Lecture 11: Data transformations

Criminology 1200

Prof Maria Cuellar

University of Pennslyvania

My diagnostic plots look bad, now what?

You could try using a more complex model. But first, try transforming the data **before** fitting it, by using an invertible tranformation (powers and logs) of your response variable (y). Then, you can transform x as well.

- **Transforming a variable** involves using a mathematical operation to change its measurement scale. This is useful in linear regression because you can use the transformation to fit a model better, and then untransform the results.

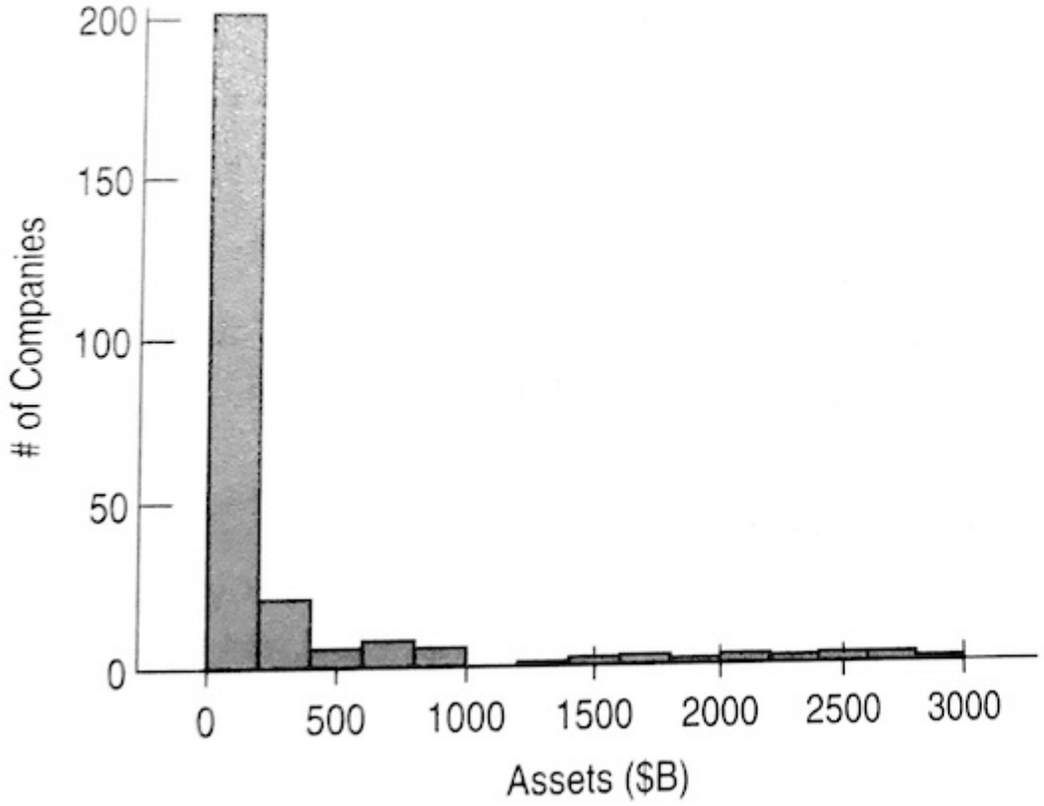
- There are many types of transformations, but the most common ones are taking the log of y and squaring y.

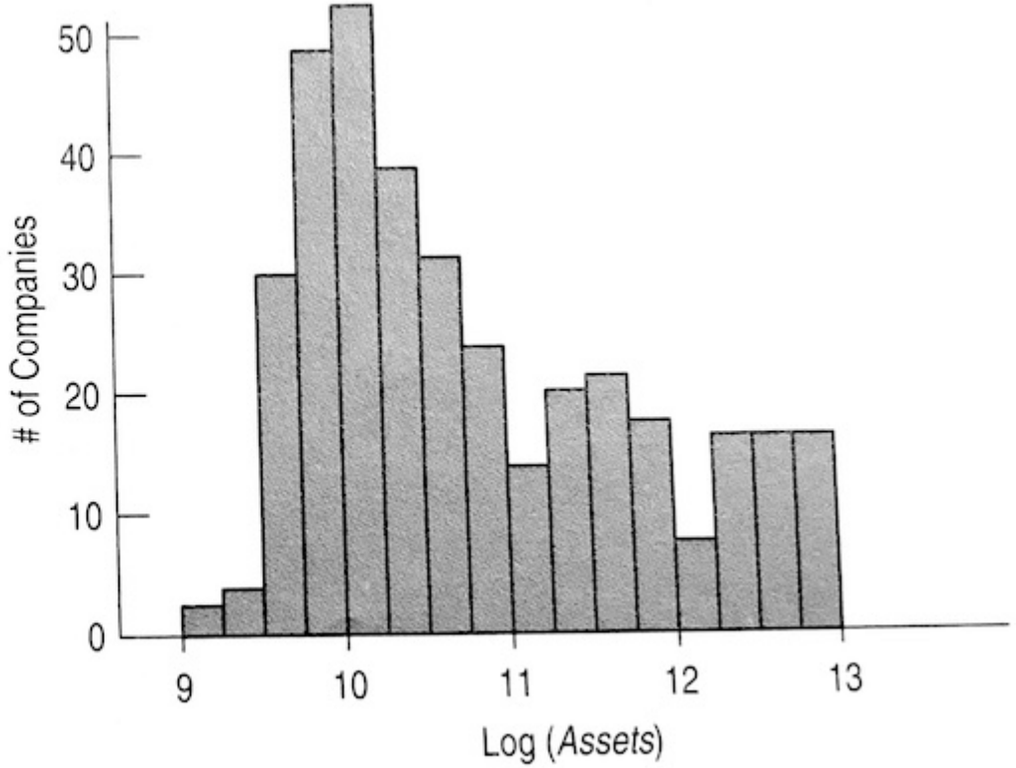
Transformations to achieve linearity

Method	Transform	Regression equation	Predicted value ($\hat{y}$)
Standard linear regression	None	$y = b_0 + b_1x$	$\hat{y} = b_0 + b_1x$
Exponential model	$DV = \log(y)$	$\log(y) = b_0 + b_1x$	$\hat{y} = 10^{b_0 + b_1x}$
Quadratic model	$DV = \text{sqrt}(y)$	$\text{sqrt}(y) = b_0 + b_1x$	$\hat{y} = (b_0 + b_1x)^2$
Reciprocal model	$DV = 1/y$	$1/y = b_0 + b_1x$	$\hat{y} = 1 / (b_0 + b_1x)$
Logarithmic model	$IV = \log(x)$	$y= b_0 + b_1\log(x)$	$\hat{y} = b_0 + b_1\log(x)$
Power model	$DV = \log(y)$ $IV = \log(x)$	$\log(y)=$ $b_0 + b_1\log(x)$	$\hat{y} = 10^{b_0 + b_1\log(x)}$

The last column shows the "back transformation" equation used to restore the dependent variable to its original, non-transformed measurement scale.

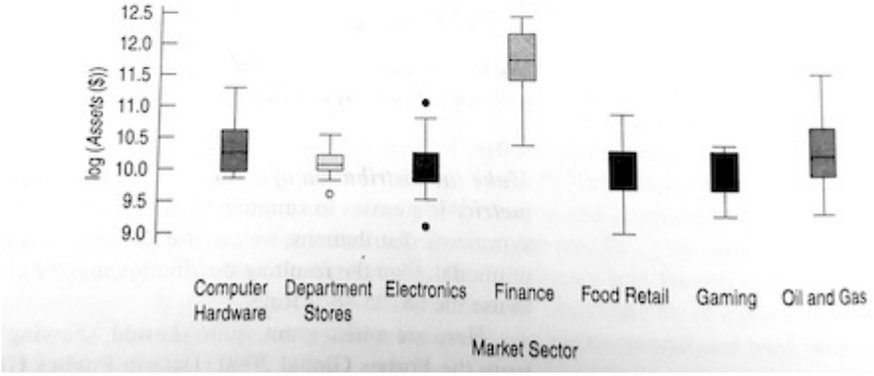
How do transformations look in the EDA?





How do transformations look in the EDA?

Taking logs makes the individual boxplots more symmetric and gives them spreads that are more nearly equal. When a re-expression simplifies the structure in several groups at once, it seems to be a natural choice for those data.



Steps to perform a transformation to achieve linearity

1. Conduct a standard regression analysis on the raw data. Draw the residuals vs. x and calculate the coefficient of determination, R^2. If the data look linear, do not transform it. If it does not, continue.

2. Choose a transformation method. Transform the independent variable, dependent variable, or both.

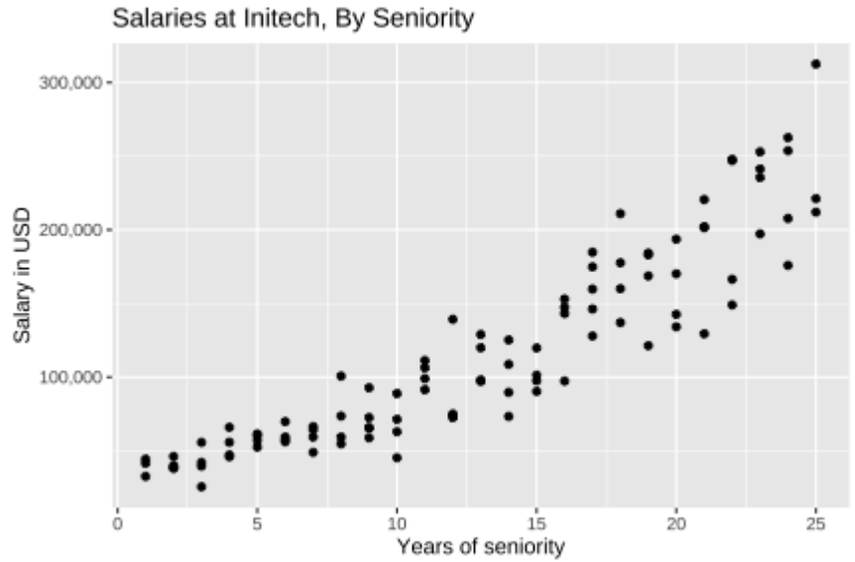
3. Repeat step 1, by running the regression analysis on the transformed data. If the transformed data looks more linear, the transformation was successful. Congratulations! Now the relationship between the independent variable (x) and the transformed dependent variable (e.g., log(y)) is linear. If not, try a different transformation method.

4. Now that you've transformed the data, how do you interpret the coefficients? You need to "undo" the transformation. See column 4 of the table in the last slide.

Example

How transforming the data can help fix the diagnosics

Let’s look at some (fictional) salary data from the (fictional) company Initech. We will try to model salary as a function of years of experience. Here is the scatterplot:

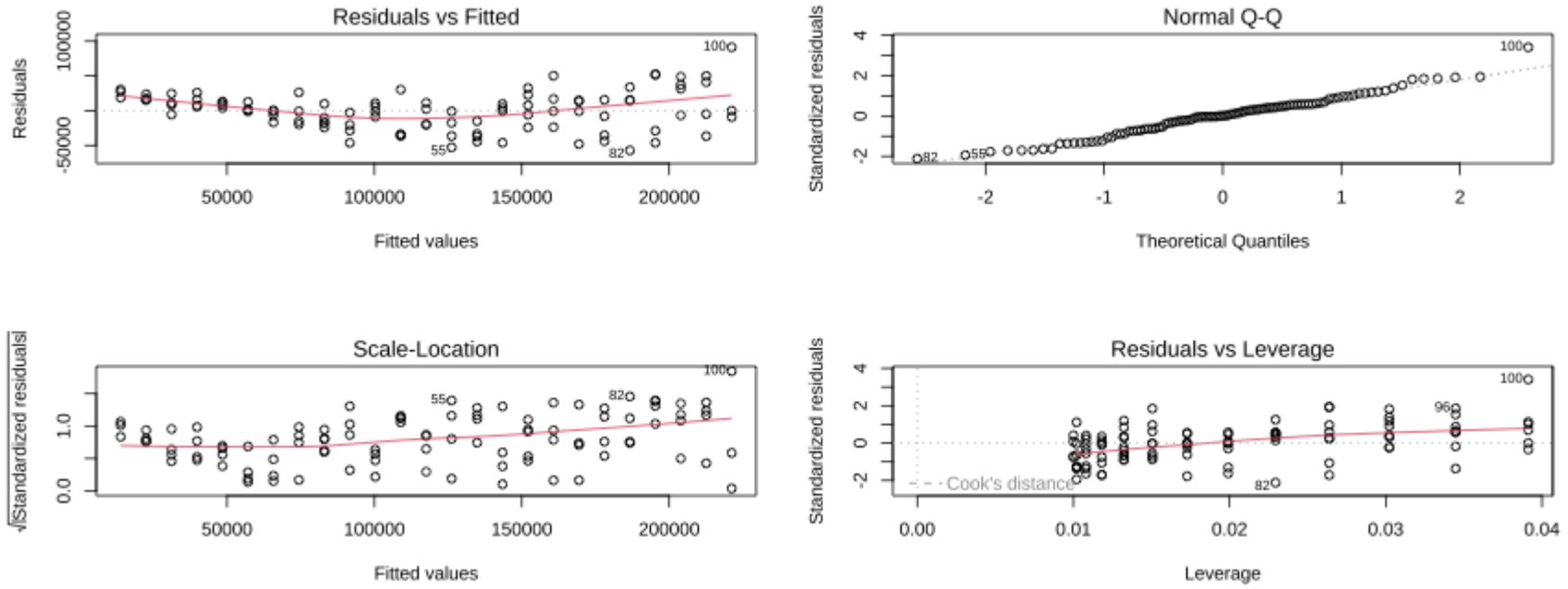


Untransformed linear model

This model appears significant, but does it meet the model assumptions?

```
##
## Call:
## lm(formula = salary ~ years, data = initech)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57225 -18104    241  15589  91332
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)     5302       5750   0.922               0.359
## years           8637        389  22.200 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27360 on 98 degrees of freedom
## Multiple R-squared:  0.8341,    Adjusted R-squared:  0.8324
## F-statistic: 492.8 on 1 and 98 DF,  p-value: < 0.00000000000000022
```

Diagnostics don't look so good



From the fitted versus residuals plot it appears there is non-constant variance. Specifically, the variance increases as the fitted value increases.

Variance-stabilizing transformation

- We will now use a model with a log transformed response for the Initech data,

$$log(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i.$$

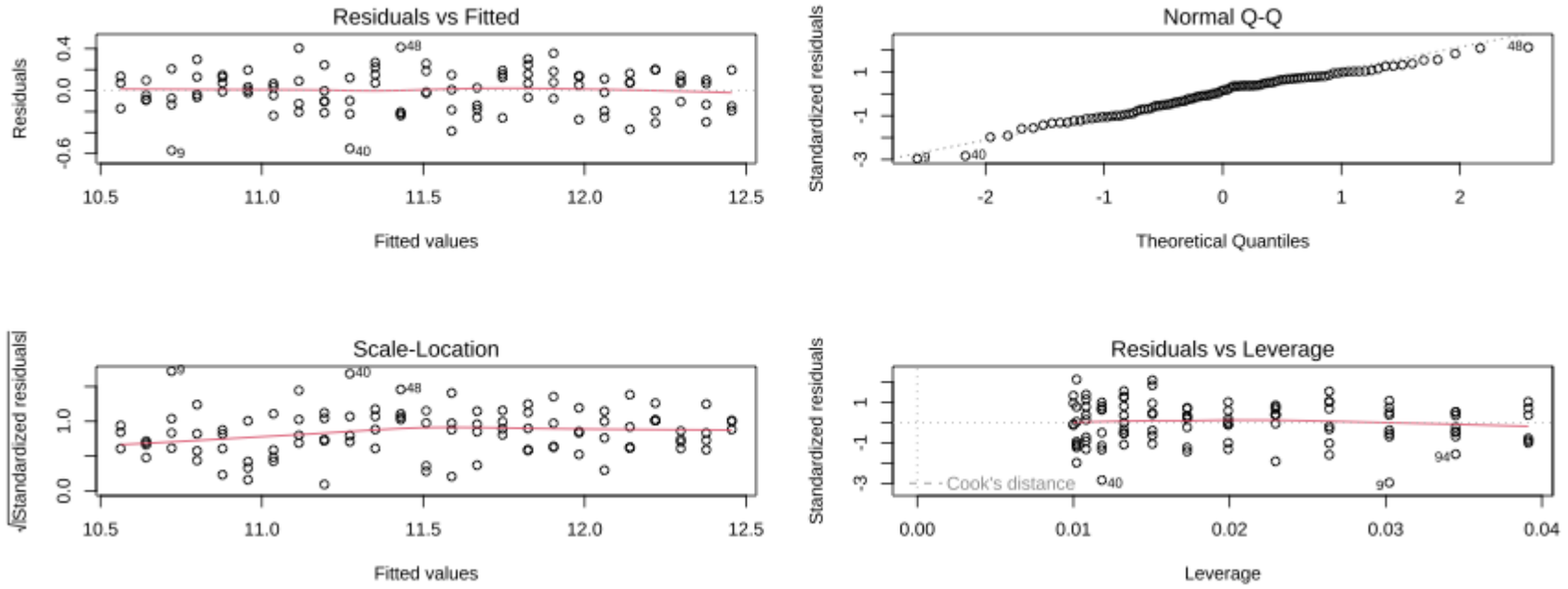
- Note, if we re-scale the model from a log scale back to the original scale of the data, we now have

$$Y_i = exp(\beta_0 + \beta_1 x_i) exp(\epsilon_i),$$

- Fitting this model in R requires only a minor modification to our formula specification.

```
initech_fit_log = lm(log(salary) ~ years, data = initech)
```

Scattertplots of the transformed data



Left: Transformed y vs. x with regression line that was fit to transformed y, Right: Untransformed y vs. x with regression line that was fit to transformed y.

How to interpret transformed regression outcome?

```
##
## Call:
## lm(formula = log(salary) ~ years, data = initech)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57022 -0.13560  0.03048  0.14157  0.41366
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 10.48381    0.04108  255.18 <0.0000000000000002 ***
## years        0.07888    0.00278   28.38 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1955 on 98 degrees of freedom
## Multiple R-squared:  0.8915,    Adjusted R-squared:  0.8904
## F-statistic: 805.2 on 1 and 98 DF,  p-value: < 0.00000000000000022
```

How to interpret transformed regression outcome?

- The transformed response is a linear combination of the predictors:

$$log(\hat{y}(x)) = \hat{\beta}_0 + \hat{\beta}_1 x = 10.848 + 0.079x,$$

- If we re-scale the data from a log scale back to the original scale of the data, we now have

$$\hat{y}(x) = exp(\hat{\beta}_0)exp(\hat{\beta}_1 x) = exp(10.848) + exp(0.079x).$$

- We see that every one additional year of experience is associated with average salary increases of $exp(0.079) = 1.082$ times. We are now multiplying, not adding.

- Prediction: If you want to make a prediction, you can plug in the value of x that you're interested in to see what the corresponding predicted value of y would be, according to your model. Suppose we'd like to know, for x=10 years, what is fitted y, salary?

$$exp(10.848) + exp(0.079x) = exp(10.848) + exp(0.079 * 10) = 51,433$$

■

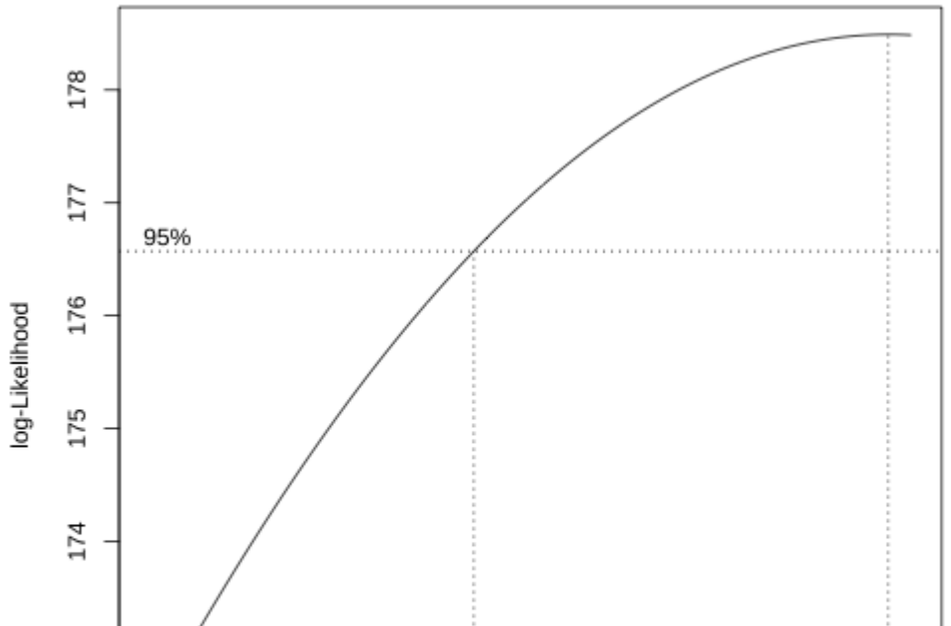
An Automated Method to Achieve Normality (Box-Cox)

- Trial and error: In practice, these methods need to be tested on the data to which they are applied to be sure that they increase rather than decrease the linearity of the relationship.
- It's hard to know which transformation is best for your data. Thankfully, there's a test that can serve as a guide: the Box-Cox method.
- The Box-Cox transformation is named after statisticians George Box and Sir David Roxbee Cox who collaborated on a 1964 paper and developed the technique.
- The Box-Cox transformation is a power transformation that corrects asymmetry of a variable, different variances or non linearity between variables.

The Box-Cox method

```
library(MASS)
initech_fit_log = lm(log(salary) ~ years, data = initech)

boxcox(initech_fit_log)
```



Box Cox transformations

Common Box-Cox Transformations	
Lambda value (λ)	Transformed data (Y')
-3	$Y^{-3} = 1/Y^3$
-2	$Y^{-2} = 1/Y^2$
-1	$Y^{-1} = 1/Y^1$
-0.5	$Y^{-0.5} = 1/(\sqrt{(Y)})$
0	log(Y)**
0.5	$Y^{0.5} = \sqrt{(Y)}$
1	$Y^1 = Y$
2	Y^2
3	Y^3

**Note: the transformation for zero is log(0), otherwise all data would transform to $Y^0 = 1$.

Transforming the right-hand-side of the equation

- You can also transform x by taking the log. Here, the coefficient is the estimated percent change in your dependent variable for a percent change in your independent variable.
- Or, you can add more variables to the right-hand-side of the equation, e.g.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i.$$

  Now the response y is a linear function of "two" variables: x and x squared.

- To interpret this we need to use multiple linear regression, which we haven't studied yet.

ANOVA

- You can think of ANOVA (analysis of variance) as a more general version of the t-test, or a special case of linear regression in which all covariates are factors (i.e. categorical variables).
- Question: Is there a significant association between race and birthweight?

```
## # A tibble: 3 × 3
##   race  mean.bwt se.bwt
##   <fct>    <dbl>  <dbl>
## 1 white     3103     74
## 2 black     2720    125
## 3 other     2805     88
```

- It looks like there's some association, but we don't yet know if it's statistically significant.
- Note that if we had just two racial categories in our data, we could run a t-test. Since we have more than 2, we need to run a 1-way analysis of variance (ANOVA).

ANOVA

- Terminology: a k-way ANOVA is used to assess whether the mean of an outcome variable is constant across all combinations of k factors. The most common examples are 1-way ANOVA (looking at a single factor), and 2-way ANOVA (looking at two factors).

```
##              Df   Sum Sq Mean Sq F value  Pr(>F)
## race          2  5015725 2507863   4.913 0.00834 **
## Residuals   186 94953931  510505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```