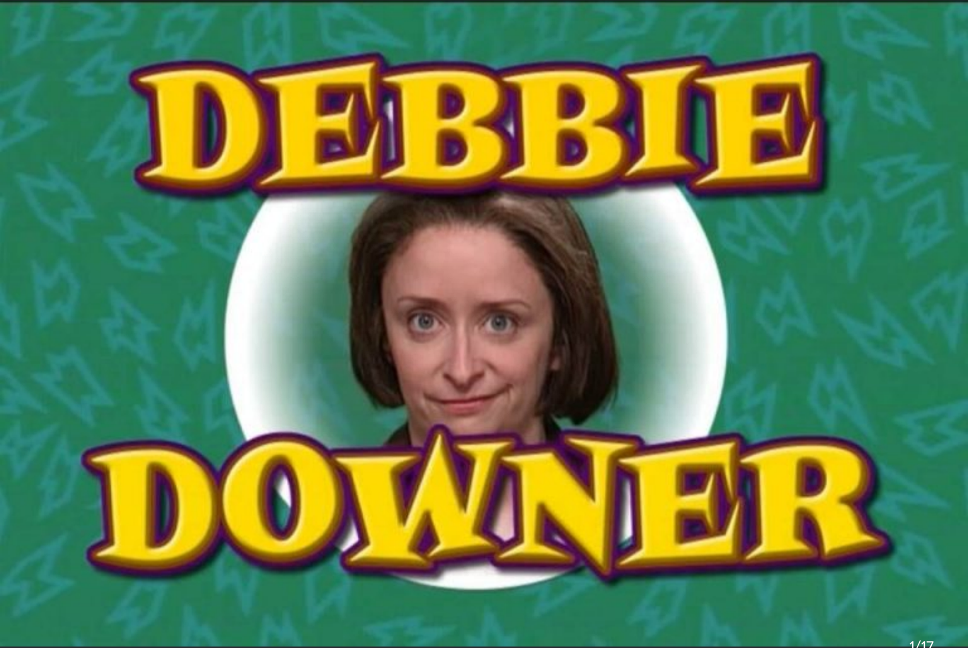
THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

What can go wrong, and how to fix it 2

Lecture 14

STA 371G

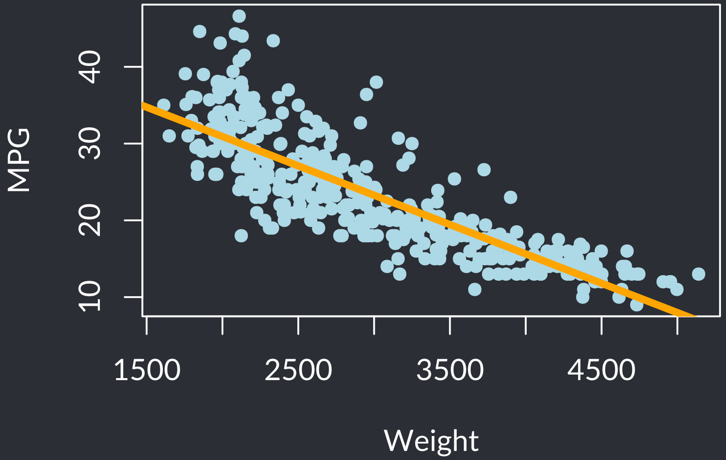

DEBBIE
DOWNER

Mileage efficiency data set

The data set `cars` contains specs for 392 different cars. We'll focus on two variables:

- **MPG** is fuel efficiency, measured in miles per gallon
- **Weight** is the weight of the car, in pounds

What problems do you see here?

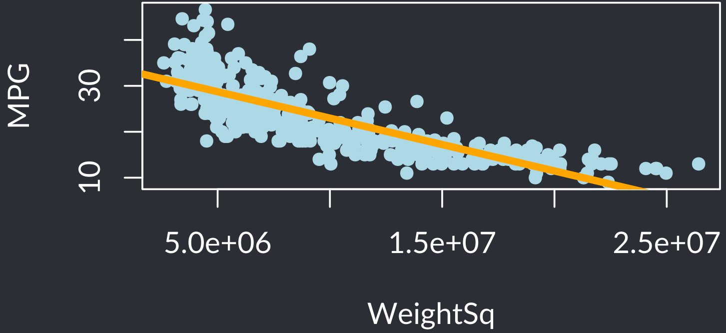


Using transformations to fix problems

- Sometimes, a violation of regression assumptions can be fixed by transforming one or the other of the variables (or both).
- When we transform a variable, we have to also transform our interpretation of the equation.

A bad example

```
cars$WeightSq <- cars$Weight^2
plot(MPG ~ WeightSq, data=cars, pch=16, col="lightblue")
sq.model <- lm(MPG ~ WeightSq, data=cars)
abline(sq.model, col="orange", lwd=4)
```



The log transformation

The log transformation is frequently useful in regression, because many nonlinear relationships are naturally exponential.


LOG!

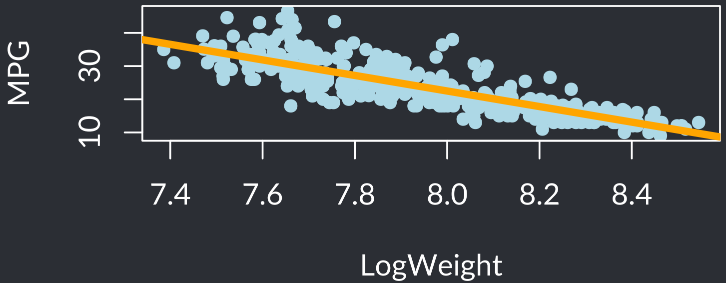
The log transformation

The log transformation is frequently useful in regression, because many nonlinear relationships are naturally exponential.

- $\log_b x = y$ when $b^y = x$
- For example, $\log_{10} 1000 = 3$, $\log_{10} 100 = 2$, and $\log_{10} 10 = 1$
- The natural log is $\log_e$, where $e \approx 2.72$ — when we say "log" we will usually mean "natural log" (although for our purposes the base doesn't matter)

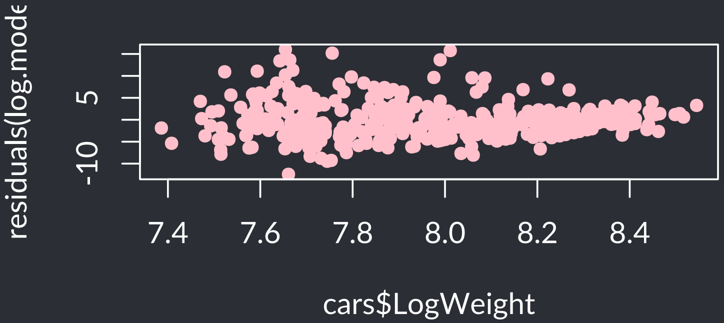
Applying a log transformation

```
cars$LogWeight <- log(cars$Weight)
plot(MPG ~ LogWeight, data=cars, pch=16, col="lightblue")
log.model <- lm(MPG ~ LogWeight, data=cars)
abline(log.model, col="orange", lwd=4)
```



Checking assumptions of our new model

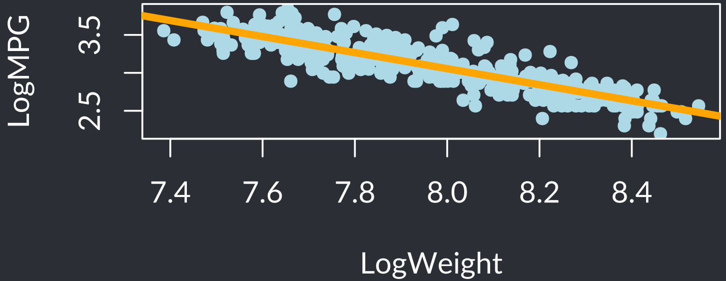
```
plot(cars$LogWeight, residuals(log.model), pch=16, col="pink")
```



Linearity looks good, but homoscedasticity is still not satisfied!

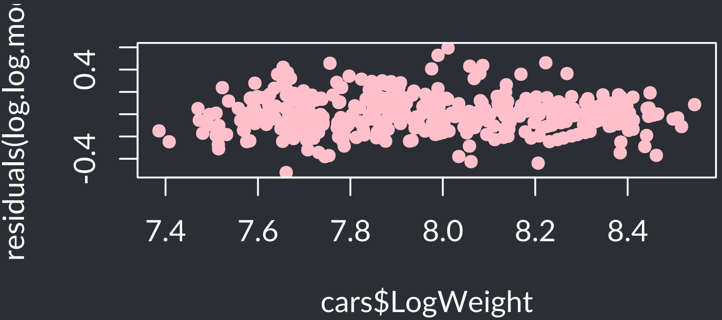
Applying a second log transformation

```
cars$LogMPG <- log(cars$MPG)
plot(LogMPG ~ LogWeight, data=cars, pch=16, col="lightblue")
log.log.model <- lm(LogMPG ~ LogWeight, data=cars)
abline(log.log.model, col="orange", lwd=4)
```



Checking assumptions of the log-log model

```
plot(cars$LogWeight, residuals(log.log.model), pch=16, col="pink"
```

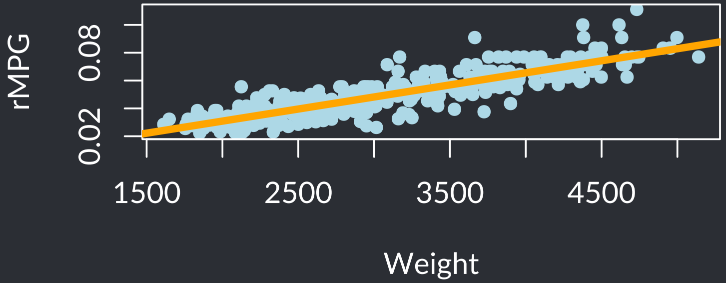


Much better—transforming MPG to log(MPG) gives us both linearity and homoscedasticity!

Another way to transform

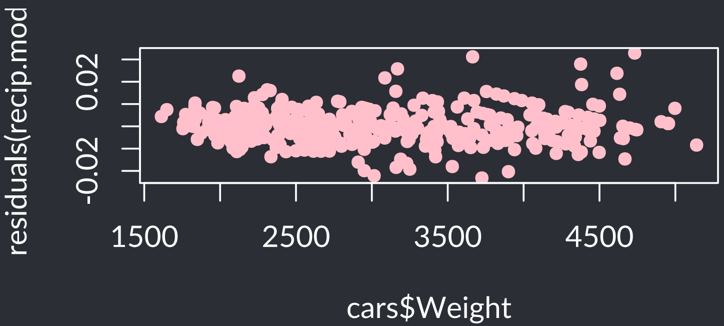
- Instead of transforming using logs, we can also transform using the reciprocal ($x \longrightarrow 1/x$)
- Let's transform Y by making MPG $\longrightarrow$ 1/MPG

```r
cars$rMPG <- 1/(cars$MPG)
plot(rMPG ~ Weight, data=cars, pch=16, col="lightblue")
recip.model <- lm(rMPG ~ Weight, data=cars)
abline(recip.model, col="orange", lwd=4)
```



Checking assumptions of the reciprocal model

```
plot(cars$Weight, residuals(recip.model), pch=16, col="pink")
```



Which transformation is better?

- We could compare R^2 values: $R^2 = 0.78$ for the reciprocal model, vs $R^2 = 0.76$ for the log-log model

Which transformation is better?

- We could compare R^2 values: $R^2 = 0.78$ for the reciprocal model, vs $R^2 = 0.76$ for the log-log model
- There is also a conceptual reason to prefer the reciprocal model: MPG is already a ratio, which tips us off that a reciprocal transformation would be appropriate

Which transformation is better?

- We could compare R^2 values: $R^2 = 0.78$ for the reciprocal model, vs $R^2 = 0.76$ for the log-log model
- There is also a conceptual reason to prefer the reciprocal model: MPG is already a ratio, which tips us off that a reciprocal transformation would be appropriate
- The reciprocal model is also simpler to interpret since only one of the variables is transformed

Which transformation is better?

- We could compare R^2 values: $R^2 = 0.78$ for the reciprocal model, vs $R^2 = 0.76$ for the log-log model
- There is also a conceptual reason to prefer the reciprocal model: MPG is already a ratio, which tips us off that a reciprocal transformation would be appropriate
- The reciprocal model is also simpler to interpret since only one of the variables is transformed
- In general, log transformations tend to work best when the spacing between values increases as the values increase (e.g., salaries, city population sizes)

Thinking about transformations

- Thinking about whether you want to stretch or squeeze one of the axes, and apply a transformation accordingly (e.g., $\sqrt{x}$ or $\log x$ to squeeze; x^2 or e^x to stretch).

Thinking about transformations

- Thinking about whether you want to stretch or squeeze one of the axes, and apply a transformation accordingly (e.g., $\sqrt{x}$ or $\log x$ to squeeze; x^2 or e^x to stretch).
- Transformations of Y can address both heteroscedasticity and nonlinearity; transformations of X can only address nonlinearity.

Thinking about transformations

- Thinking about whether you want to stretch or squeeze one of the axes, and apply a transformation accordingly (e.g., $\sqrt{x}$ or $\log x$ to squeeze; x^2 or e^x to stretch).
- Transformations of Y can address both heteroscedasticity and nonlinearity; transformations of X can only address nonlinearity.
- You might need to transform both X and Y; if so, start by transforming Y to address the heteroscedasticity, and then transform X to address nonlinearity if necessary.

Thinking about transformations

- Thinking about whether you want to stretch or squeeze one of the axes, and apply a transformation accordingly (e.g., $\sqrt{x}$ or $\log x$ to squeeze; x^2 or e^x to stretch).
- Transformations of Y can address both heteroscedasticity and nonlinearity; transformations of X can only address nonlinearity.
- You might need to transform both X and Y; if so, start by transforming Y to address the heteroscedasticity, and then transform X to address nonlinearity if necessary.
- It's OK to do a little trial and error!