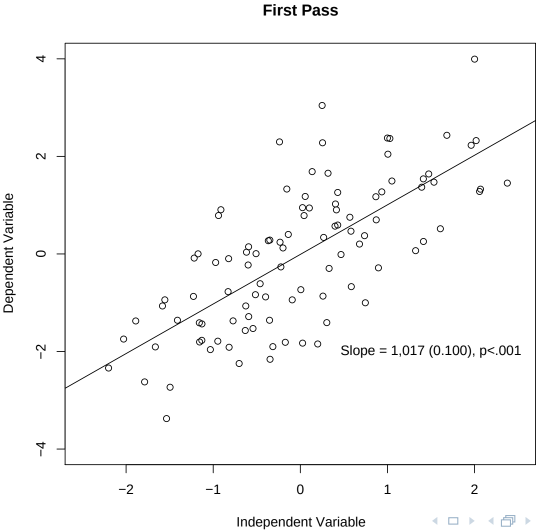
I think you mean Xtreme!

- The most common discussion point for residual checks are extreme observations, or outliers.
- We can talk about outliers in a few different ways:
  - *Leverage:* How extreme is this observation's set of predictor variables?
  - *Distance:* How far is an observation from the regression line (How big is the residual)?
  - *Influence:* Combining distance and leverage, how much can this observation move the regression line/plane/n-space term for MR prediction.
- There are approximately 11 pounds of statistics used for the analysis of extreme observations.

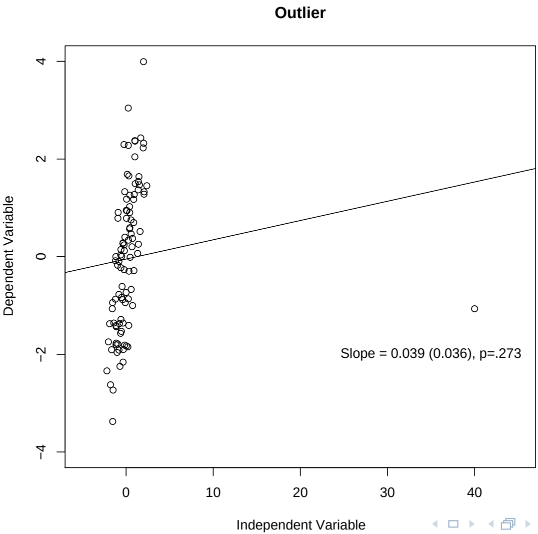
Simple Illustration

- Let's see how big a deal an extreme observation can be before we get started.
  - We'll also hint at future statistical methods for outlier detection.
- I'll make up some data for a simple regression involving x and y.
- Then we'll add an outlier and see what happens.

Simple Illustration



Simple Illustration



Simple Illustration

- The slope went down a bunch, loosing significance in the process.
  - Difference of .978 in raw units (1.017-.039).
  - Difference of .606 in standard units (.717-.111).
- Model fit, as judged by R^2, fell of the table (.516-.012).
- Oh, and we violated a bunch of assumptions.
- How do we deal with it?
  - Let's diagnose the problems first. They won't all be this obvious.
- In general, residual diagnostics are calculated for every observation.

Leverage

- Leverage describes how unusual an observation is with respect to its set of X variables.
- Leverage is often described by the statistic h_{ii}.
- For one independent variable, leverage can be defined as:

$$Leverage = h_{ii} = \frac{1}{n} + \frac{(X_i - \hat{\mu}_i)^2}{\sum x^2}$$

- Centered leverage (h_{ii}^*) is calculated as:

$$h_{ii}^* = h_{ii} - \frac{1}{n} = \frac{(X_i - \hat{\mu}_i)^2}{\sum x^2}$$

- For multivariate predictors, one must consider distance from the centroid (set of means) and the correlations between variables.

Leverage

- Greater leverage values have a greater potential to affect the regression equation.
  - Think of the regression line as a lever that each data point can push.
  - Higher leverage=greater potential to push the line.
- Which kind of leverage to use?
  - Traditional accounts for sample size, such that a large deviation in a small sample has more leverage than the same deviation in a large sample.
  - Centered is easier to understand, and more useful in other formulas.
  - Leverage usually refers to the basic h_{ii}.

Leverage

Mathy Stuff. Impress your friends.

- This h is related to something called a hat matrix, which is a crucial part of error variance calculation.
  - $\epsilon_i = \sigma^2(1 - h_{ii})$
  - The sum of all of the h_{ii} values is n.
- Mahalonobis Distance is a related measure:
  - $(n - 1) * h_{ii}^*$
  - Just in case you see it.

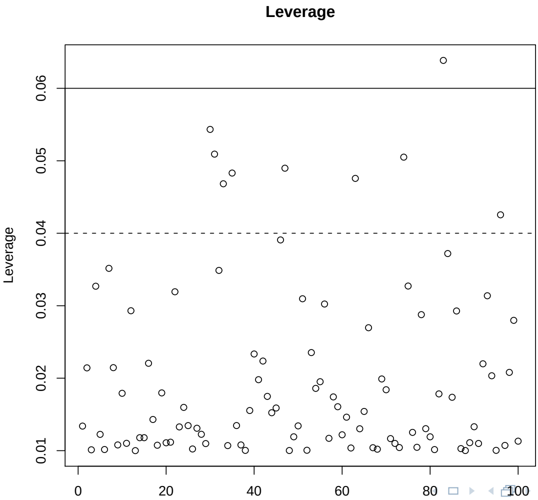
Leverage

Making R work for you.

- You can get leverage and some other residual diagnostics from the `influence()` function.
  - This function returns several things, one of which is the hat results.
  - You run this function on an `lm()` object.
  - `model<-lm(y~x)`
  - `influence(model)$hat` OR `hatvalues(model)`
- We also need a criterion.
  - 2(p+1)/n (Belsley, Kuh & Welsch, 1980); good benchmark.
  - 3(p+1)/n (Stevens, 1992); more stringent for smaller samples.
  - You should *consider* deleting or otherwise dealing with residuals above these cut-offs. It's not automatic.

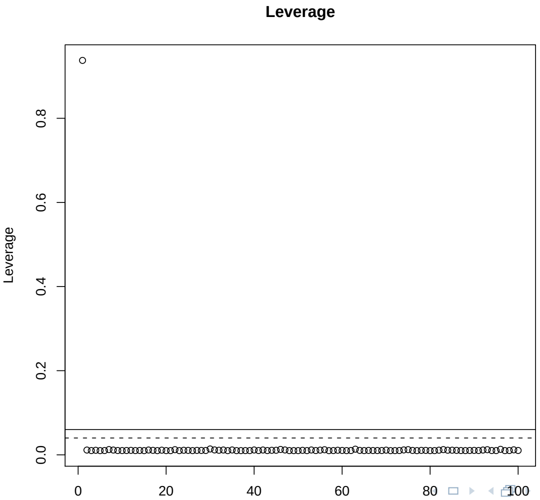
Plotting Leverage Values

Original Example



Plotting Leverage Values

With an outlier



Question Slide

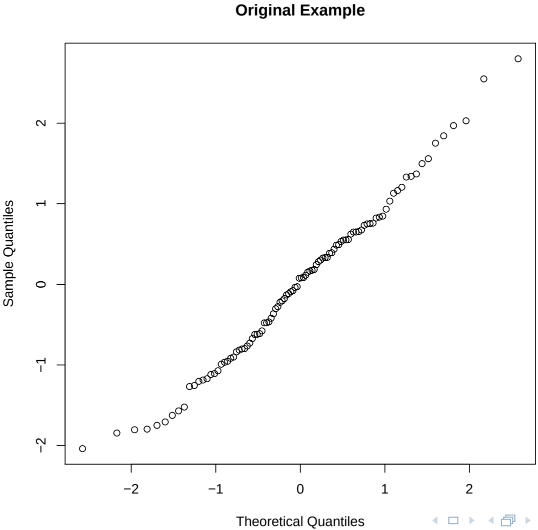
Summing up Leverage

- If you want some measure of extremity with regards to your independent variables, leverage (h_{ii}) is it.
  - All of the leverage values in your dataset sum to n.
  - We have a few criteria for when leverage is pretty high.
- It's not the end-all, though.
  - You're looking for one (or maybe a few) observations very out of sync with the rest of your data.
  - We still haven't talked about the dependent variable or residuals yet; these are just descriptions of our predictor variables.

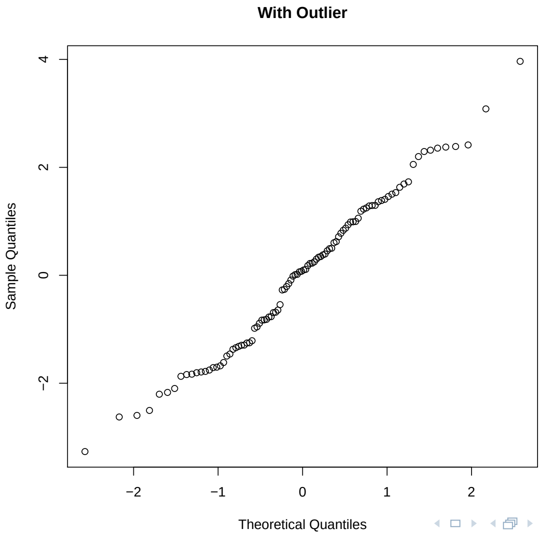
Distance

- The simplest way of talking about extreme residuals is distance, or just how big the residual is.
  - Every observation as a residual, which we can calculate and analyze.
  - We can analyze them in raw units, if those are meaningful.
  - $\epsilon_i = y_i - \hat{y}_i$
- What if we don't want to look at raw units?
  - We can (and will) look at some type of standardized residual, but the presence of outliers presents a problem.

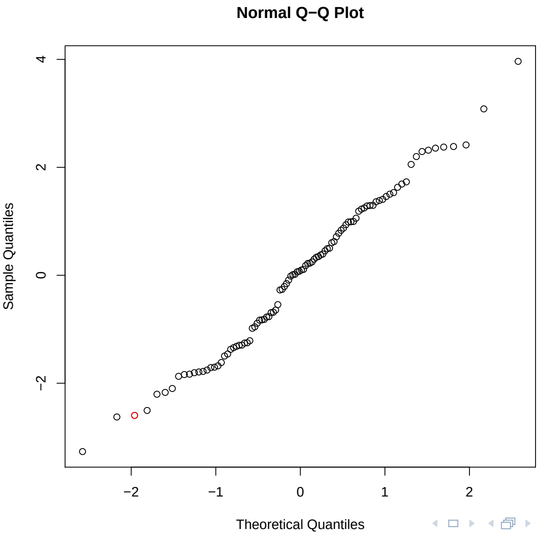
Distance


Original Example
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2

Distance


With Outlier
Sample Quantiles
Theoretical Quantiles
4
2
0
−2
−2
−1
0
1
2

Distance


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
4
2
0
−2
−2
−1
0
1
2

Distance

- What if we don't want to look at raw units?
  - We can look at some type of standardized residual, but the presence of outliers presents a problem.
  - We can get around this by looking at Studentized residuals, named in honor of William Gosset (Brilliant!).
- There are two types of studentized residuals:
  - Internal Studentized Residuals (shorthand: Standardized Residuals) compare any one residual to the entire set including itself.
  - External Studentized Residuals (shorthand: Studentized Residuals) compare any one residual to the entire set excluding itself.

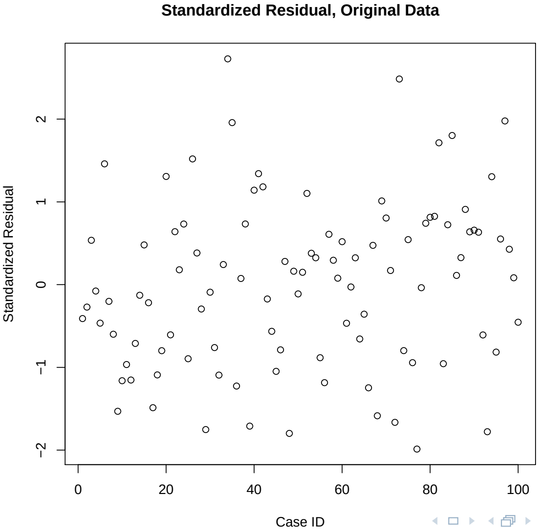
Studentized Residuals

- Internally Studentized Residuals compare the value of a residual to the residual variance.
  - $ISR = \frac{e_i}{\sigma_e \sqrt{1-h_{ii}}}$
- Pro:
  - Easy to calculate and provides the purest degree of standardization.
- Con:
  - There's no distribution that will be followed when no outliers exist. It's purely for inspection.

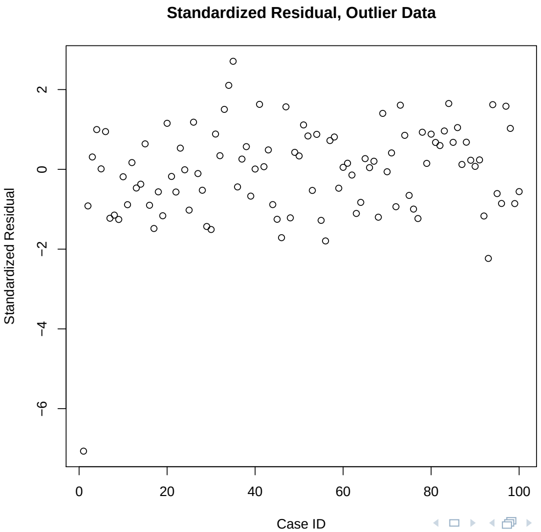
Studentized Residuals

- Externally Standardized Residuals compare a residual to the residual variance when that case is excluded.
- The regression is then *recalculated* without that case, and the residual in question is compared to the new regression line.
- Because of the properties of h_{ii}, there's a way to do this without actually estimating that line.
  - R will do it for you.
  - `rstandard()` will run the ISRs,
  - `rstudent()` will run the ESRs (Studentized).
- When will this make a difference?
  - When an outlier moves a regression line a lot, ESRs will catch it when ISRs don't.
- Bonus: this will be *t*-distributed, with (n-k-1) df.

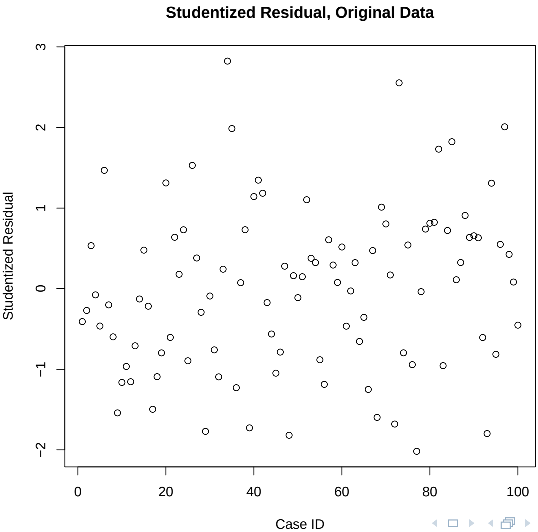
Internally Standardized Over Case ID



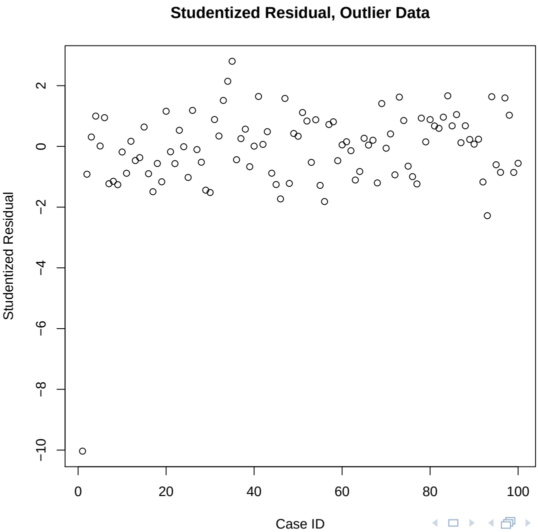
Internally Standardized Over Case ID



Externally Standardized Over Case ID



Externally Standardized Over Case ID



Distance Questions

- The second way to talk about extreme observations is the size of the residual.
- Standardization and Studentization are ways of turning residuals from a raw scale to one with known or interpretable distributional properties.
  - `rstudent()` and `plot.lm()` will be your friends here.
- Now we just need some ways to talk about these things together.
- What observations have a lot of leverage (extremity on X) and a lot of distance (extremity on Y)?

Influence

- Influence is a very well named term. It describes the degree to which any observation affects the regression line.
- Why are we talking about this?
  - Because we use regression to describe a population.
  - If the entirety of an effect is related to one or a few observations, we want to know.
- There are several different measures of influence.
  - Cook's *D*
  - DFFITS
  - DFBETAS

Influence

- Cook's *D* or Cook's Distance is a global measure of influence, meaning it measures how much an observation influences the model as a whole.

$$Cook'sD = \frac{\sum(\hat{Y} - \hat{Y}_{(i)})}{(k+1)\sigma_e^2} = \frac{h_{ii}}{1-h_{ii}} ISR_i^2(k+1)$$

- $\hat{Y}$ refers to the predicted value of Y for person i when they're included in the model.
- $\hat{Y}_{(i)}$ refers to the predicted value of Y for person i when they're omitted from the model (that's what the parentheses around i mean).
  - This looks a little like Studentized Residuals and a little like leverage, because it's proportional to their product.
- `cooks.distance()`, again used on an `lm()` object.

Influence

- DFFITS is a closely related function to Cook's D.

$$DFFITS = \frac{\sum(\hat{Y} - \hat{Y}_{(i)})}{\sqrt{\sigma^2_{e_{(i)}} h_{ii}}}$$

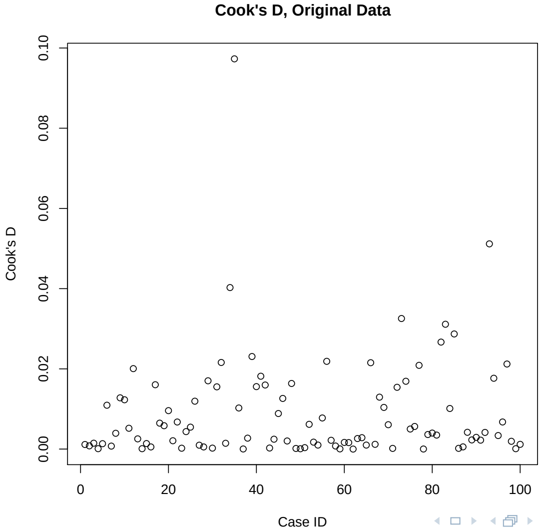
- We're scaling in terms of leverage instead of parameters (h_{ii} instead of k).
- The residual variance is from the new model (omitting observation i), rather than the full model.
- It's more rare, and can be estimated from Cook's D.
- `dffits()`

$$Cook'sD = \frac{(DFFITS)^2 \sigma^2_{e_{(i)}}}{(k+1)\sigma^2_e}$$

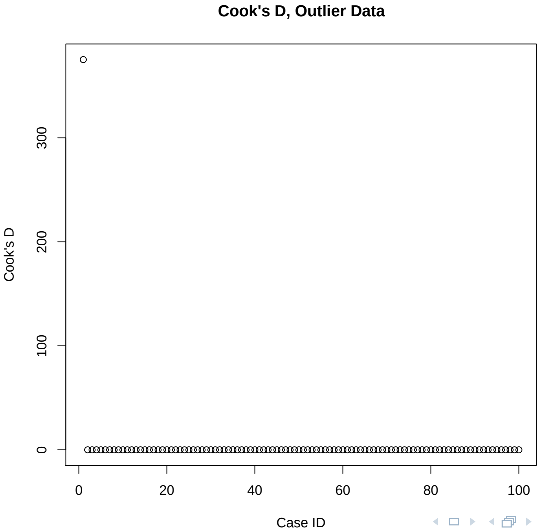
Influence Criteria

- Cook's D criteria:
  - 1.0, or:
  - The median of an F-distribution (p=.50) for (k+1,n-k-1) df.
- DFFITS criteria:
  - 1.0, or:
  - $2 * \sqrt{\frac{k-1}{n}}$
- In all cases, higher values are more influential observations.

Cook's Distance Over Case ID



Cook’s Distance Over Case ID



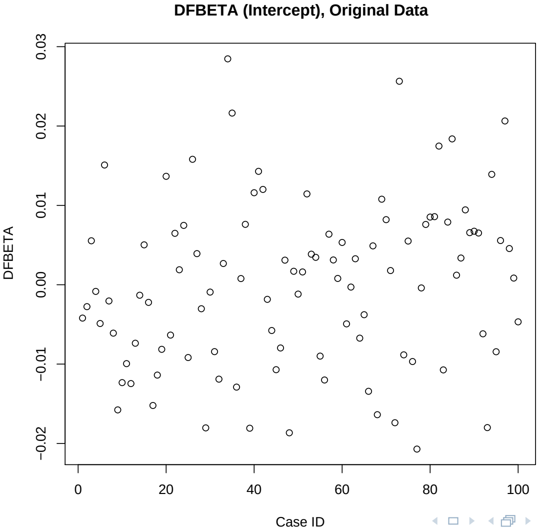
Specific Measures of Influence

- Cook's D and DFFITS deal with global measures of influence.
- What if you care about specific regression coefficients in multiple regression?
- We have DFBETAS for this, which we calculate for each person (i) and coefficient (j):

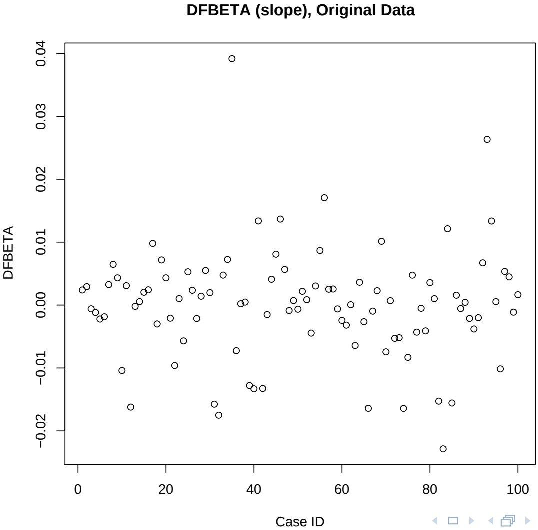
$$DFBETAS_{ij} = \frac{\beta_j - \beta_{j(i)}}{SE_{B_{(i)}}}$$

- While this resembles a *t*- or *z*-statistic, common usage treats it more like an effect size, using ± 1 or $\pm\frac{2}{\sqrt{n}}$ as criteria values.
- `dfbetas()` is your command. `dfbeta()` gives raw differences, as does `influence()$coefficients`.

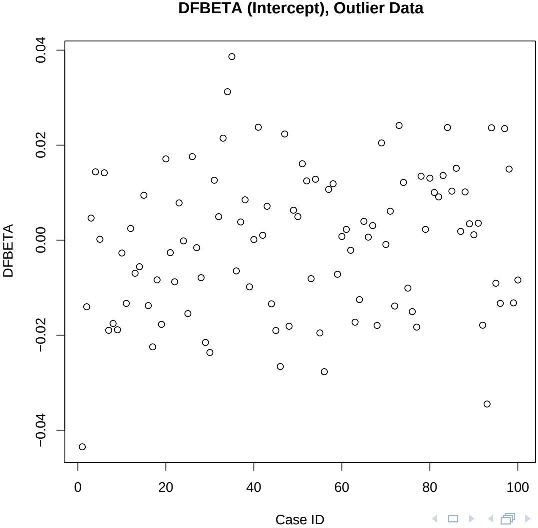
DFBETAS Over Case ID



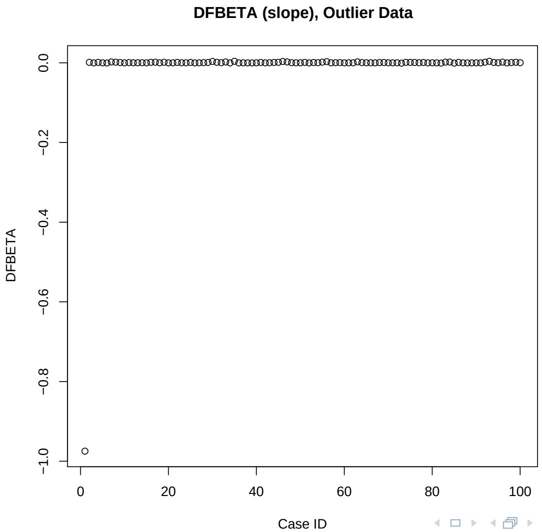
DFBETAS Over Case ID



DFBETAS Over Case ID



DFBETAS Over Case ID



Influence Summary

Questions

- So now we can talk about how extreme observations affect our model.
  - We can combine leverage and (more prominently) distance to describe the influence of every observation on the overall model.
- Cook's Distance is a fairly standard measure of influence.
  - DFFITS is a rarer but equivalent measure, with similar criterion values.
- DFBETAS allows you to look at how individual parameters are affected by extreme observations.
  - Which helps you understand your data a little better.

What to do with outliers?

It Depends! HAHAHAHA.

- Ok, now what?
- Treat as “contaminated observations."
  - Just delete, because you think it’s due to an error or problem.
- Delete, treating these observations as not representative of your data.
- Respecify the model to improve fit, usually through additional variables and transformations.
- Switch to a robust regression approach (LAD, or least absolute deviation regression)

The case for deletion

It's easy!

- If its reasonable to attribute an outlier to a coding problem, participant error or equipment malfunction, there's no problem deleting them.
  - These can be attributed to MCAR in the case of error.
  - If your participant clocked out on you, then that's a sampling issue.
- Just deleting problem people is more of an ordeal.
  - By deleting someone, you're arguing that they're not representative of the population you're studying.
  - You have to be careful that the population you end up with is not "the one who supports my hypothesis."

You can respecify your model

That sounds suspiciously like “work"

- An extreme observation (or observations) can be indicative of several things:
  - Sampling issue or error.
  - A violation of the linearity or specification assumption.
- Different ways you can deal with this:
  - Interaction terms.
  - Transforming one’s predictors.
  - Both of which we’ve just covered.

Questions

Diagnostics

- Regression has assumptions.
  - When we violate them, we can usually spot it in residual diagnostics.
- When we spot a violation (either through simple visual inspection or a diagnostic tool), we have to fix it.
  - Either delete the case or change the model.
- Assumption checking is an important part of running regression or ANOVA.
  - Just because residual checks aren't often published doesn't mean they're not important.

Closing Up

We're done!

- We've covered a whole lot today.
  - Regression and ANOVA are equivalent versions of the GLM.
  - Use whichever you want!
  - I think that regression is the more flexible technique, particularly when you get into complex modeling extensions.
- We can fit models and diagnose them.
  - Use global tests for nested models, and whenever possible.
  - Use model diagnostics to check your models. You may require more terms, transformations, or just a different model.

Closing Up

Stuff we didn't cover

- There's a lot of stuff I didn't even touch.
  - Logistic & Poisson regression.
  - Autoregressive & Difference score models.
  - Automated model selection.
  - Missing Data.
  - Mixed effect/random effect models.
  - And much much more.
- And you should learn these. Why?
  - You shouldn't use the wrong model for your data just because you're most familiar with that model.
- All models are wrong; some are useful.

Closing Up

Books

- Texts:
  - Cohen, Cohen, Aiken & West: Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.
  - Hays: Statistics for Psychologists (out of print, I think).
  - Howell: Statistical Methods for Psychology.
  - Maxwell & Delaney: Designing Experiments and Analyzing Data.
- And don't forget:
  - Wolfram MathWorld.
  - Wikipedia & Google.

Thank you!

- Friday, July 31: SEM.
- Friday, August 7: Survival Analysis.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003.

Stevens, S. S. (1939). Operationism and logical positivism. *Psychological Bulletin*, *36*, 221–236.