STA130H1S - Week #10

Another variable can affect the nature of a relationship

Prof. A. Gibbs

March 19, 2018

STA130H1S - Week #10

Another variable can affect the nature of a relationship

Prof. A. Gibbs

March 19, 2018

Today

Big idea:

Examining the affect of another variable on a relationship

Important concepts:

1. Inference for regression parameters
2. Regression when the independent variable is a categorical variable
3. Is the regression line the same for two groups?
4. An example of a variable affecting a relationship in a non-regression setting
5. Confounding

Recommended reading:

Section 7.6 of *Modern Data Science with R*
Section 1.4.1 of [*Introductory Statistics with Randomization and Simulation* from OpenIntro]

Inference for regression parameters

What affects course evaluations?

… other than the quality of the course …

- Data from course evaluations for a random sample of courses at the University of Texas at Austin.
- Each observation corresponds to a course.
- `score` is the average student evaluation for the course.
- `bty_avg` is the average beauty rating of the professor, based on ratings of physical appear from 6 students in the course.

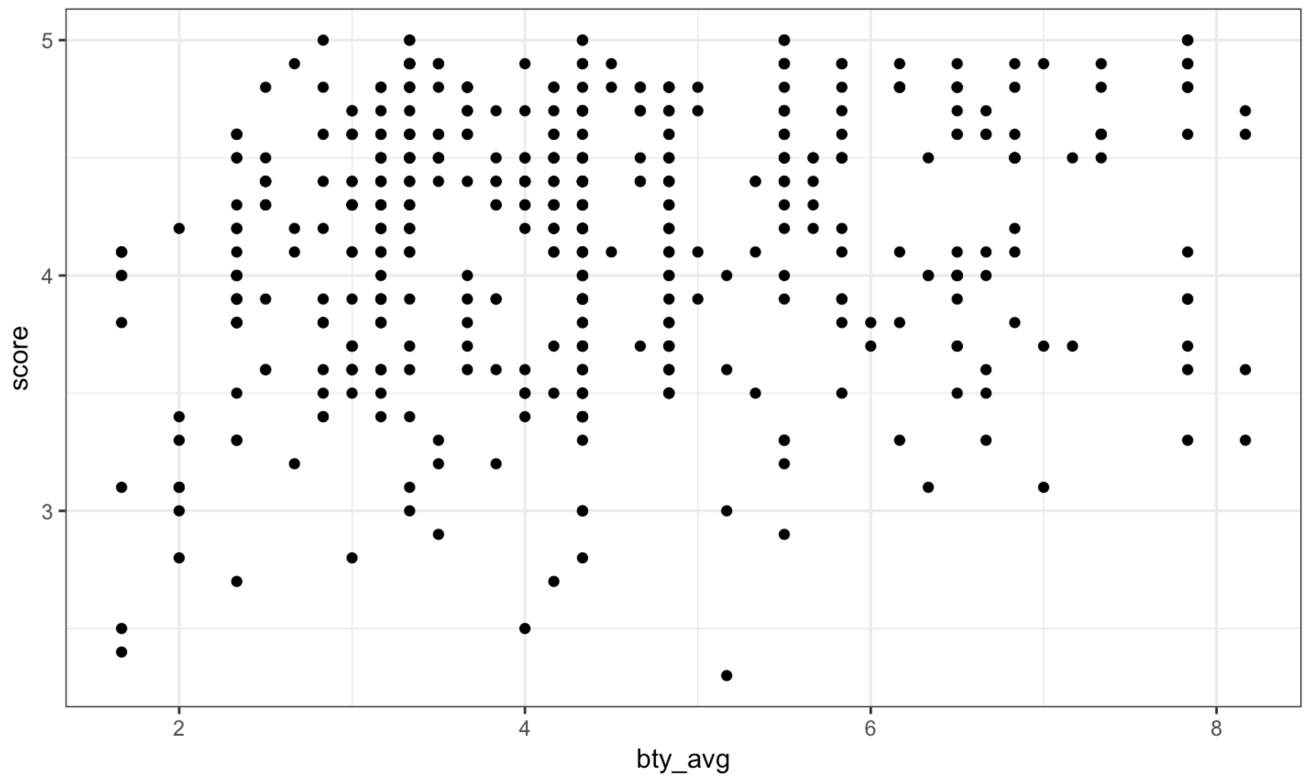
```
download.file("http://www.openintro.org/stat/data/evals.RData", destfile = "evals.RData")
load("evals.RData")
```

```
glimpse(evals)
```

```
## Observations: 463
## Variables: 21
## $ score         <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5...
## $ rank          <fctr> tenure track, tenure track, tenure track, tenur...
## $ ethnicity     <fctr> minority, minority, minority, minority, not min...
## $ gender        <fctr> female, female, female, female, male, male, mal...
## $ language      <fctr> english, english, english, english, english, en...
## $ age           <int> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, ...
## $ cls_perc_eval <dbl> 55.81395, 68.80000, 60.80000, 62.60163, 85.00000...
## $ cls_did_eval  <int> 24, 86, 76, 77, 17, 35, 39, 55, 111, 40, 24, 24,...
## $ cls_students  <int> 43, 125, 125, 123, 20, 40, 44, 55, 195, 46, 27, ...
## $ cls_level     <fctr> upper, upper, upper, upper, upper, upper, upper...
## $ cls_profs     <fctr> single, single, single, single, multiple, multi...
## $ cls_credits   <fctr> multi credit, multi credit, multi credit, multi...
## $ bty_f1lower   <int> 5, 5, 5, 5, 4, 4, 4, 5, 5, 2, 2, 2, 2, 2, 2, 2, ...
## $ bty_f1upper   <int> 7, 7, 7, 7, 4, 4, 4, 2, 2, 5, 5, 5, 5, 5, 5, 5, ...
## $ bty_f2upper   <int> 6, 6, 6, 6, 2, 2, 2, 5, 5, 4, 4, 4, 4, 4, 4, 4, ...
## $ bty_m1lower   <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, ...
## $ bty_m1upper   <int> 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...
## $ bty_m2upper   <int> 6, 6, 6, 6, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, ...
## $ bty_avg       <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000,...
## $ pic_outfit    <fctr> not formal, not formal, not formal, not formal,...
```

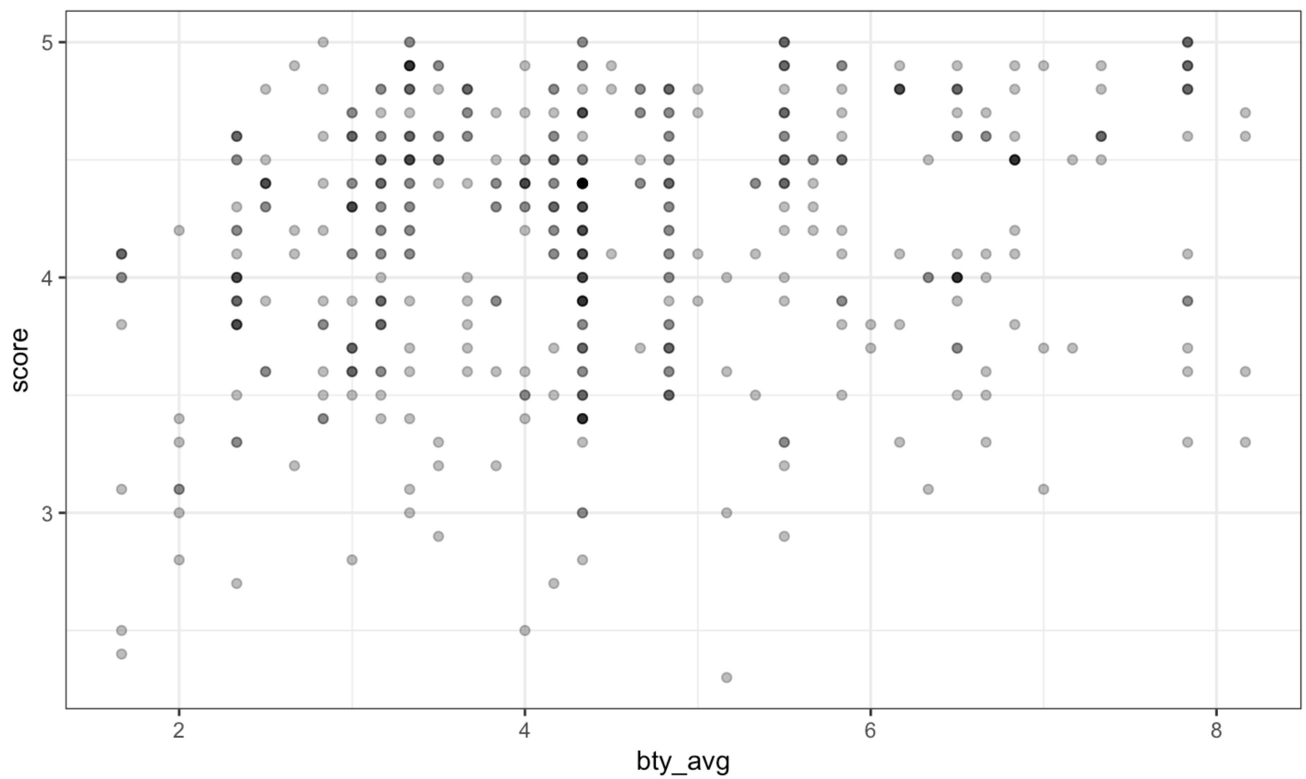
Relationship between `score` and `bty_avg`?

```
ggplot(evals, aes(x=bty_avg, y=score)) + geom_point() + theme_bw()
```



Use some transparency so we can see where there are overlapping points

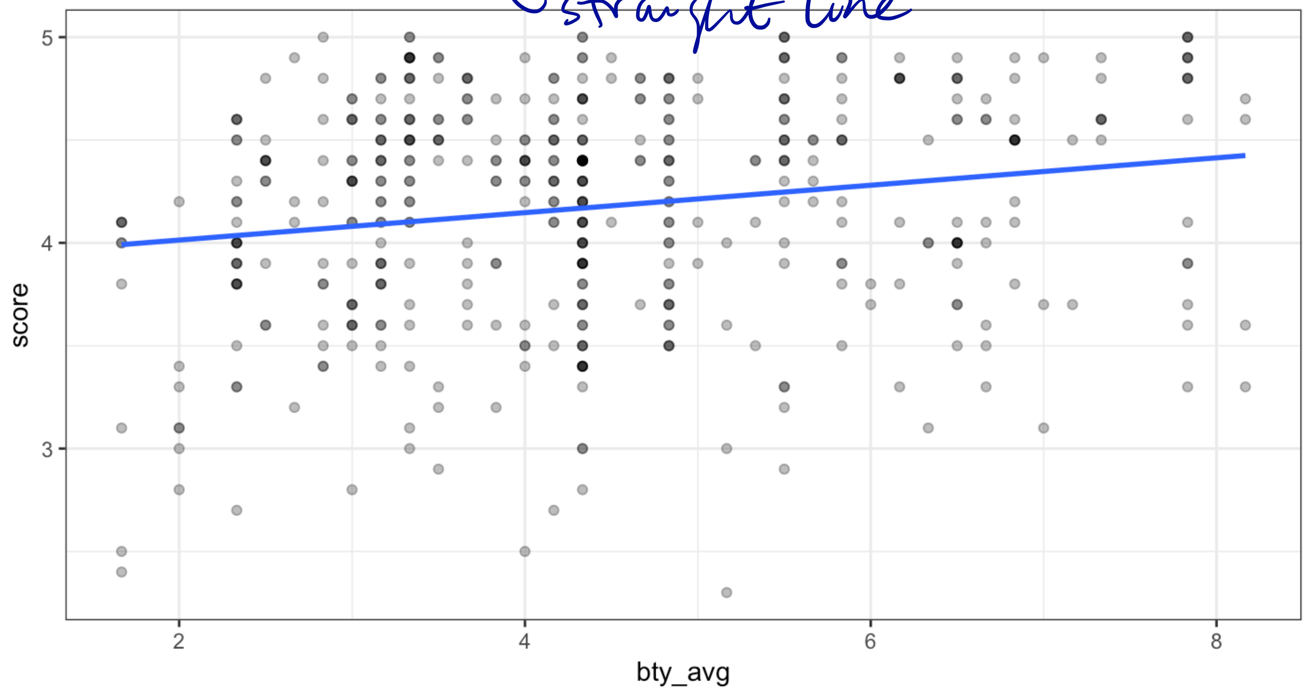
```
ggplot(evals, aes(x=bty_avg, y=score)) + geom_point(alpha=0.3) + theme_bw()
```



Is there a relationship between **score** and **bty_avg**?

```
ggplot(evals, aes(x=bty_avg, y=score)) + geom_point(alpha=0.3)  + theme_bw() +
  geom_smooth(method = "lm", fill=NA)
```

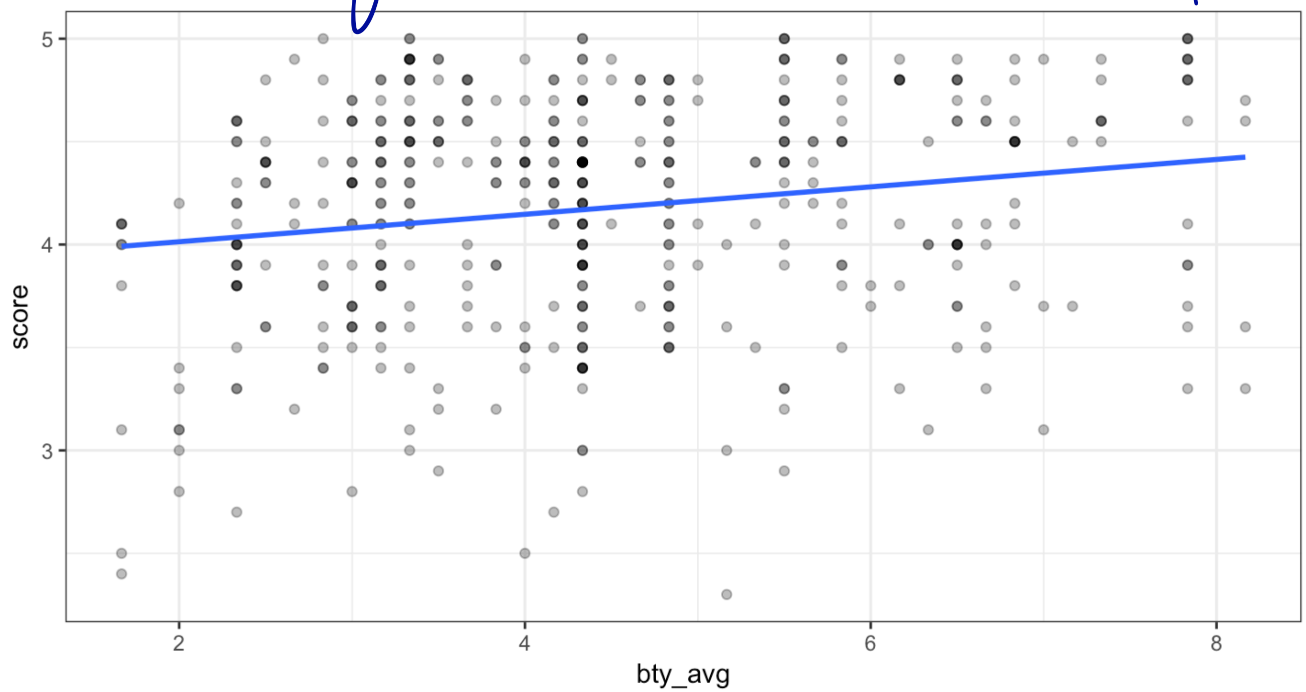
straight line



What would the slope be if there was no relationship?

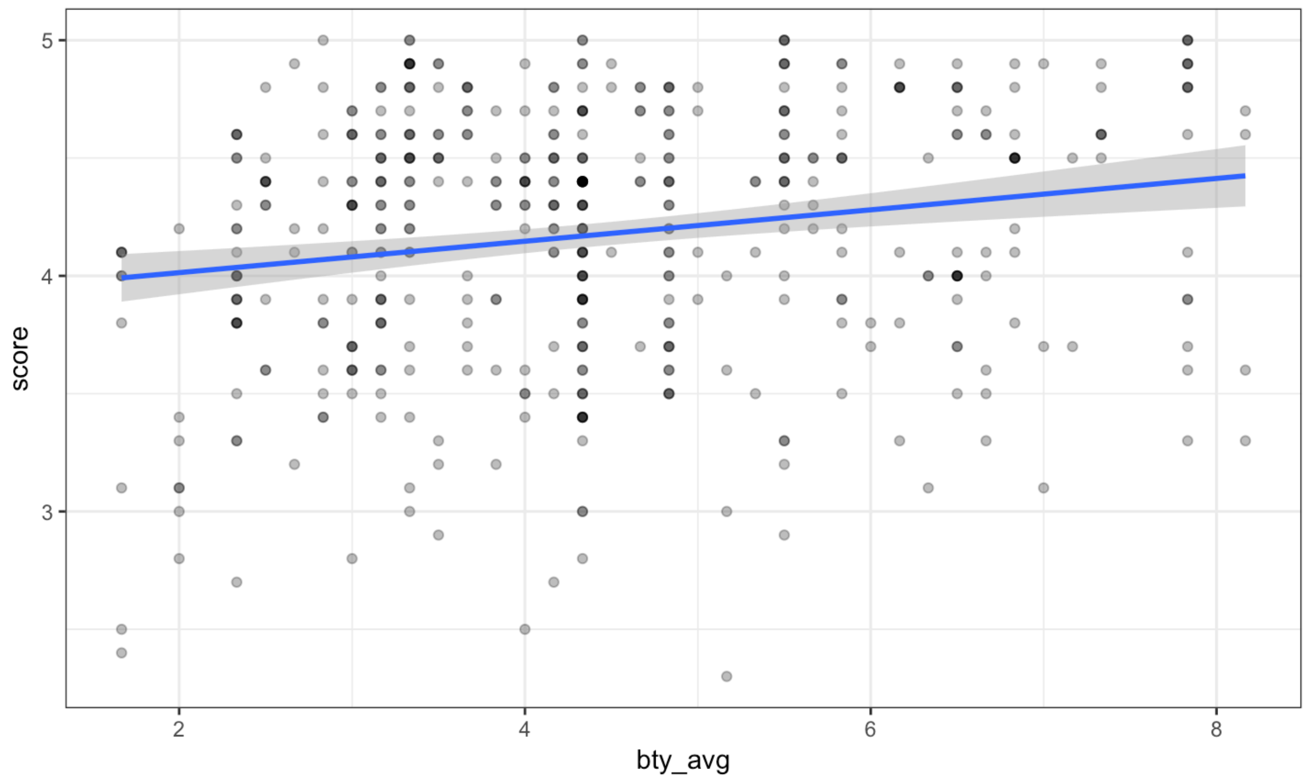
slope would be 0

Horizontal line — line predicts same value of y for every x



By default, geom_smooth gives a confidence interval for the fitted line (or curve)

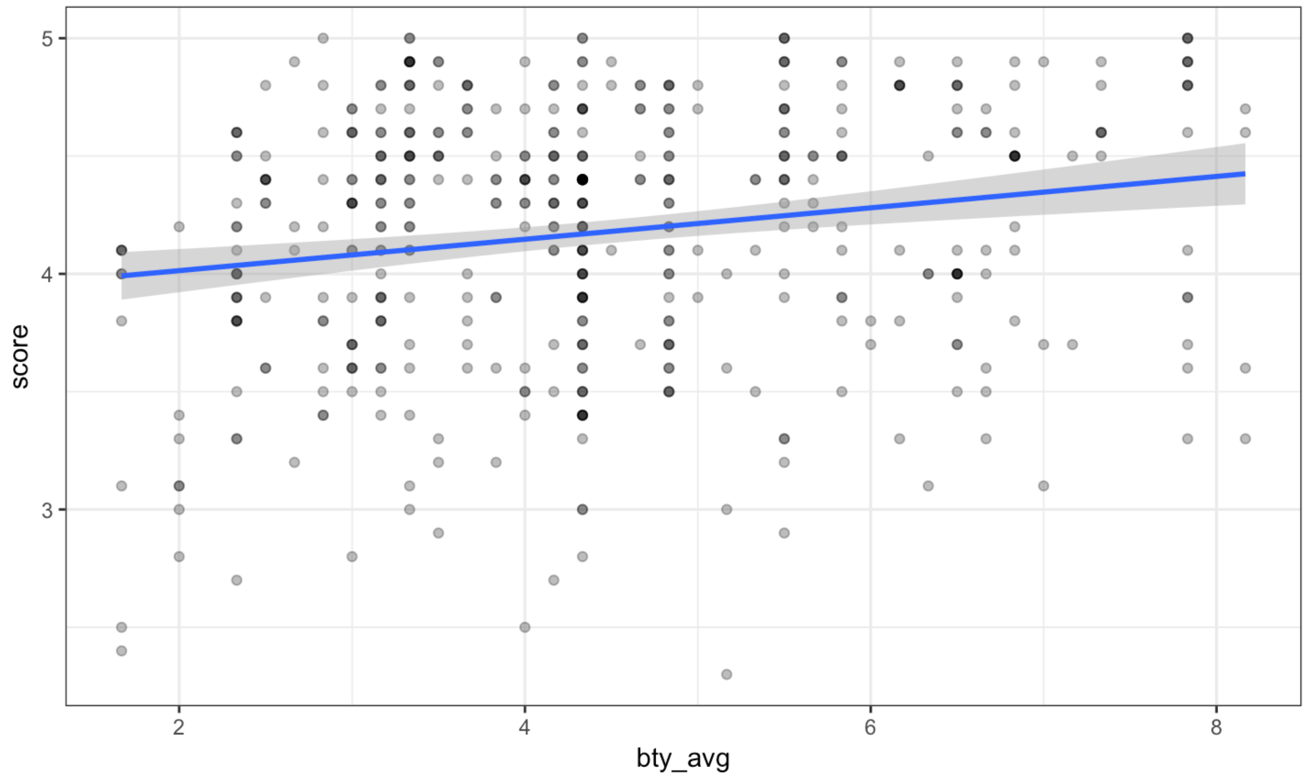
```
ggplot(evals, aes(x=bty_avg, y=score)) + geom_point(alpha=0.3)  + theme_bw() +
  geom_smooth(method = "lm")
```



Inference for regression part 1: Confidence interval for the slope

The grey shaded area around the fitted regression line is a 95% confidence interval for the slope.

- The width of the confidence interval varies with the independent variable `bty_avg`.
- The confidence interval is wider at the extremes; the regression is estimated most precisely near the mean of the independent variable.
- The confidence interval for the slope shown is calculated based on a probability model that assumes all observations are independent and that the error terms have a symmetric, bell-shaped distribution.
- Confidence intervals for the slope can also be calculated using the bootstrap.



Does the confidence interval indicate that 0 is a possible value for β_1 (the parameter for the slope)?

No. Because a horizontal doesn't fit within the confidence interval (grey bands)

Inference for regression part 2: Hypothesis test for the slope

$score = \beta_0 + \beta_1 \, bty\_avg + \varepsilon$

Output from the summary command for the estimated regression coefficients:

```
summary(lm(score ~ bty_avg, data=evals))$coefficients
```

P-values

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 3.88033795 0.07614297 50.961212 1.561043e-191
## bty_avg     0.06663704 0.01629115  4.090382  5.082731e-05
```

$\widehat{score} = 3.88 + 0.0666 \, bty\_avg$

R gives results for an hypothesis test with hypotheses:

$$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0$$

P-value

For $H_0 : \beta_0 = 0$ vs $H_a : \beta_0 \neq 0$

P-value = 1.561043e-191

```
##                 Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 3.88033795 0.07614297 50.961212 1.561043e-191
## bty_avg     0.06663704 0.01629115  4.090382  5.082731e-05
```

- The estimate of the slope is 0.06664.
- When using the `lm` function, by default the P-value is calculated based on a probability model that assumes all observations are independent and that the error terms have a symmetric, bell-shaped distribution.
- The P-value is $5.08 \times 10^{-5} = 0.0000508$

Does the hypothesis test for the slope indicate that the slope is different from 0?

Null hypothesis : $H_0 : \beta_1 = 0$ (slope = 0)

Observed slope from data : 0.06664

Assuming slope is 0 chance of getting a slope as different from 0 as 0.06664 is .0000508

We conclude that we have strong evidence against the null hypothesis, that is, there is a relationship.

What other factors might affect course evaluations?

The Economist

Topics Current edition More

Academic sexism

Research suggests students are biased against female lecturers

How long does that prejudice last?

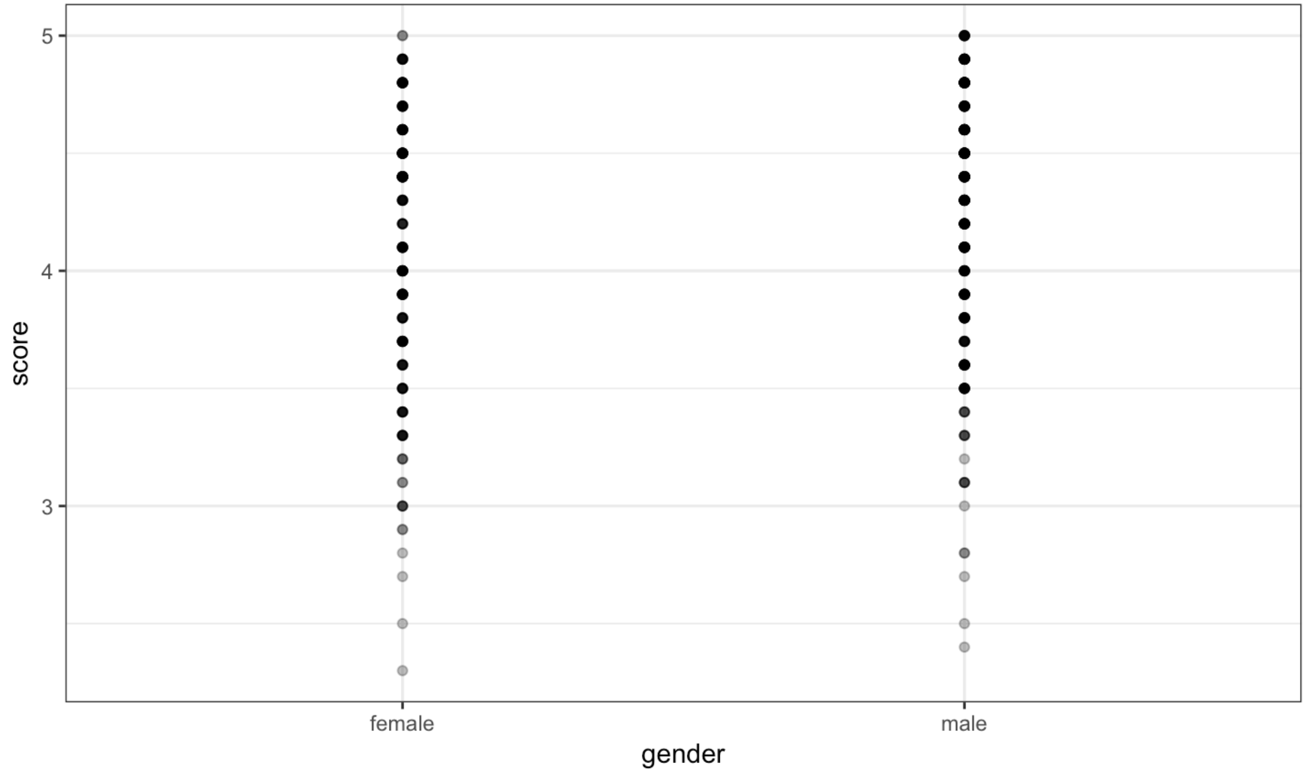
Print edition | Science and technology ›

Sep 21st 2017

Regression when the independent variable is a categorical variable

Relationship between **score** and **gender**?

```
ggplot(evals, aes(x=gender, y=score)) + geom_point(alpha=0.3) + theme_bw()
```



Regression with `gender` as the independent variable

```
lm(score ~ gender, data=evals)$coefficients
```

```
## (Intercept)  gendermale
##   4.0928205   0.1415078
```

$$\widehat{score} = 4.09 + 0.14\, gender\_is\_male$$

How to interpret the slope:
On average, course evaluation scores for male professors are 0.14 higher than for female professors.

$$\boxed{\widehat{score} = 4.09 + 0.14\, gender\_is\_male}$$

- In regression, `R` encodes categorical independent variables as **indicator variables** (also called **dummy variables**).
- `R` picks a baseline value of the categorical variable. Here the baseline level is `female`.
- The indicator variable `gender_is_male` is 1 for observations for which `gender` is male and 0 otherwise.
- For females,

gender_is_male = 0,

$$\widehat{score} = 4.09$$

- For males,

gender_is_male = 1,

$$\widehat{score} = 4.09 + 0.14 = 4.23$$

Could the difference between the mean score for males and females just be due to chance?

The regression model is

$$score = \beta_0 + \beta_1\ gender\_is\_male + \epsilon$$

where

$$gender\_is\_male = \begin{cases} 1 & \text{if } gender \text{ is } male \\ 0 & \text{if } gender \text{ is } female \end{cases}$$

We can answer the question with an hypothesis test with hypotheses

$$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0$$

```
summary(lm(score ~ gender, data=evals))$coefficients
```

P-values

```
##               Estimate Std. Error    t value    Pr(>|t|)
## (Intercept) 4.0928205 0.03866539 105.852305 0.000000000
## gendermale  0.1415078 0.05082127   2.784422 0.005582967
```

What conclusion do we make?

P-Value is 0.00558

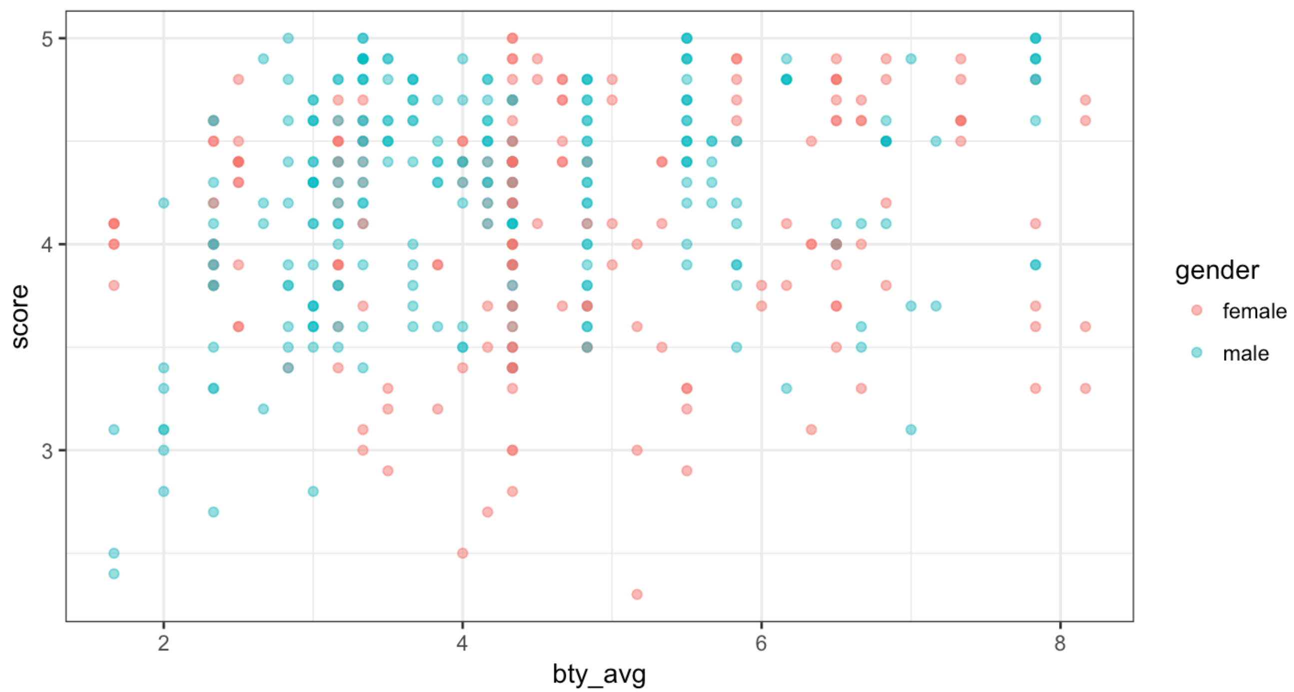
Strong evidence against the null hypothesis

Conclude there is a difference in predicted score between male and female professors

Is the regression line the same for two groups?

Is the relationship between **`score`** and **`bty_avg`** the same for male and female professors?

```
ggplot(evals, aes(x=bty_avg, y=score, colour=gender)) +
  geom_point(alpha=0.5)  + theme_bw()
```



Model 1: Multiple regression:

$$score = \beta_0 + \beta_1 \; gender\_is\_male + \beta_2 \; bty\_avg + \epsilon$$

Model 1 for male professors:

$$score = (\beta_0 + \beta_1) + \beta_2 \; bty\_avg + \epsilon$$

Model 1 for female professors:

$$score = \beta_0 + \beta_2 \; bty\_avg + \epsilon$$

How would you describe these two lines?

2 parallel lines (same slope, different intercepts)

Fitted parallel lines

```
parallel_lines <- lm(score ~ gender + bty_avg, data=evals)
parallel_lines$coefficients
```

```
## (Intercept)  gendermale     bty_avg
##  3.74733824  0.17238955  0.07415537
```

$\hat{\beta}_0$ $\hat{\beta}_1$ $\hat{\beta}_2$

For males: $\widehat{score} = (3.747 + 0.172) + 0.074\, bty\_avg$

For females: $\widehat{score} = 3.747 + 0.074\, bty\_avg$

Plotting the parallel lines

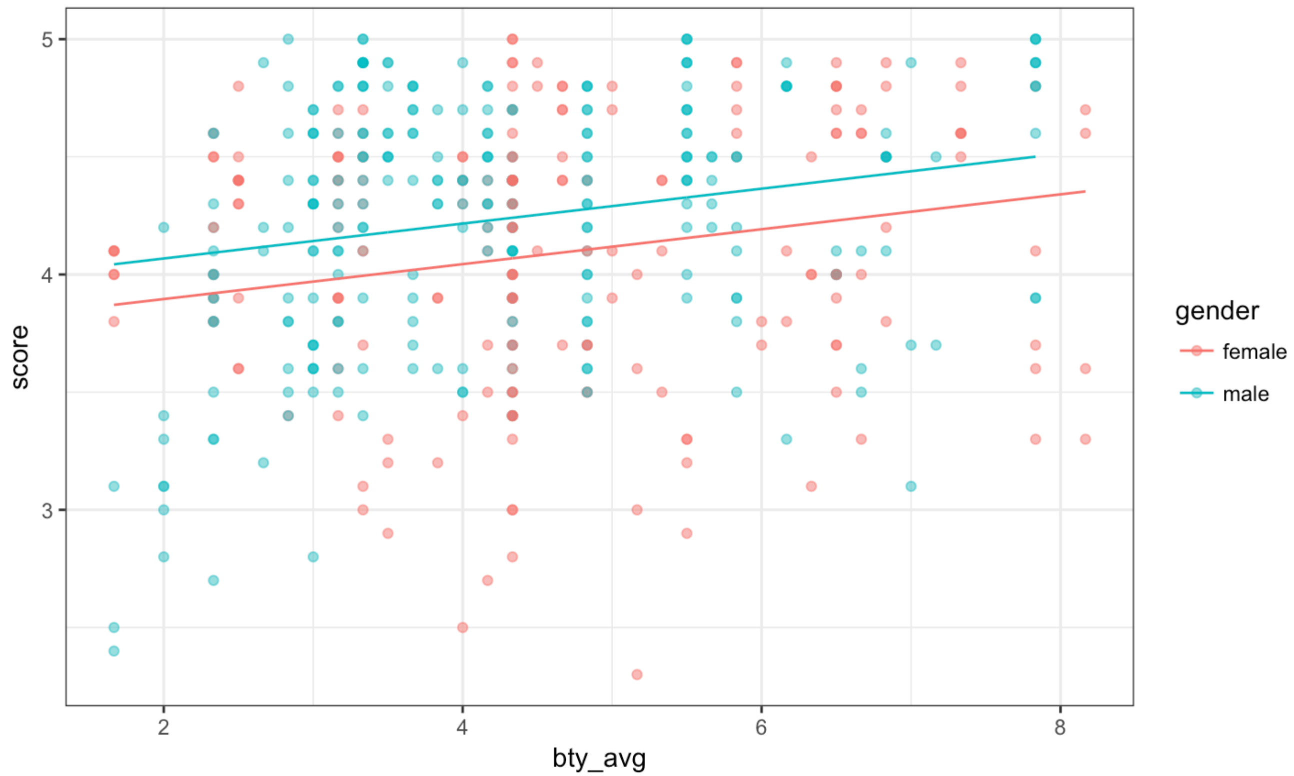
The `augment` function (in the library `broom`) creates a data frame with predicted values (`.fitted`), residuals, etc. for linear model output.

```
library(broom)
augment(parallel_lines)
```

```
##      score gender bty_avg  .fitted    .se.fit       .resid        .hat
## 1      4.7 female   5.000 4.118115 0.03826383  0.581884899 0.005238155
## 2      4.1 female   5.000 4.118115 0.03826383 -0.018115101 0.005238155
## 3      3.9 female   5.000 4.118115 0.03826383 -0.218115101 0.005238155
## 4      4.8 female   5.000 4.118115 0.03826383  0.681884899 0.005238155
## 5      4.6   male   3.000 4.142194 0.03808791  0.457806096 0.005190100
## 6      4.3   male   3.000 4.142194 0.03808791  0.157806096 0.005190100
## 7      2.8   male   3.000 4.142194 0.03808791 -1.342193904 0.005190100
## 8      4.1   male   3.333 4.166888 0.03551641 -0.066887644 0.004512941
## 9      3.4   male   3.333 4.166888 0.03551641 -0.766887644 0.004512941
## 10     4.5 female   3.167 3.982188 0.04495870  0.517811698 0.007231509
## 11     3.8 female   3.167 3.982188 0.04495870 -0.182188302 0.007231509
## 12     4.5 female   3.167 3.982188 0.04495870  0.517811698 0.007231509
## 13     4.6 female   3.167 3.982188 0.04495870  0.617811698 0.007231509
## 14     3.9 female   3.167 3.982188 0.04495870 -0.082188302 0.007231509
## 15     3.9 female   3.167 3.982188 0.04495870 -0.082188302 0.007231509
```

Join up the fitted values to plot the parallel lines model

```
ggplot(evals, aes(x=bty_avg, y=score, colour=gender)) +
  geom_point(alpha=0.5) + theme_bw() +
  geom_line(data = augment(parallel_lines), aes(y=.fitted, colour=gender))
```



Lines for each gender that aren't parallel

Add an independent variable to the model that is the product of `gender_is_male` and `bty_avg`. This is called an **interaction term**.

Model 2:

$$score = \beta_0 + \beta_1\ gender\_is\_male + \beta_2\ bty\_avg + \beta_3\ (gender\_is\_male \times bty\_avg) + \epsilon$$

Model 2 for male professors:

$$score = \beta_0 + \beta_1 + \beta_2\ bty\_avg + \beta_3\ bty\_avg + \epsilon$$

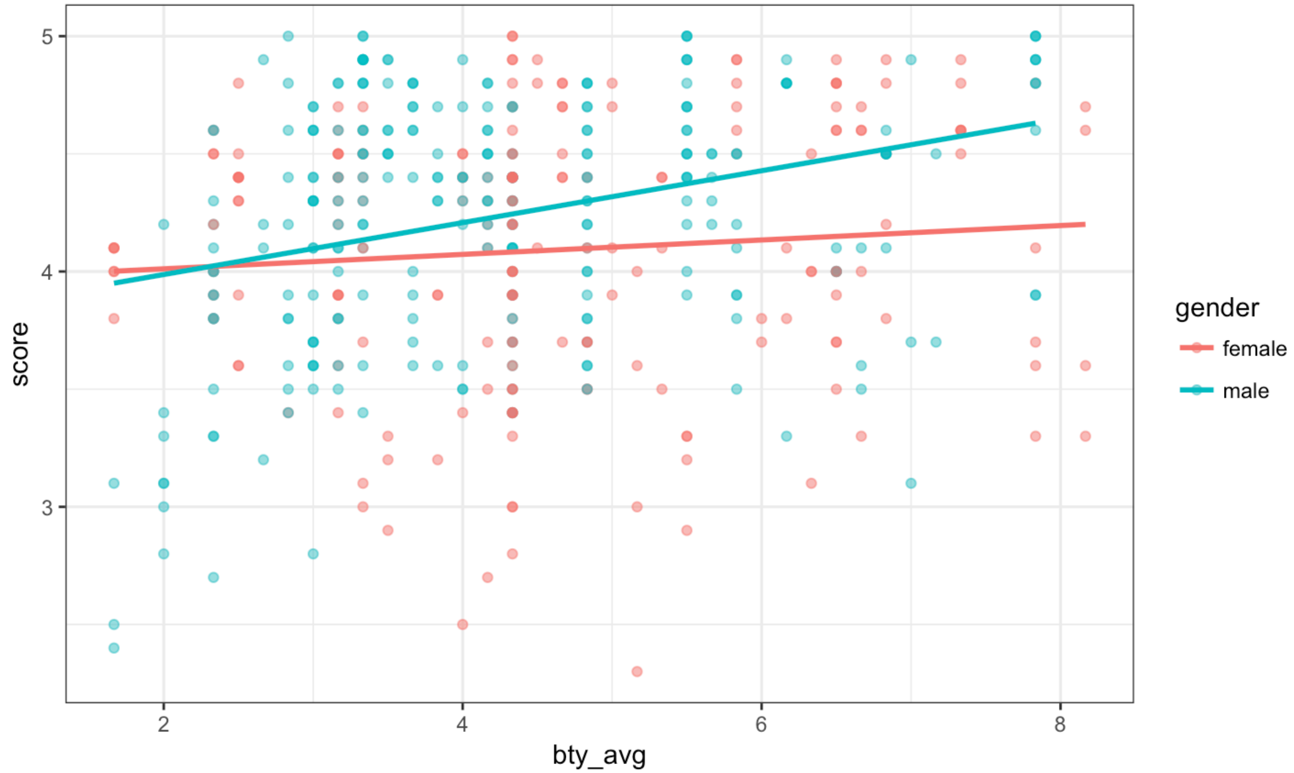
$$score = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)\ bty\_avg + \epsilon$$

Model 2 for female professors:

$$score = \beta_0 + \beta_2\ bty\_avg + \epsilon$$

Plot of non-parallel lines

```
ggplot(evals, aes(x=bty_avg, y=score, colour=gender)) + geom_point(alpha=0.5) +
        geom_smooth(method=lm, fill=NA) + theme_bw()
```



Fitted lines for male and female professors

Including the term **bty_avg*gender** on the right-side of the model specification in **lm** includes the interaction term plus both of the variables in the model.

```
summary(lm(score ~ bty_avg*gender, data=evals))$coefficients
```

```
##                        Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)          3.95005984 0.11799986 33.475124 2.920267e-125
## bty_avg              0.03064259 0.02400361  1.276582  2.023952e-01
## gendermale          -0.18350903 0.15349459 -1.195541  2.324931e-01
## bty_avg:gendermale   0.07961855 0.03246948  2.452105  1.457376e-02
```

times

Null hypothesis!

$\beta_0 = 0$

$\beta_1 = 0$

$\beta_2 = 0$

$\beta_3 = 0$

What are the fitted lines for male and for female professors?

Females

$$\widehat{score} = 3.95 + 0.031\ bty\_avg$$

Males

$$\widehat{score} = (3.95 - 0.18) + (0.031 + .08)\ bty\_avg$$

Could the difference in the slopes for male and female professors just be due to chance?

Model:

$$score = \beta_0 + \beta_1 \ gender\_is\_male + \beta_2 \ bty\_avg + \beta_3 \ (gender\_is\_male \times bty\_avg) + \epsilon$$

What would be appropriate hypotheses to test?

$H_0: \beta_3 = 0$
$H_a: \beta_3 \neq 0$

What do you conclude?

P-value = 0.0146

Some evidence that slopes are different for male & female professors.

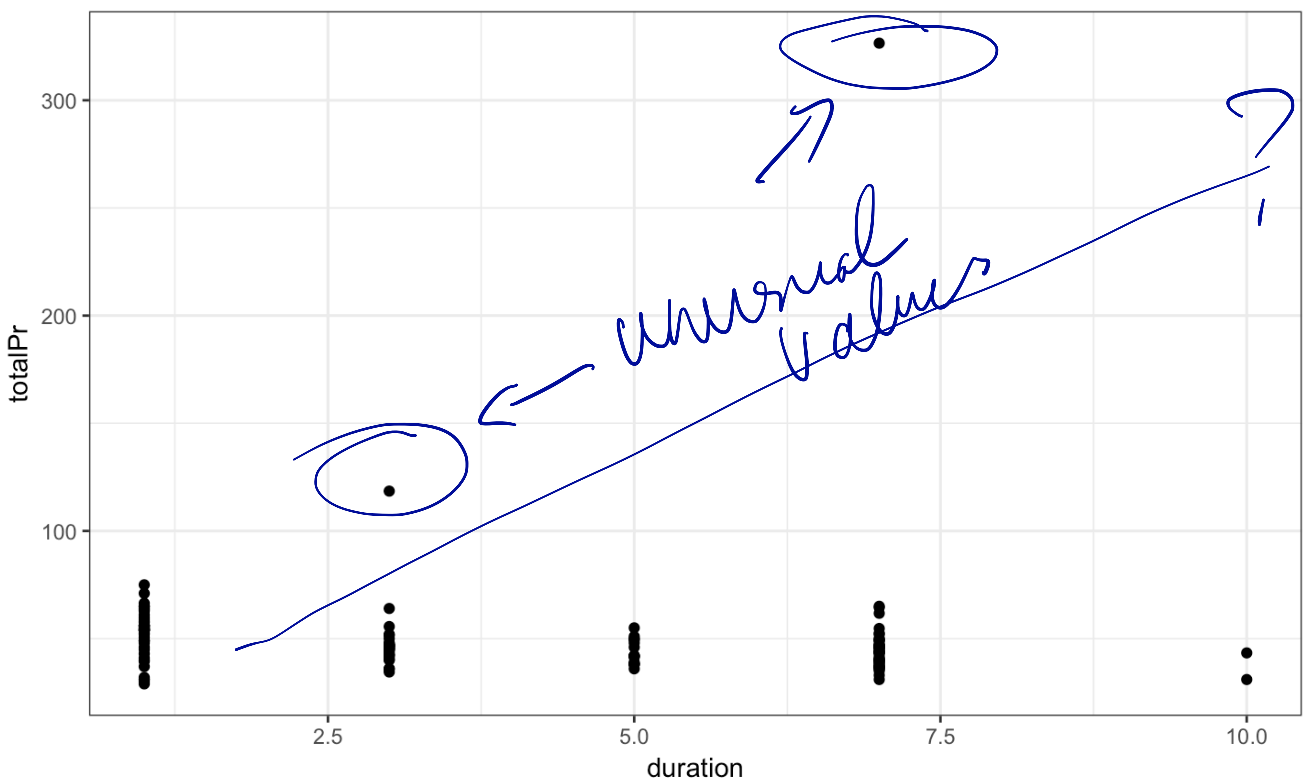
Example: eBay auctions of *Mario Kart*

- Items can be sold on ebay.com through an auction.
- The person who bids the highest price before the auction ends purchases the item.
- The `marioKart` dataset in the `openintro` package includes eBay sales of the game *Mario Kart* for Nintendo Wii in October 2009.
- Do longer auctions (`duration`, in days) result in higher prices (`totalPr`)?

```r
library(openintro)
glimpse(marioKart)
```

```
## Observations: 143
## Variables: 12
## $ ID         <dbl> 150377422259, 260483376854, 320432342985, 280405224...
## $ duration   <int> 3, 7, 3, 3, 1, 3, 1, 1, 3, 7, 1, 1, 1, 1, 7, 7, 3, ...
## $ nBids      <int> 20, 13, 16, 18, 20, 19, 13, 15, 29, 8, 15, 15, 13, ...
## $ cond       <fctr> new, used, new, new, new, new, used, new, used, us...
## $ startPr    <dbl> 0.99, 0.99, 0.99, 0.99, 0.01, 0.99, 0.01, 1.00, 0.9...
## $ shipPr     <dbl> 4.00, 3.99, 3.50, 0.00, 0.00, 4.00, 0.00, 2.99, 4.0...
## $ totalPr    <dbl> 51.55, 37.04, 45.50, 44.00, 71.00, 45.00, 37.02, 53...
## $ shipSp     <fctr> standard, firstClass, firstClass, standard, media,...
## $ sellerRate <int> 1580, 365, 998, 7, 820, 270144, 7284, 4858, 27, 201...
## $ stockPhoto <fctr> yes, yes, no, yes, yes, yes, yes, yes, yes, no, ye...
## $ wheels     <int> 1, 1, 1, 1, 2, 0, 0, 2, 1, 1, 2, 2, 2, 2, 1, 0, 1, ...
## $ title      <fctr> ~~ Wii MARIO KART &amp; WHEEL ~ NINTENDO Wii ~ BRA...
```

```
ggplot(marioKart, aes(x=duration, y=totalPr)) + geom_point() + theme_bw()
```

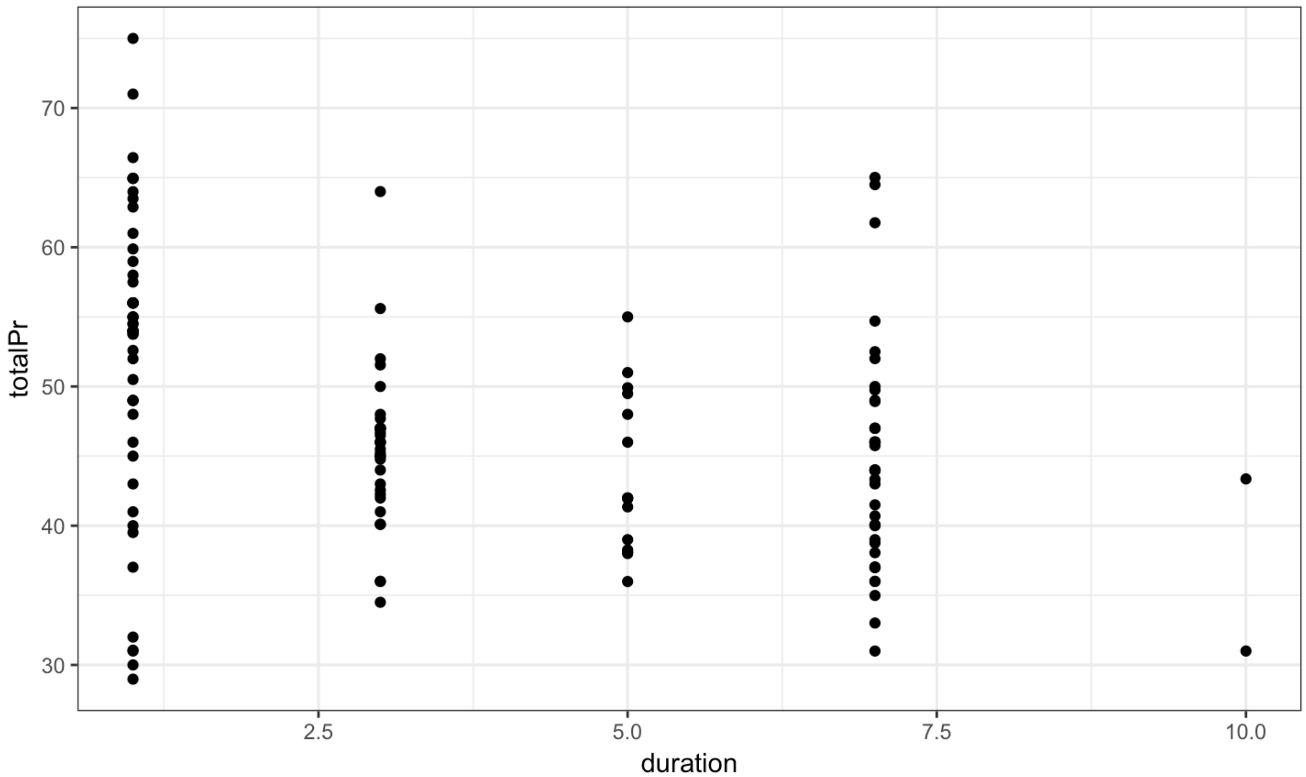


What should we do with the two outlying values of `totalPr`?

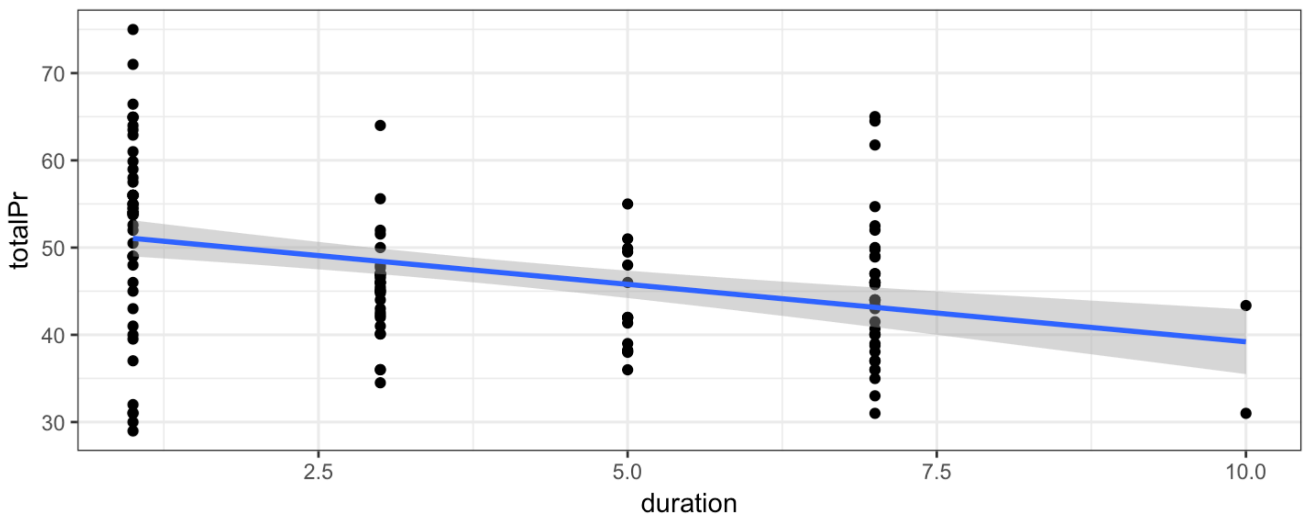
- Remove outliers only if there is a good reason.
- In these two auctions, and only these two auctions, the game was sold with other items.

```
# create a data set without the outliers
marioKart2 <- marioKart %>% filter(totalPr < 100)
```

```
ggplot(marioKart2, aes(x=duration, y=totalPr)) + geom_point() + theme_bw()
```



```
ggplot(marioKart2, aes(x=duration, y=totalPr)) + geom_point() + theme_bw() +
                      geom_smooth(method = "lm")
```



There appears to be a negative relationship between `totalPr` and `duration`. That is, the longer an item is on auction, the lower the price.

Does this make sense?

You would think having a longer auction gives the price a chance to rise.

Maybe there actually isn't a relationship.

We can investigate if the data are consistent with a slope of 0.

```
summary(lm(totalPr ~ duration, data=marioKart2))$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 52.373584  1.2607560 41.541411 3.010309e-80
## duration    -1.317156  0.2769021 -4.756756 4.866701e-06
```

P-value 0.0000049 for test with $H_0: \beta_1 = 0$

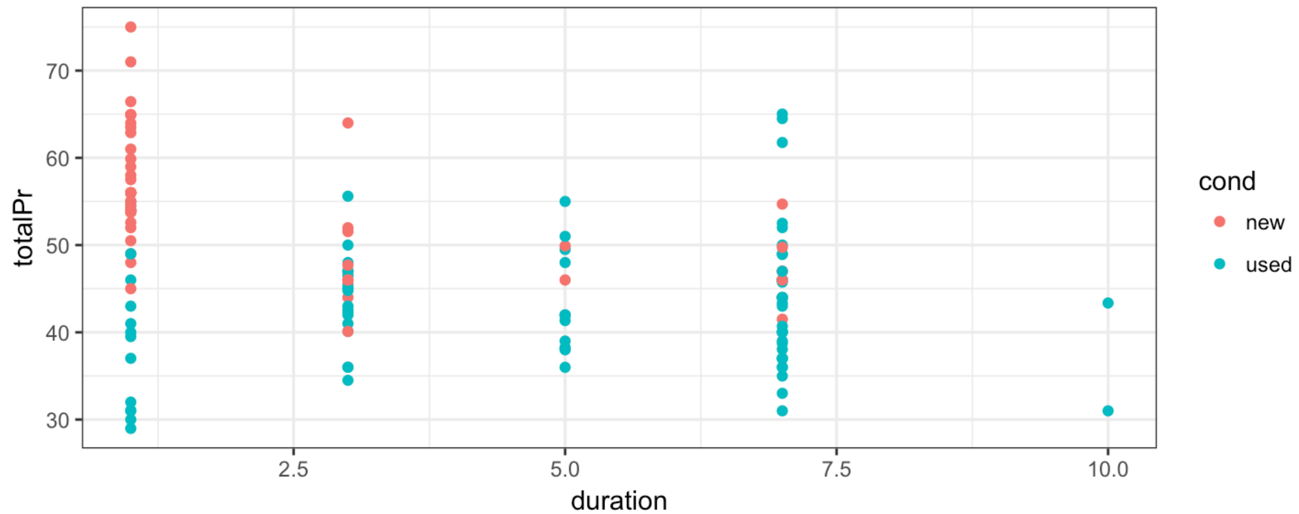
We have strong evidence that the slope is not 0.

There must be something else affecting the relationship …

Consider the role of `cond`.

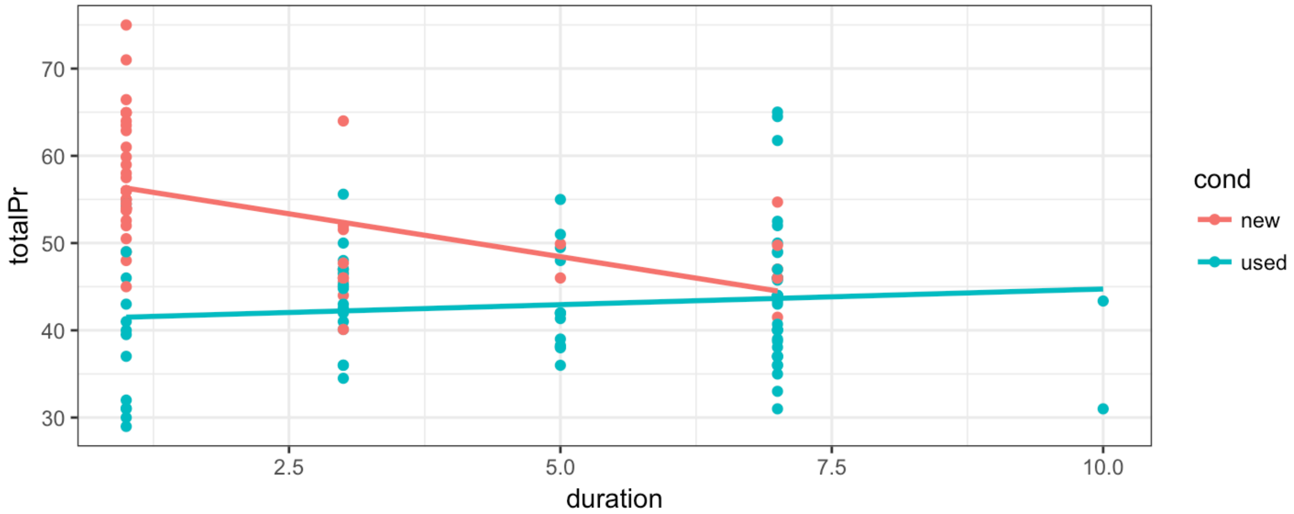
`cond` is a categorical variable for the game's condition, either `new` or `used`.

```
ggplot(marioKart2, aes(x=duration, y=totalPr, color=cond)) +
  geom_point() + theme_bw()
```



New games, which are more desirable, were mostly sold in one-day auctions.

```r
ggplot(marioKart2, aes(x=duration, y=totalPr, color=cond)) +
  geom_point() + geom_smooth(method="lm", fill=NA) + theme_bw()
```



- Considering `cond` changes the nature of the relationship between `totalPr` and `duration`.
- This is an example of **Simpson's Paradox** in which the nature of a relationship that we see in all observations changes when we look at sub-groups.

The fitted lines

```
summary(lm(totalPr ~ duration*cond, data=marioKart2))$coefficients
```

```
##                        Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)           58.268226  1.3664729 42.641332 5.832075e-81
## duration              -1.965595  0.4487799 -4.379865 2.341705e-05
## condused             -17.121924  2.1782581 -7.860374 1.013608e-12
## duration:condused      2.324563  0.5483731  4.239016 4.101561e-05
```

— Small P-value
⇒ strong evidence slopes are different for new & used games

condused = { 1 if game is used
             { 0 if game is new

An example of a variable affecting a relationship between two variables in a non-regression setting: Data in two-way tables

A Classic Example: Treatment for kidney stones

Source of data: *British Medical Journal (Clinical Research Edition)* March 29, 1986

- Observations are patients being treated for kidney stones.
- `treatment` is one of 2 treatments (`A` or `B`)
- `outcome` is `success` or `failure` of the treatment

```
kidney_stones %>% count(treatment, outcome)
```

```
## # A tibble: 4 x 3
##   treatment outcome     n
##       <chr>   <chr> <int>
## 1         A failure    77
## 2         A success   273
## 3         B failure    61
## 4         B success   289
```

Which treatment would you choose?

The table below shows counts of patients being treated for kidney stones. Which treatment would you choose?

Counts of number of observations

Treatment

Outcome

	Treatment A	Treatment B	TOTAL
Success	273	289	562
Failure	77	61	138
TOTAL	350	350	700

What would make it easier to decide which treatment is better?

Proportion of observations in each cell in the table:

	Treatment A	Treatment B	TOTAL
Success	0.39	0.41	0.80
Failure	0.11	0.09	0.20
TOTAL	0.50	0.50	1.00

Proportion of observations in each row and column:

Column Percentages

	Treatment A	Treatment B
Success	0.78	0.83
Failure	0.22	0.17

% success / failure for treatment A

Row Percentages

	Treatment A	Treatment B
Success	0.49	0.51
Failure	0.56	0.44

% A / B that were success

+ summarize in one

Proportion of successes in each column:

group_by

```
kidney_stones %>%
  count(treatment, outcome) %>%
  group_by(treatment) %>%
  mutate(perc_success = n / sum(n)) %>%
  filter(outcome=="success")
```

code to produce success rates for each treatment

```
## # A tibble: 2 x 4
## # Groups:   treatment [2]
##   treatment outcome     n perc_success
##       <chr>   <chr> <int>        <dbl>
## 1         A success   273    0.7800000
## 2         B success   289    0.8257143
```

Which treatment would you choose?

treatment B has a higher % of success (83% vs 78%)

Some vocabulary

Recall: The distribution of a variable is the pattern of values in the data for that variable, showing the frequency or relative frequency (proportions) of the occurrence of the values relative to each other.

We can also look at the **joint distribution** of two variables. If both variables are categorical, we can see their joint distribution in a **contingency table** showing the counts of observations in each way the data can be cross-classifed.

Counts:

	Treatment A	Treatment B	TOTAL
Success	273	289	**562**
Failure	77	61	**138**
TOTAL	**350**	**350**	**700**

Proportions in the joint distributions:

(1st table on slide 46)

A **marginal distribution** is the distribution of only one of the variables in a contingency table.

Counts:

	Treatment A	Treatment B	TOTAL
Success	273	289	562
Failure	77	61	138
TOTAL	350	350	700

Proportions:

	Treatment A	Treatment B	TOTAL
Success	0.39	0.41	0.80
Failure	0.11	0.09	0.20
TOTAL	0.50	0.50	1.00

→ marginal distribution of outcome

→ marginal distribution of treatment

A **conditional distribution** is the distribution of a variable within a fixed value of a second variable.

Column Percentages

	Treatment A	Treatment B
Success	0.78	0.83
Failure	0.22	0.17

Row Percentages

	Treatment A	Treatment B
Success	0.49	0.51
Failure	0.56	0.44

Conditional distribution of outcome given treatment = A

What percentage of successes were Treatment A?

49%

Given success, what's prob. trt A

What percentage of Treatment A surgeries resulted in a success?

78%

Given trt A, whats prob of success

Some notation:

Not responsible for.

$P(E_1)$ is the probability of an event E_1

$P(E_1 \mid E_2)$ is the probability of E_1 **given** that event E_2 has occurred. It is a **conditional probability**.

Example:

- What is the probability it will rain tomorrow?
- What is the probability it will rain tomorrow given that it is raining today?

Not responsible for.

	Treatment A	Treatment B	TOTAL
Success	0.39	0.41	0.80
Failure	0.11	0.09	0.20
TOTAL	0.50	0.50	1.00

Column Percentages

	Treatment A	Treatment B
Success	0.78	0.83
Failure	0.22	0.17

Row Percentages

	Treatment A	Treatment B
Success	0.49	0.51
Failure	0.56	0.44

From the tables, we estimate:

$$P(\text{success}) = 0.80$$

$$P(\text{success} \mid \text{treatment A}) = 0.78$$

$$P(\text{success} \mid \text{treatment B}) = 0.83$$

Does there appear to be a relationship between success and treatment?

Yes! Success is more likely with treatment B.

Independence

Not responsible for

E_1 and E_2 are **independent** if $P(E_1 \mid E_2) = P(E_1)$.

That is, the conditional distribution of one variable is the same for all values of the other variable.

It appears that success and treatment are not independent.

Some additional information

- A is an invasive open surgery treatment
- B is a new less invasive treatment
- Doctors get to choose the treatment, depending on the patient
- What might influence how a doctor chooses a treatment for their patient?

Kidney stones come in various sizes

```
kidney_stones %>%
  count(size, treatment, outcome) %>%
  group_by(size, treatment) %>%
  mutate(per_success = n / sum(n)) %>%
  filter(outcome=="success")
```

```
## # A tibble: 4 x 5
## # Groups:   size, treatment [4]
##    size treatment outcome     n per_success
##   <chr>     <chr>   <chr> <int>       <dbl>
## 1 large         A success   192   0.7300380
## 2 large         B success    55   0.6875000
## 3 small         A success    81   0.9310345
## 4 small         B success   234   0.8666667
```

Column percentages:

All Stones

	A	B
Success	0.78	0.83
Failure	0.22	0.17

Small Stones

A	B
0.93	0.87
0.07	0.13

Large Stones

A	B
0.73	0.69
0.27	0.31

Which treatment is better?

A for both small and large stones

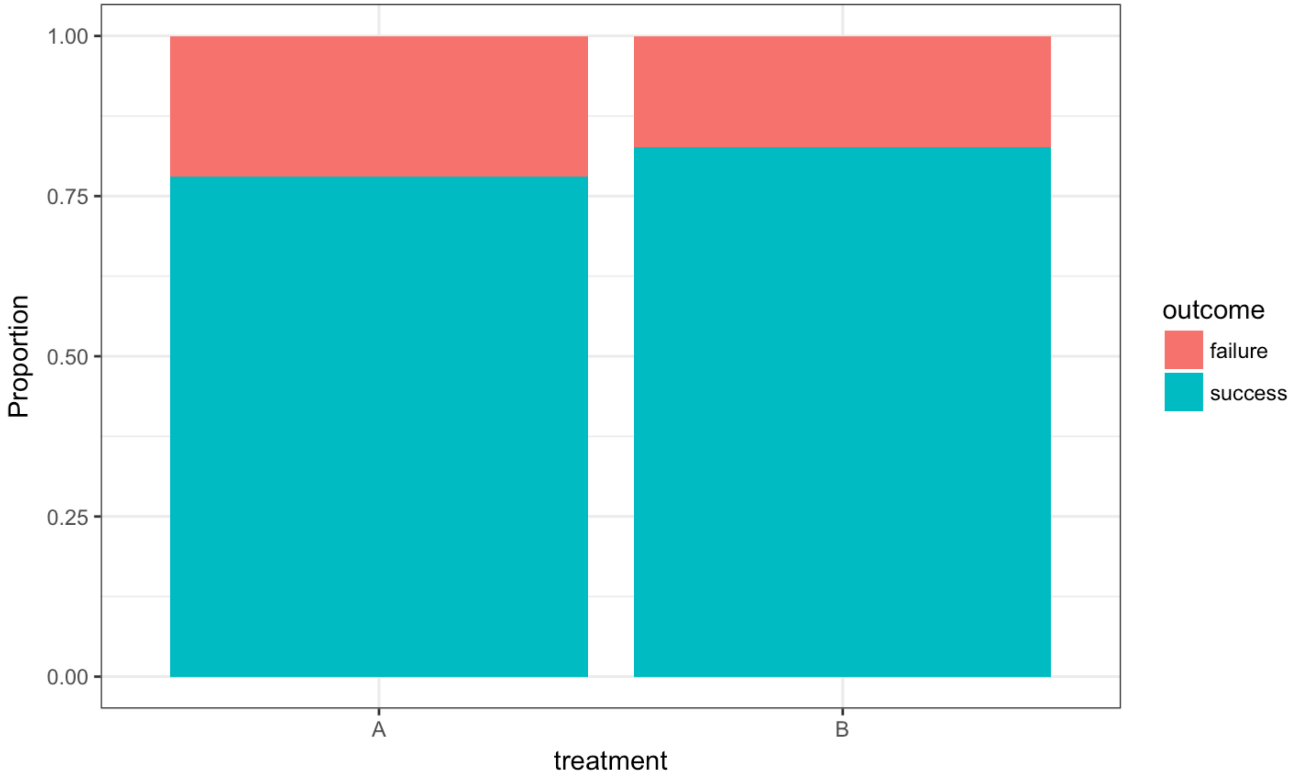
This example is another case of **Simpson's paradox**.

Moral of the story:

Be careful drawing conclusions from data!
It's important to understand how the data were collected and what other factors might have an affect.

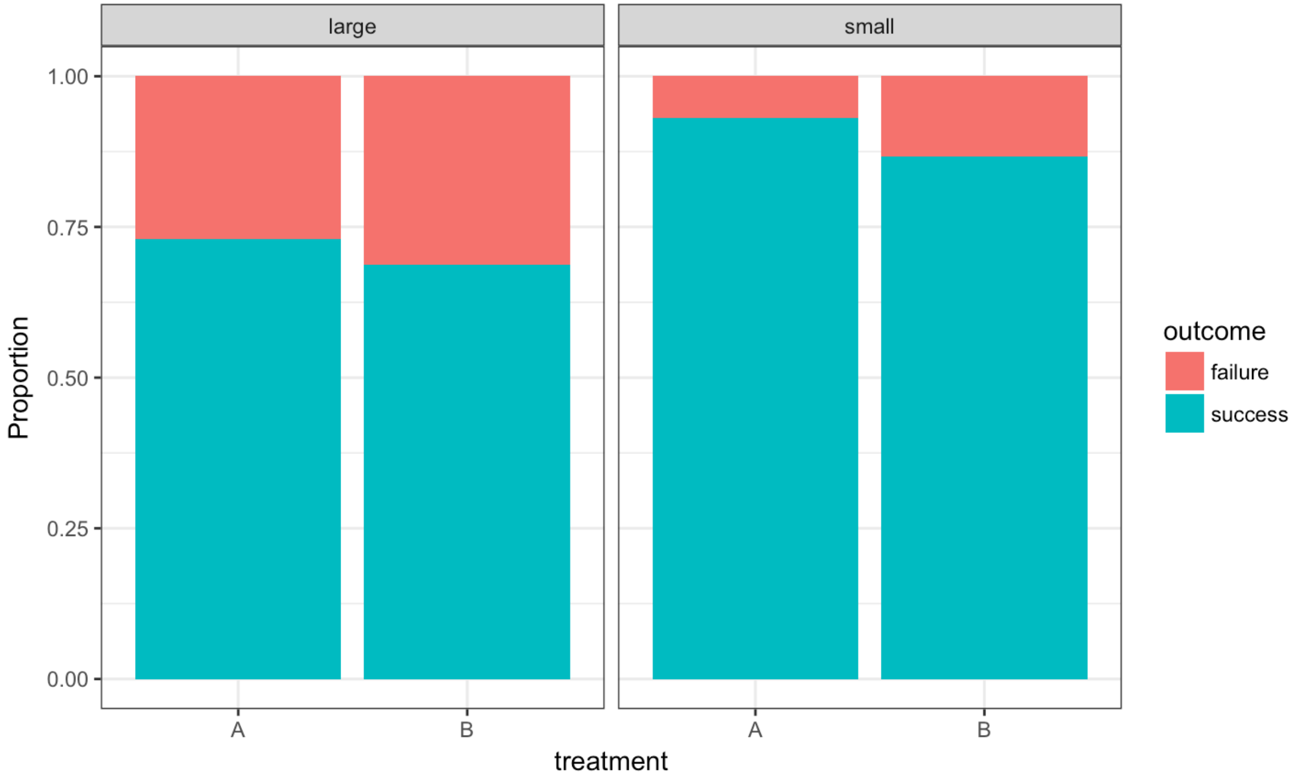
Visualizing the kidney stone data: treatment and outcome

```
ggplot(kidney_stones, aes(x=treatment, fill=outcome)) +
  geom_bar(position = "fill") + labs(y="Proportion") + theme_bw()
```



Visualizing the kidney stone data: treatment and outcome by size

```
ggplot(kidney_stones, aes(x=treatment, fill=outcome)) + geom_bar(position = "fill") +
  labs(y="Proportion") + facet_grid(. ~ size) + theme_bw()
```



Confounding

What is a confounding variable?

- When examining the relationship between two variables in observational studies, it is important to consider the possible effects of other variables.
- A third variable is a **confounding variable** if it affects the nature of the relationship between two other variables, so that it is impossible to know if one variable causes another, or if the observed relationship is due to the third variable.
- The possible presence of confounding variables means we must be cautious when interpreting relationships.

Examples of situations that may have confounding variables:

- A 2012 study showed that heavy use of marijuana in adolescence can negatively affect IQ.
  Is it possible that there are other variables, such as socioeconomic status, that is associated with both marijuana use and IQ?
- Another 2012 study showed that coffee drinking was inversely related to mortality.
  Should we all drink more coffee so we will live longer? Or is it possible that healthy people, who will live longer because they are healthy, are also more likely to drink coffee than unhealthy people?
- Many nutrition studies.
  Are people who are likely to stick to a diet different than those who won't in important ways?

How can confounding be avoided?

- Data can be collected through *experiments* or *observational studies*.
- In **observational studies**, data are collected without intervention. The data are measurements of existing characteristics of the individuals being measured.
- In **experiments**, an investigator imposes an intervention on the individuals being studied, randomly assigning some individuals to one treatment and randomly assigning other individuals to another treatment (sometimes this other treatment is a *control*).
    - Randomized experiments are often used when we want to be able to say a treatment **causes** a change in a measurement.
    - Other than the difference in treatment received, any differences between the individuals in the treatment and control groups are just due to random chance in their group assignment.

- In a randomized experiment, if there is a difference in our measurement of interest, we can conclude it was caused by the treatment, and not due to some other systematic difference that can confound our interpretation of the effect of the treatment.
- Example experiment from Week 5 lecture:
  Students were randomly assigned to be sleep-deprived or to have unrestricted sleep and how they learned a visual discrimination task was compared between these two groups.
- It's not always practical or ethical to carry out an experiment. For example, you can't randomly assign people to smoke marijuana.

Great care must be taken to deal with potential confounders in observational studies.

- In a randomized experiment, if there is a difference in our measurement of interest, we can conclude it was caused by the treatment, and not due to some other systematic difference that can confound our interpretation of the effect of the treatment.
- Example experiment from Week 5 lecture:
  Students were randomly assigned to be sleep-deprived or to have unrestricted sleep and how they learned a visual discrimination task was compared between these two groups.
- It's not always practical or ethical to carry out an experiment. For example, you can't randomly assign people to smoke marijuana.

Great care must be taken to deal with potential confounders in observational studies.