Introduction to Linear Models (Regression)

Hunter Glanz

OUTLINE

Motivation

Univariate

Bivariate

Multivariate

Data Exploration to Data Analysis

- What are the observations?
- What variables do we have?

Data Exploration to Data Analysis

- What are the observations?
- What variables do we have?
- What are the values of these variables like?

Data Exploration to Data Analysis

- What are the observations?
- What variables do we have?
- What are the values of these variables like?
- What kinds of relationships are there among the variables we have?

Data Exploration to Data Analysis

- What are the observations?
- What variables do we have?
- What are the values of these variables like?
- What kinds of relationships are there among the variables we have?

Storytelling with data!

The Carseats Dataset in the ISLR package for R

- 400 observations on the following variables:
  - Sales (in thousands) at each location
  - CompPrice
  - Income
  - Advertising
  - Population
  - Price
  - ShelveLoc
  - Age
  - Education
  - Urban
  - US
- More info here: https://rdrr.io/cran/ISLR/man/Carseats.html

Research Questions

- What kinds of questions might you ask of this dataset?
- What kinds of questions might have caused you to collect/obtain these data?

Research Questions

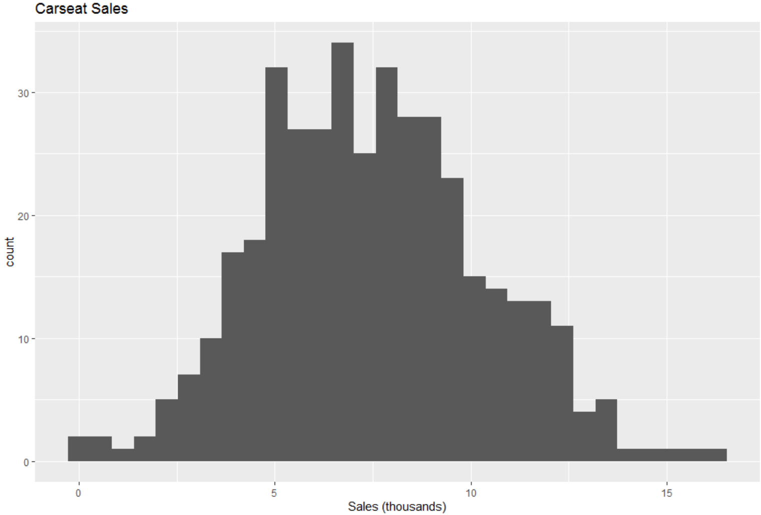
- What kinds of questions might you ask of this dataset?
- What kinds of questions might have caused you to collect/obtain these data?
- Primary question:

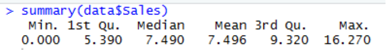
Research Questions

- What kinds of questions might you ask of this dataset?
- What kinds of questions might have caused you to collect/obtain these data?
- Primary question:

Can we predict Sales using the other information in this dataset?

What do we know about Sales?

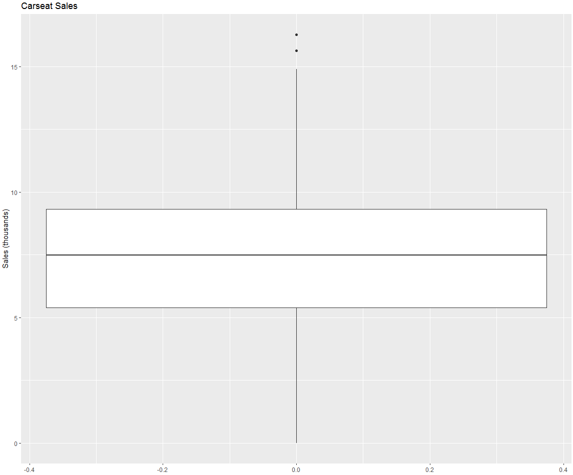




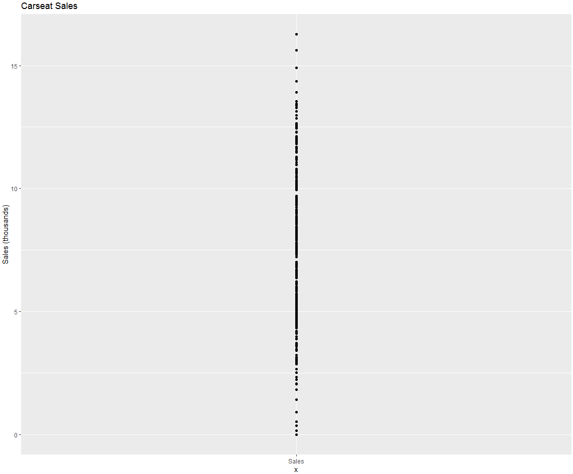
```
> summary(data$Sales)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   5.390   7.490   7.496   9.320  16.270 
```

Other Possible Visualizations...?

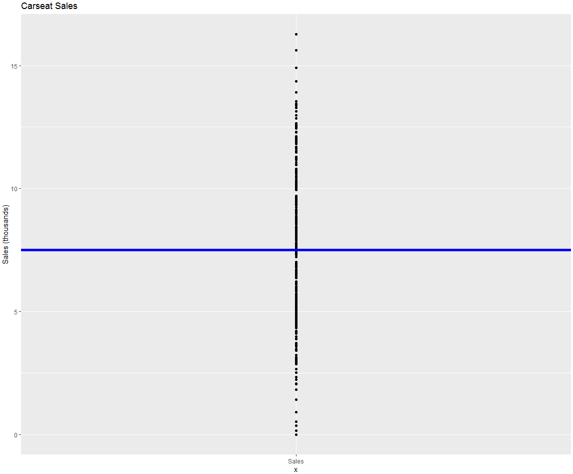
Other Possible Visualizations...?



Another Visualization?



Another Visualization? With the Mean...



Predicting Sales Part I

- Without knowing any other information or using any other data, what would your prediction for Sales be?

Predicting Sales Part I

- Without knowing any other information or using any other data, what would your prediction for Sales be?
  - The most representative value of Sales that we have access to, right?!

Predicting Sales Part I

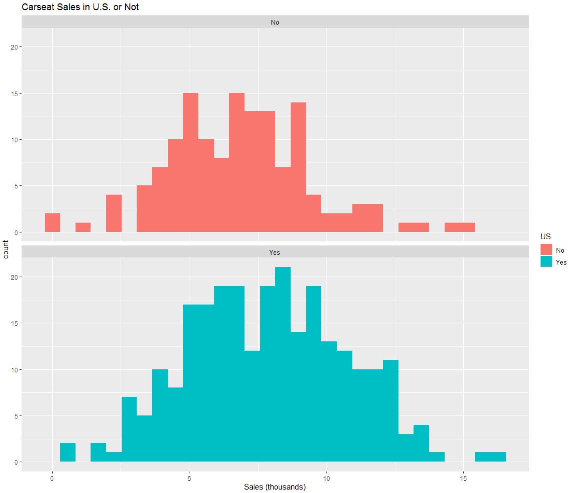
- Without knowing any other information or using any other data, what would your prediction for Sales be?
    - The most representative value of Sales that we have access to, right?!
- The mean or average of Sales is a good start: 7.5 thousand

But we DO have more data!

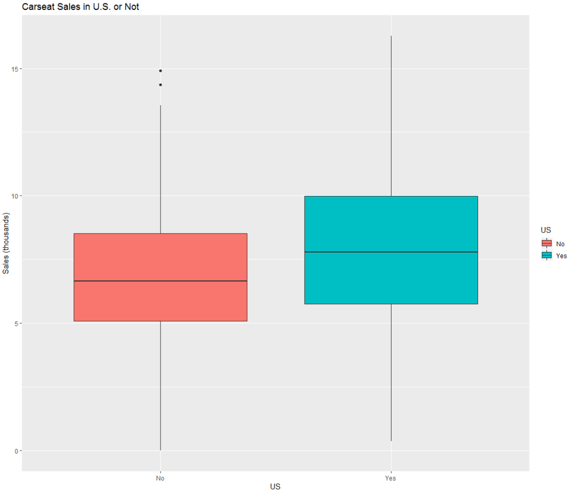
- Does knowing whether a store is in the U.S. or not help in predicting Sales?

But we DO have more data!

- Does knowing whether a store is in the U.S. or not help in predicting Sales?



Does being in the U.S. change our Sales prediction?



- Any better?

Predicting Sales Part II

- If you knew a store was in the U.S., what would your prediction for Sales be?

Predicting Sales Part II

- If you knew a store was in the U.S., what would your prediction for Sales be?
  - The most representative value of Sales for stores in the U.S. that we have access to, right?!

Predicting Sales Part II

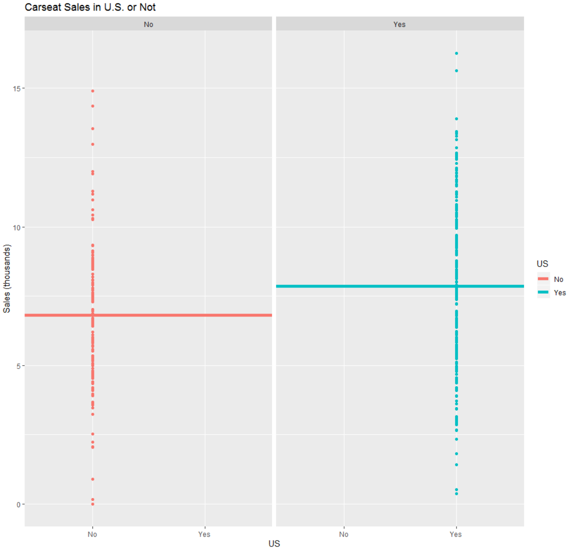
- If you knew a store was in the U.S., what would your prediction for Sales be?
  - The most representative value of Sales for stores in the U.S. that we have access to, right?!
- The mean or average of Sales in the U.S. is a good start:

Predicting Sales Part II

- If you knew a store was in the U.S., what would your prediction for Sales be?
  - The most representative value of Sales for stores in the U.S. that we have access to, right?!
- The mean or average of Sales in the U.S. is a good start:
  - Compute the average Sales for stores in the U.S.
  - Compute the average Sales for stores not in the U.S.

So what did we just do?!

So what did we just do?!



What are we doing **statistically**?

- Does knowing whether a store is in the U.S. change our prediction of Sales? **OR**

What are we doing **statistically**?

- Does knowing whether a store is in the U.S. change our prediction of Sales? **OR**
- Is there a relationship between a store's location and its Sales? **OR**

What are we doing **statistically**?

- Does knowing whether a store is in the U.S. change our prediction of Sales? **OR**
- Is there a relationship between a store's location and its Sales? **OR**
- Is there a difference in the average Sales between stores in the U.S. and stores outside the U.S.?

What are we doing **statistically**?

- Does knowing whether a store is in the U.S. change our prediction of Sales? **OR**
- Is there a relationship between a store's location and its Sales? **OR**
- Is there a difference in the average Sales between stores in the U.S. and stores outside the U.S.?
  - Hopefully this last question sounds familiar...

What are we doing **statistically**?

- Does knowing whether a store is in the U.S. change our prediction of Sales? **OR**
- Is there a relationship between a store's location and its Sales? **OR**
- Is there a difference in the average Sales between stores in the U.S. and stores outside the U.S.?
  - Hopefully this last question sounds familiar...

Two-sample ...

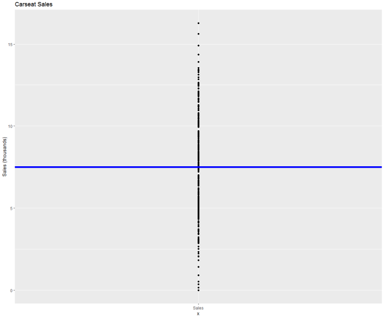
What are we doing **statistically**?

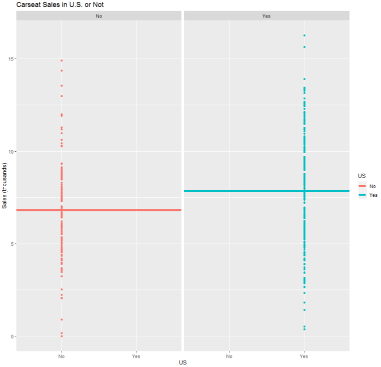
- Does knowing whether a store is in the U.S. change our prediction of Sales? **OR**
- Is there a relationship between a store's location and its Sales? **OR**
- Is there a difference in the average Sales between stores in the U.S. and stores outside the U.S.?
  - Hopefully this last question sounds familiar...

Two-sample ...

- t-test
- confidence interval

But What About the Lines on Those Graphs?!





- What's the equation of a horizontal line?

Our Models Thus Far

- Sales alone:

$$Sales = \beta_0 + \epsilon$$

Our Models Thus Far

- Sales alone:

$$Sales = \beta_0 + \epsilon$$

- Sales on US:

$$Sales = \beta_0 + \beta_1 USYes + \epsilon$$

- where we assume $\epsilon \sim N(0, \sigma^2)$.

Fitting Our Models Using Data

- Sales alone:

$$E[Sales] = \beta_0$$

$$S\hat{a}les = \hat{\beta}_0 = 7.5$$

Fitting Our Models Using Data

- Sales alone:

$$E[Sales] = \beta_0$$

$$S\hat{al}es = \hat{\beta}_0 = 7.5$$

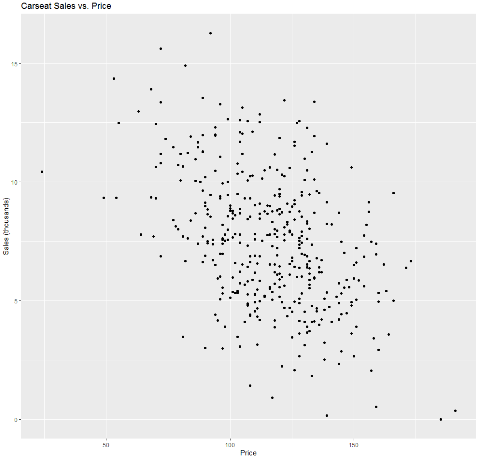
- Sales on US:

$$E[Sales|US] = \beta_0 + \beta_1 USYes$$

$$S\hat{al}es = \hat{\beta}_0 + \hat{\beta}_1 USYes = 6.823 + 1.0439 USYes$$

- We estimate the **average** Sales using the fitted model!
- $\hat{\beta}_1$: we **expect** a 1.0439 thousand unit increase in Sales if a store is in the U.S.

What is the relationship between Sales and Price?



- How do we usually describe/interpret such plots?

Correlation

- What does *correlation* measure?

Correlation

- What does *correlation* measure?
  - The **strength** and **direction** of the **linear** relationship between **two quantitative** variables.
- What are the possible values correlation, r, can take?

Correlation

- What does *correlation* measure?
  - The **strength** and **direction** of the **linear** relationship between **two quantitative** variables.
- What are the possible values correlation, r, can take?
  - Between -1 and 1

Correlation

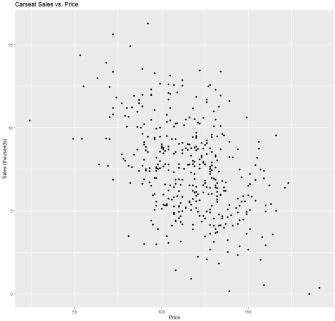
- What does *correlation* measure?
  - The **strength** and **direction** of the **linear** relationship between **two quantitative** variables.
- What are the possible values correlation, r, can take?
  - Between -1 and 1
- What else do usually hear about **correlation**?!

Correlation

- What does *correlation* measure?
  - The **strength** and **direction** of the **linear** relationship between **two quantitative** variables.
- What are the possible values correlation, r, can take?
  - Between -1 and 1
- What else do usually hear about **correlation**?!
  - *correlation does not imply causation*

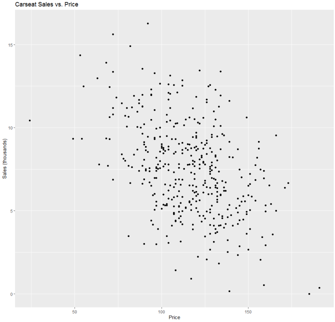
Can We Go Beyond Correlation?

- What is your estimate of the correlation between Sales and Price?



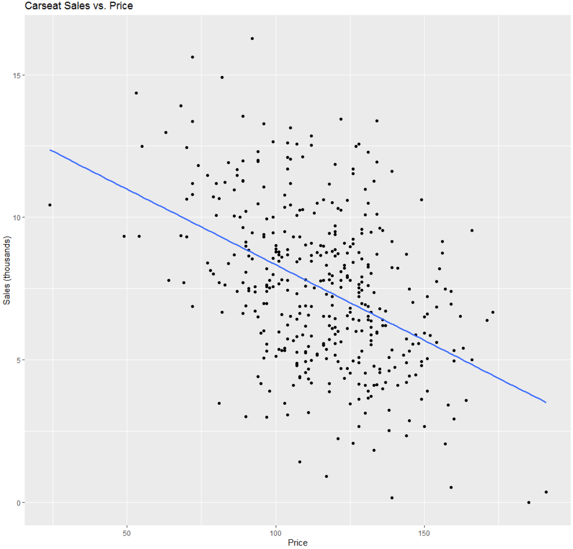
Can We Go Beyond Correlation?

- What is your estimate of the correlation between Sales and Price?



- $r = -0.445$
- How else could we describe the relationship between these two variables?

(Least Squares) Best Fit Line



The Model Equation of the Best Fit Line

- Sales on Price:

$$Sales = \beta_0 + \beta_1 Price + \epsilon$$

The Model Equation of the Best Fit Line

- Sales on Price:

$$Sales = \beta_0 + \beta_1 Price + \epsilon$$

$$E[Sales|Price] = \beta_0 + \beta_1 Price$$

The Model Equation of the Best Fit Line

- Sales on Price:

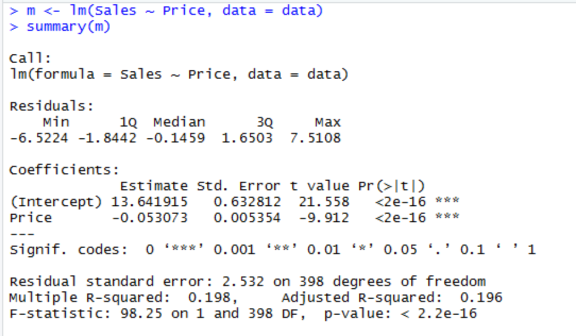
$$Sales = \beta_0 + \beta_1 Price + \epsilon$$

$$E[Sales|Price] = \beta_0 + \beta_1 Price$$

$$\hat{Sales} = \hat{\beta}_0 + \hat{\beta}_1 Price = 13.641915 - 0.053073 Price$$

- We estimate the **average** Sales using the fitted model!
- $\hat{\beta}_1$: we **expect** a 0.053073 thousand (53.073) unit decrease in Sales for every dollar increase in Price. (not causation!)

Fitting Linear Models in R



```
> m <- lm(Sales ~ Price, data = data)
> summary(m)

Call:
lm(formula = Sales ~ Price, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5224 -1.8442 -0.1459  1.6503  7.5108 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.641915   0.632812  21.558   <2e-16 ***
Price       -0.053073   0.005354  -9.912   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.532 on 398 degrees of freedom
Multiple R-squared:  0.198,	Adjusted R-squared:  0.196 
F-statistic: 98.25 on 1 and 398 DF,  p-value: < 2.2e-16
```

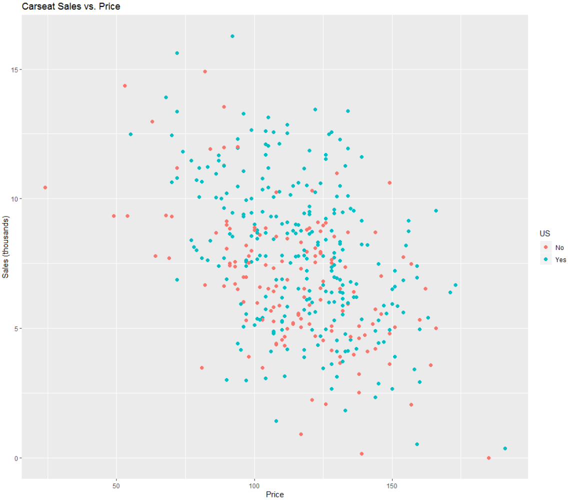
- Check out the **estimate** column for the coefficient estimates!

Our Dataset is Rich...Let’s Use It!

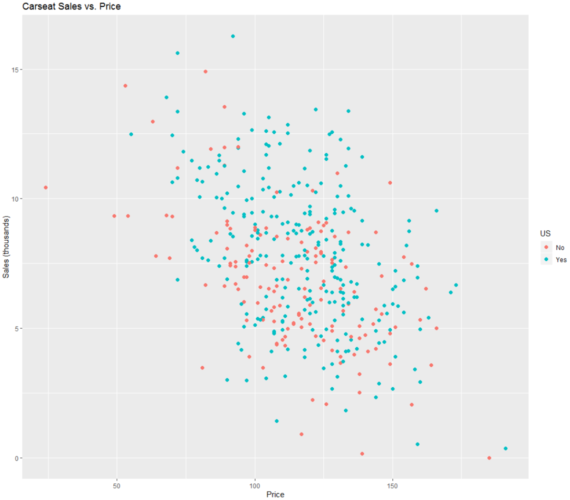
- Could we use both Price and US to help predict Sales?

Our Dataset is Rich...Let's Use It!

- Could we use both Price and US to help predict Sales?



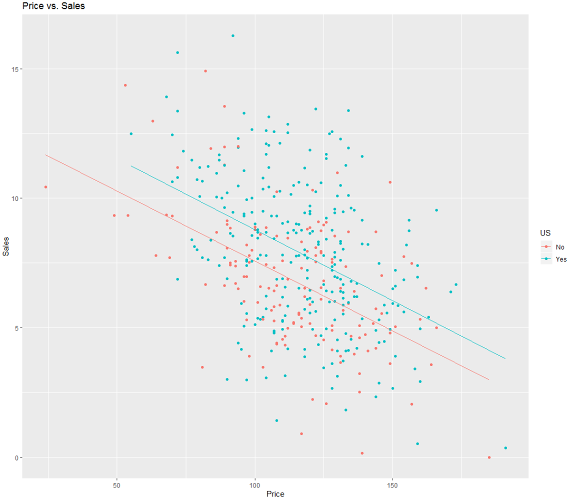
What Do We Do With Three Variables?



- What are our model options?

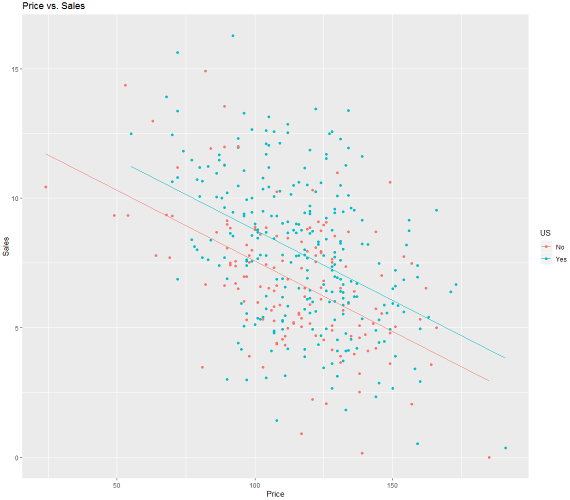
YOLO Lines!

- We could allow for completely different fitted lines for each of the two groups:



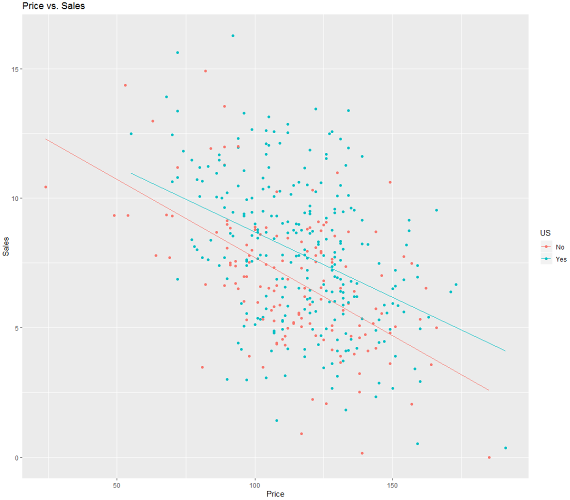
Same Slope for Both Groups

- We could force the line for each of the two groups to have the same slope:

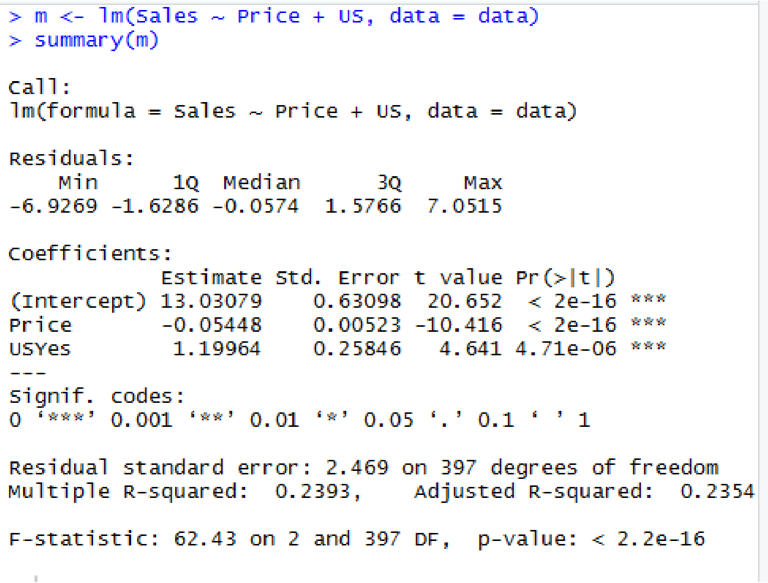


Same Intercept for Both Groups

- We could force the line for each of the two groups to have the same intercept:



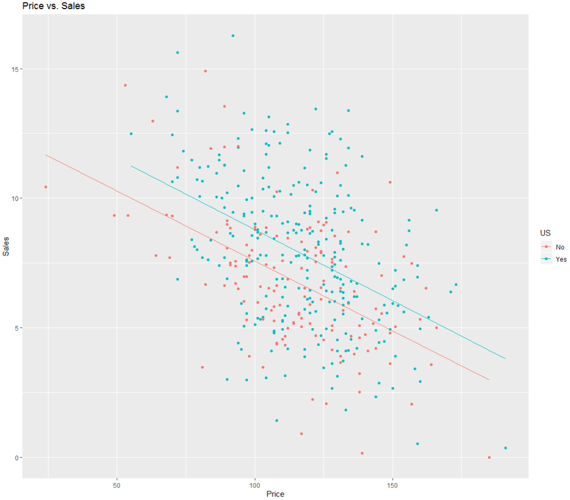
Fitting Bigger Models in R



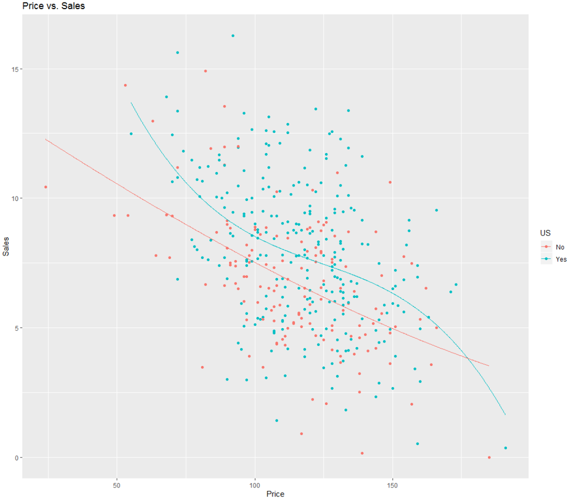
```
> m <- lm(Sales ~ Price + US, data = data)
> summary(m)

Call:
lm(formula = Sales ~ Price + US, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9269 -1.6286 -0.0574  1.5766  7.0515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
Price       -0.05448    0.00523 -10.416  < 2e-16 ***
USYes        1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2354 

F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

How do the interpretations change?

- We could allow for completely different fitted lines for each of the two groups:



We can get crazy!


Price vs. Sales
Sales
Price
US
No
Yes