Linear Regression

Presented by David John Baker
January 2020

FLATIRON SCHOOL

More Linear Regression!!

- Categorical Predictors
- Assumptions
- Standardizing Variables
- Log Transforming Data
- Missing Data
- Multicollinearity

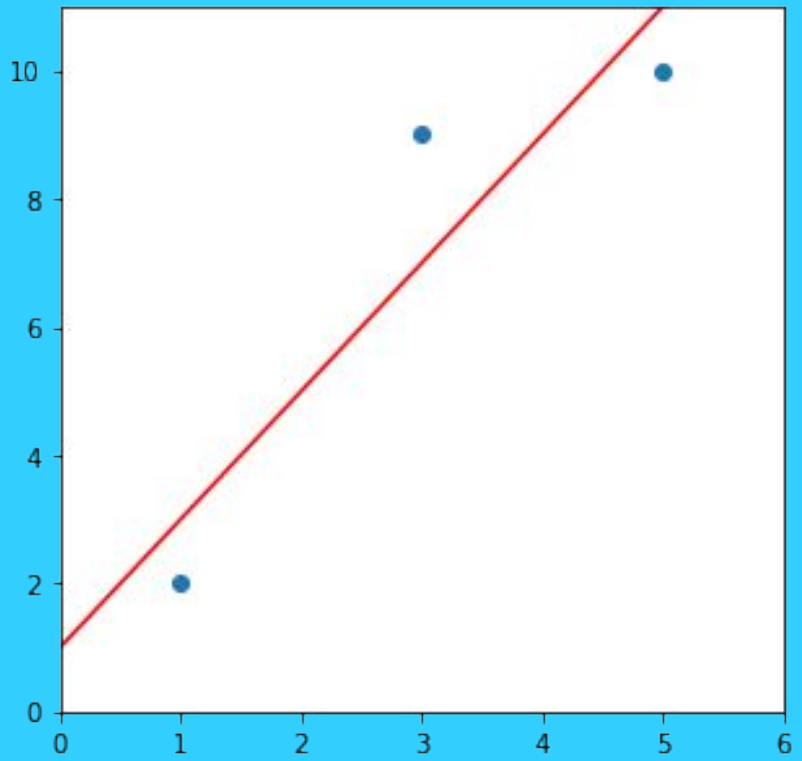
But first... some review!

- Take three minutes to talk to your partner and answer the question…
-

What is linear regression and why do we use it?

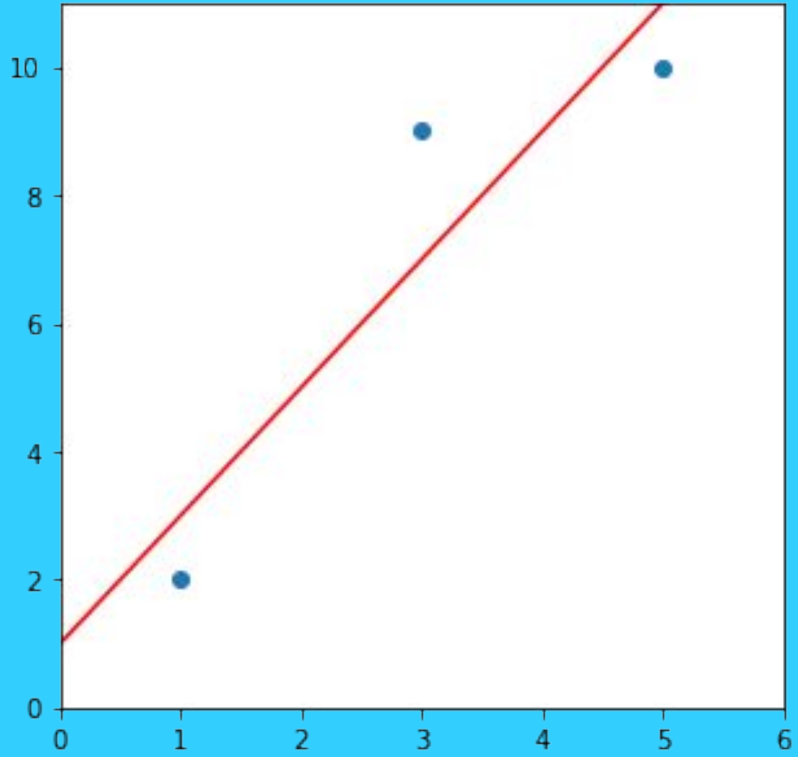
Linear regression fits a line between any number predictors and a <u>continuous dependant variable.</u> Since we are fitting a LINE we can describe it with two PARAMETERS

$$\hat{y} = \beta_1 x + \beta_0 = 2x + 1$$



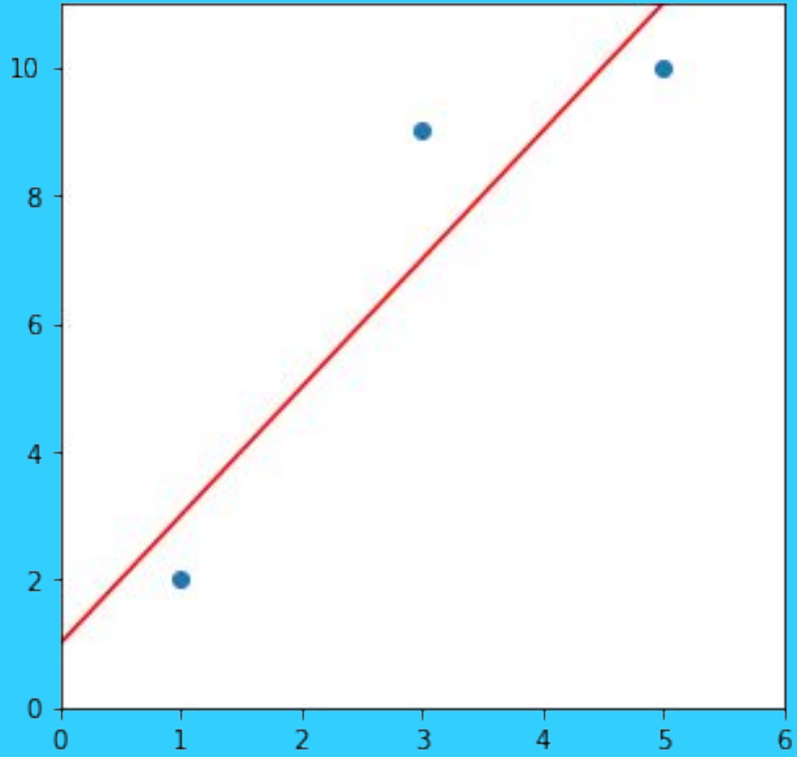
Linear regression fits a line between any number predictors and a continuous dependant variable. Since we are fitting a LINE we can describe it with two PARAMETERS

$$\hat{y} = \boxed{\beta_1 x} + \beta_0 = 2x + 1$$



Linear regression fits a line between any number predictors and a <u>continuous dependant variable.</u> Since we are fitting a LINE we can describe it with two PARAMETERS

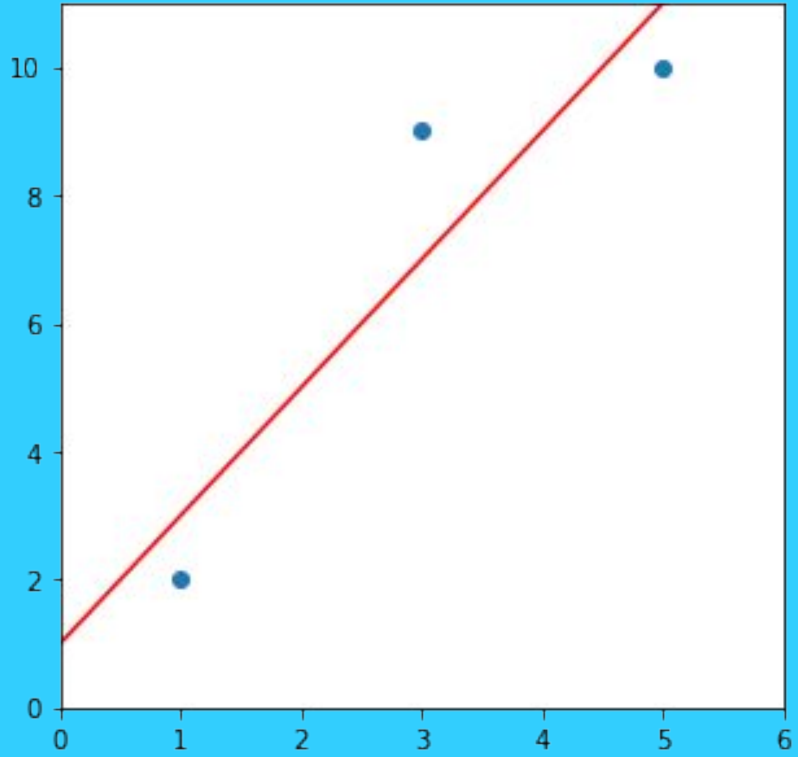
$$\hat{y} = \beta_1 x + \beta_0 = 2x + 1$$



Linear regression fits a line between any number predictors and a <u>continuous dependant variable.</u> Since we are fitting a LINE we can describe it with two PARAMETERS

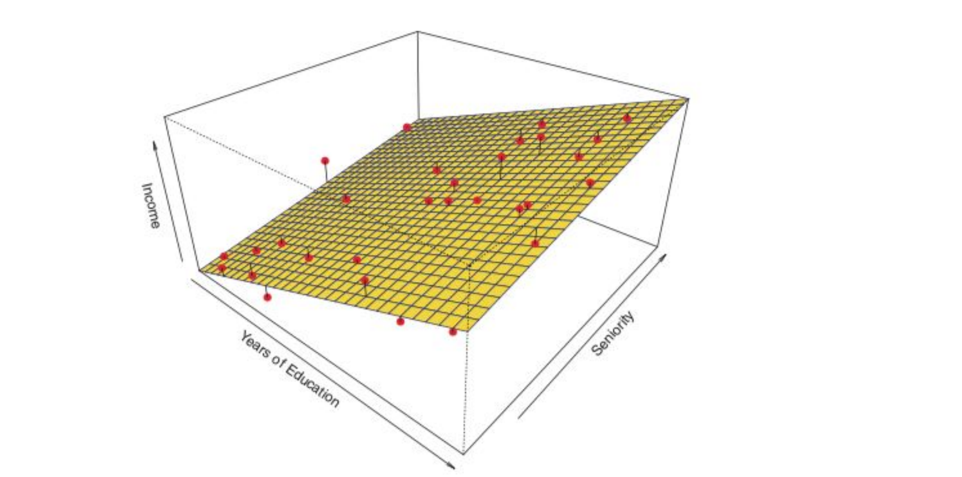
Unit change Y give 1 unit change in X, all other variables held equal!

$$\hat{y} = \beta_1 x + \beta_0 = 2x + 1$$



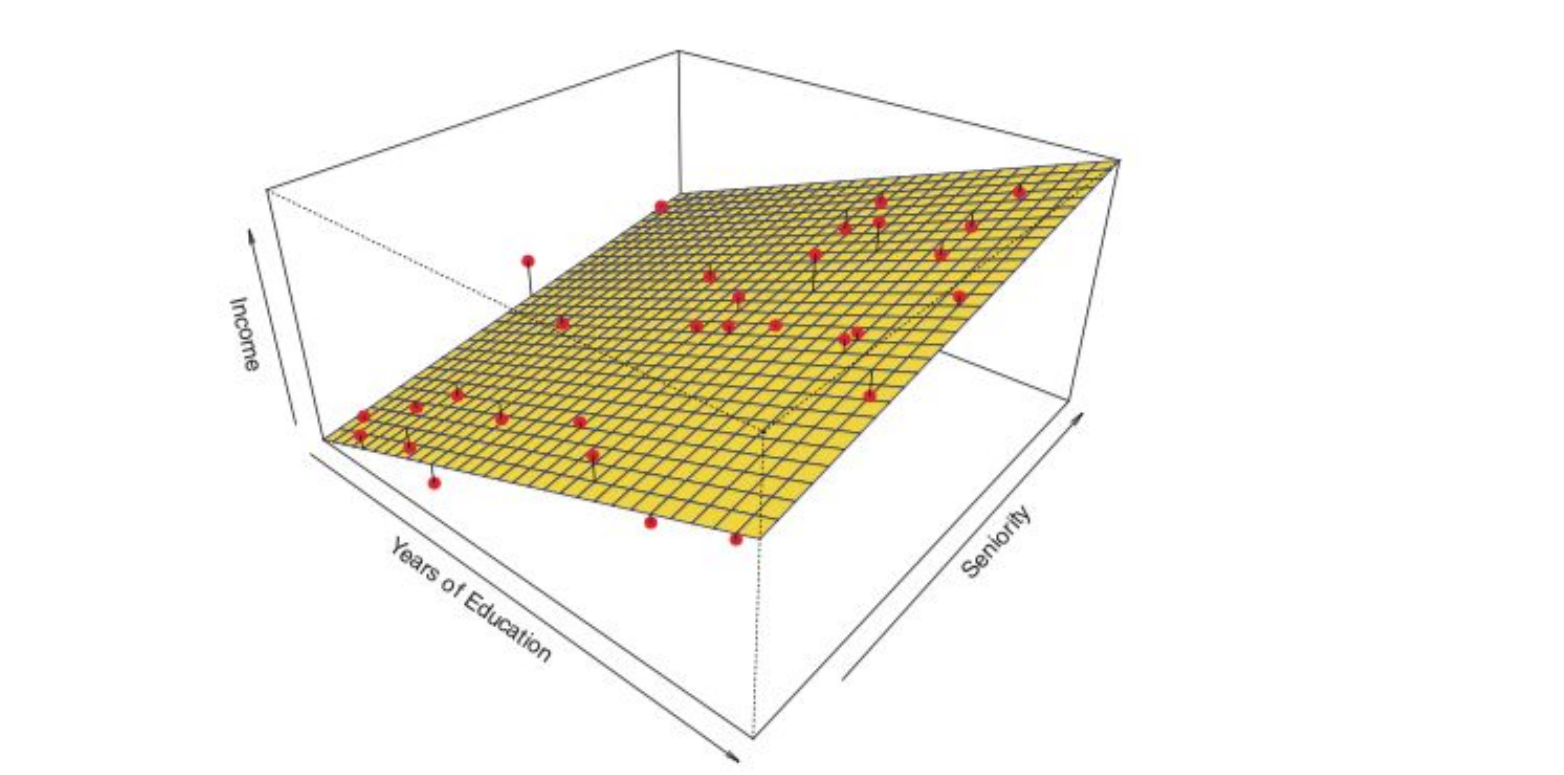
Linear regression fits a line between any number predictors and a <u>continuous dependant variable.</u> Since we are fitting a LINE we can describe it with two PARAMETERS.

Or extend it to multiple parameters (visualized here in 2D)



Unit change Y give 1 unit change in X, all other variables held equal!

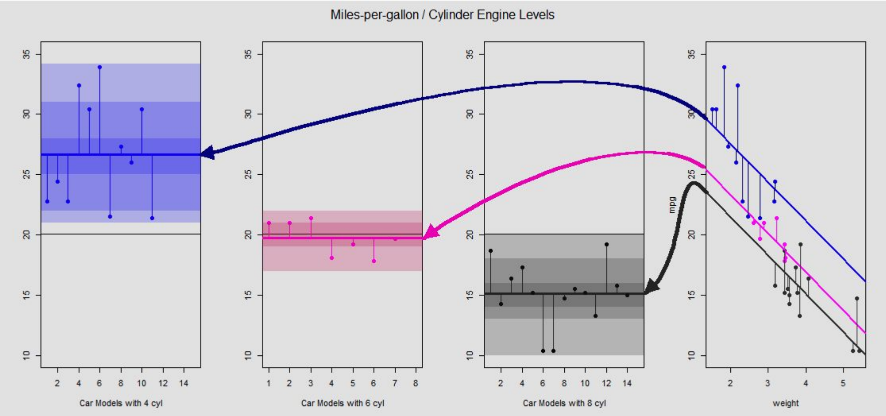
What meaningful insights could we produce with this data viz?



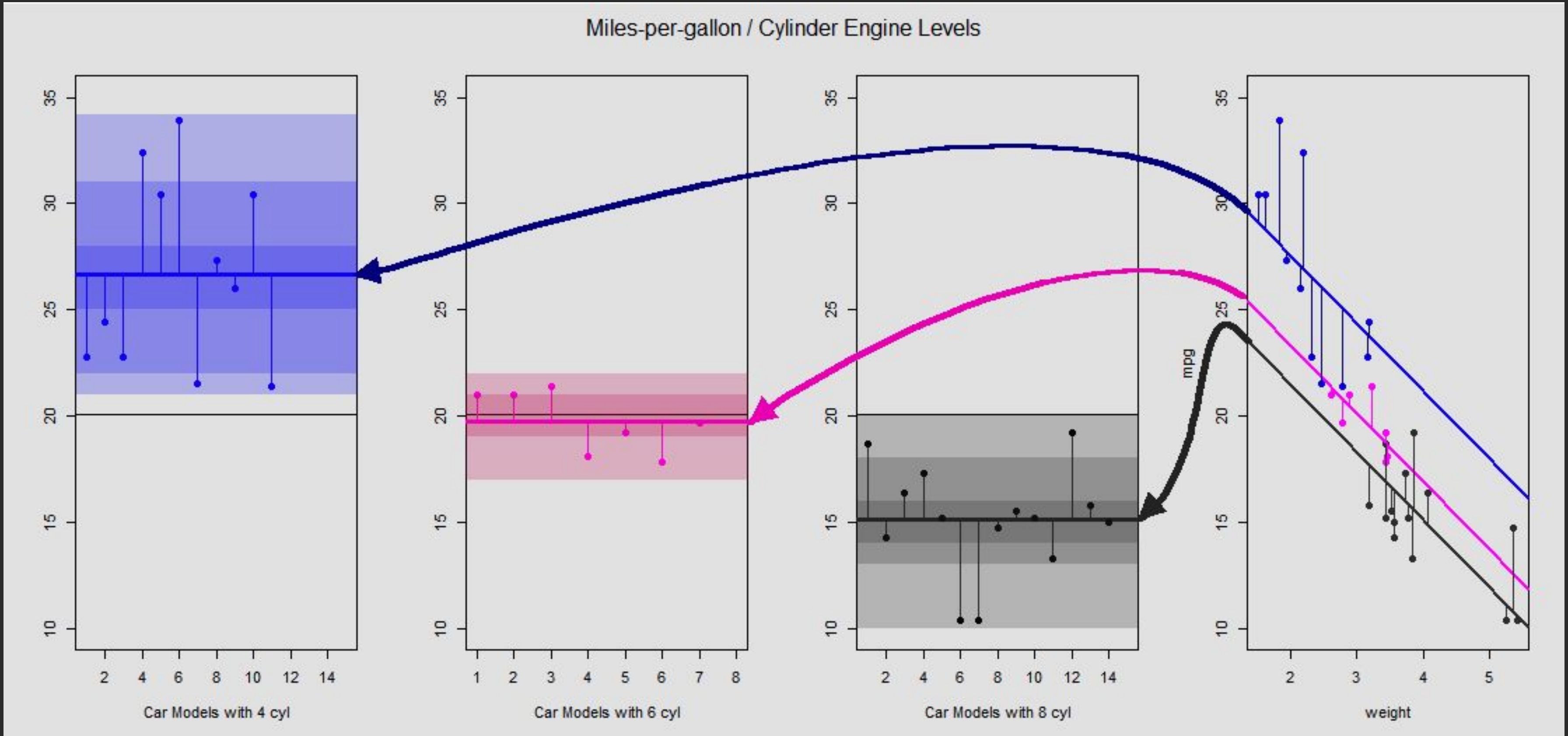
Categorical Predictors

Categorical Predictors

- Linear models are able to handle categorical predictors
- This assumes equal changes in slopes between groups
- Variables need to be “dummy coded”
- Can use both categorical and continuous variables at the same time!



https://stats.stackexchange.com/questions/175246/why-is-anova-equivalent-to-linear-regression



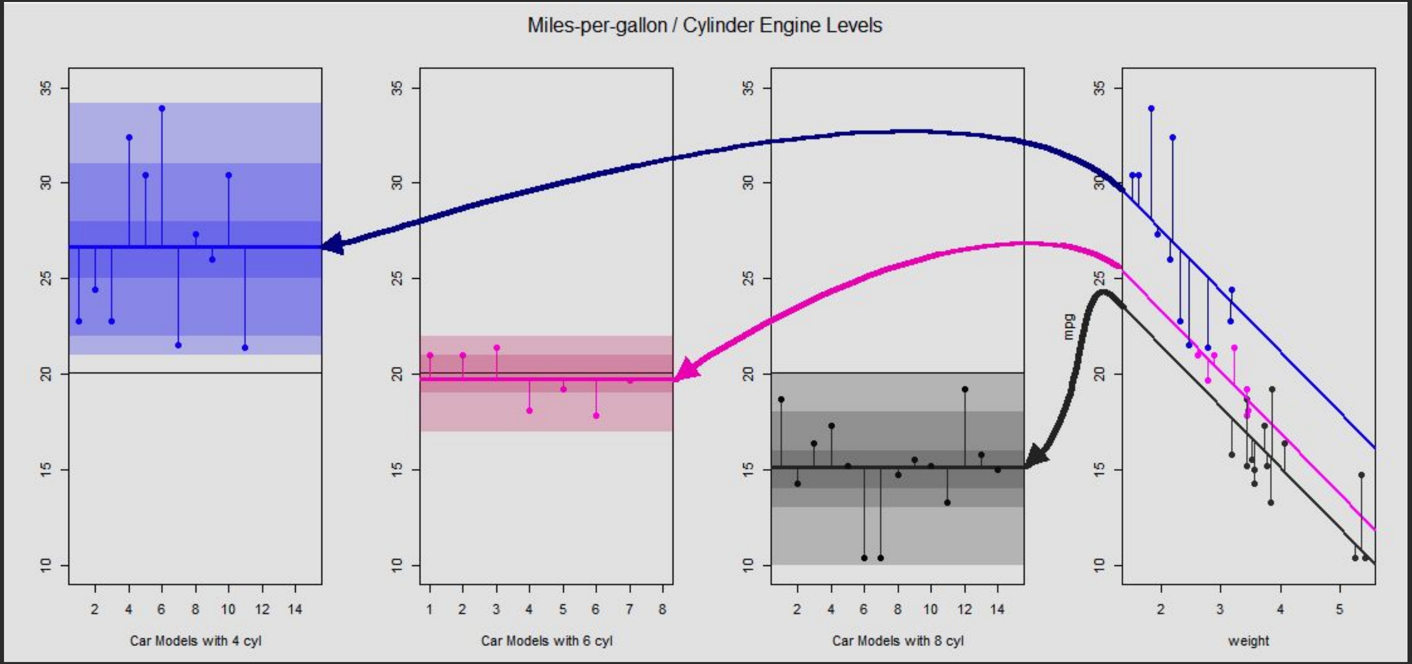
https://stats.stackexchange.com/questions/175246/why-is-anova-equivalent-to-linear-regression

Y = 33.99 + weight (-3.2056) + cycle_6 (-4.2556) + cycle_8 (-6.0709)

If 4 Cycle → Cycle 6 and 8 at 0

If 6 Cycle → Cycle 4 and 8 at 0

If 8 Cycle → Cycle 4 and 6 at 0

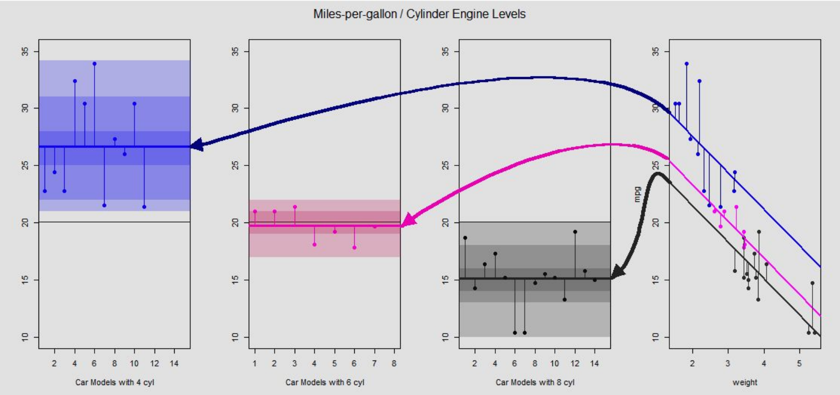


https://stats.stackexchange.com/questions/175246/why-is-anova-equivalent-to-linear-regression

Discussion Question:

How do categorical predictors work in linear regression?

What does using a categorical predictor in linear regression assume?



https://stats.stackexchange.com/questions/175246/why-is-anova-equivalent-to-linear-regression

Linear Regression Assumptions/Checks

- Errors are normally distributed!
- Zero assumptions about predictor distribution
- (Linearity...)
- Homoscedasticity
- (Missing Data)
- Additivity
- Observations Independent and Identically distributed

We’ll see this again in Mod 4...

Linear Regression Assumptions/Checks

- Errors are normally distributed!
- Zero assumptions about predictor distribution
- (Linearity…)
- Homoscedasticity
- (Missing Data)
- Additivity
- Observations Independent and Identically distributed

We'll see this again in Mod 4...

Linear Regression Checks

Linear Regression Checks

Multicollinearity

What do you think multicollinearity refers to...?

Linear Regression Checks

Multicollinearity

Variables should not be highly correlated with one another! If multiple variables measure the same “thing” then changes in one will affect changes in another, thus when one variable changes, the other will too and your model will be unstable!

Linear Regression Checks

Missing Data

- Figure out *why* our data is missing?
  - Missing completely at random?
  - Missing randomly?
  - Missing systematically?
- How to fix it…?

Linear Regression Checks

Missing Data

- Document your changes!
- Imputation Methods (mean, median, mode)
- Running Multiple Models with and without
  - it it different? Part of EDA!!
- OK for now, in future protect against Type I error!
  - Cross validation!

Linear Regression Checks

Assumptions about predictor/predicted variable?

- Linear regression DOES NOT require the independent variables to be normally distributed
- NOR do you have to log transform you dependent variables.
- You often will transform your data to make your regression model more interpretable!

Question?