# STA130H1F

## Class #11

### Prof. Nathan Taback

### 2018-11-26

# Today's Class

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

# Inference for regression parameters

What affects course evaluations?

... other than the quality of the course ...

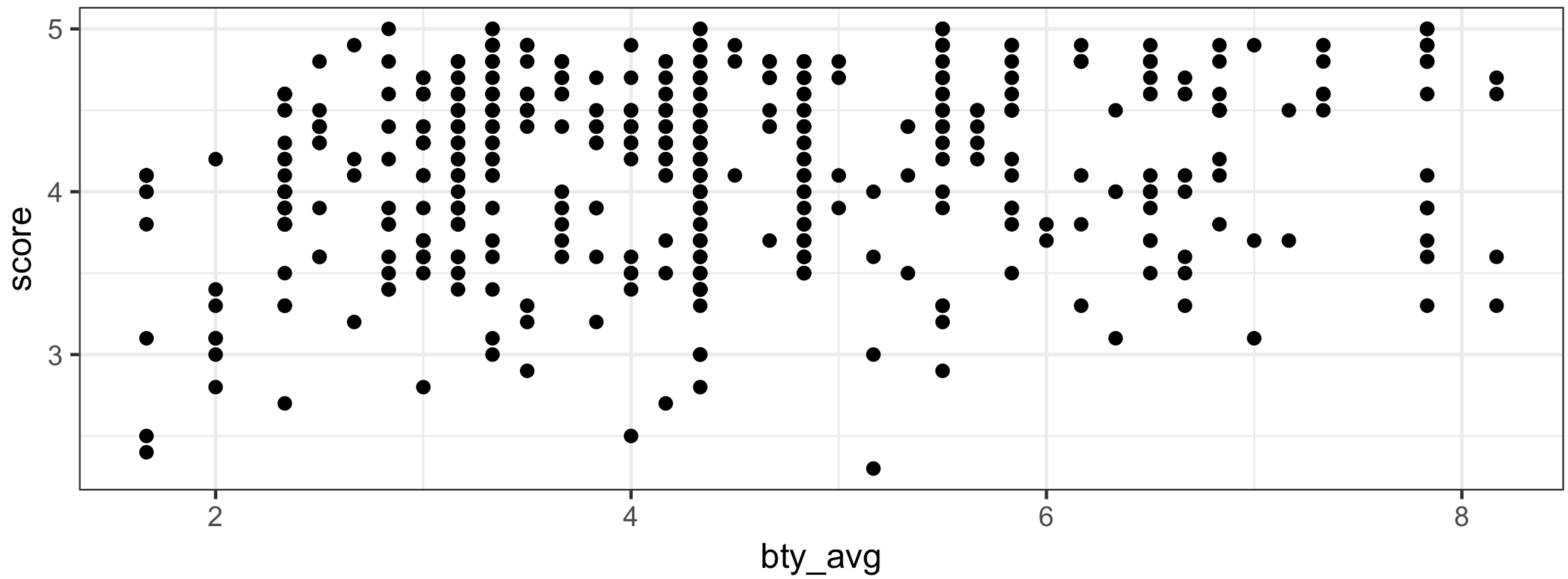
- Data from course evaluations for a random sample of courses at the University of Texas at Austin.
- Each observation corresponds to a course.
- `score` is the average student evaluation for the course.
- `bty_avg` is the average beauty rating of the professor, based on ratings of physical appear from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

```
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          <fct> tenure track, tenure track, tenure track, tenure...
## $ ethnicity     <fct> minority, minority, minority, minority, not mino...
## $ gender        <fct> female, female, female, female, male, male, male...
## $ language      <fct> english, english, english, english, english, eng...
## $ 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     <fct> upper, upper, upper, upper, upper, upper, upper,...
## $ cls_profs     <fct> single, single, single, single, multiple, multip...
## $ cls_credits   <fct> 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    <fct> not formal, not formal, not formal, not formal, ...
## $ pic_color     <fct> color, color, color, color, color, color, color,...
```

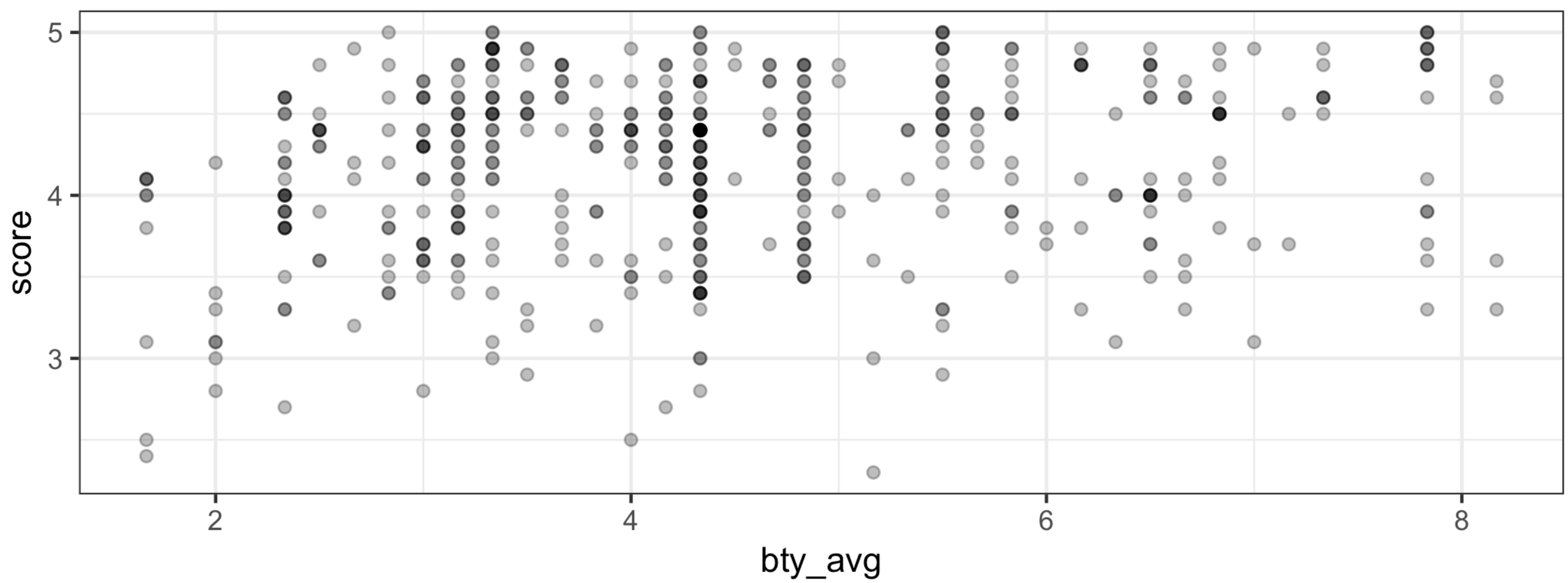
# Relationship between `score` and `bty_avg`?

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



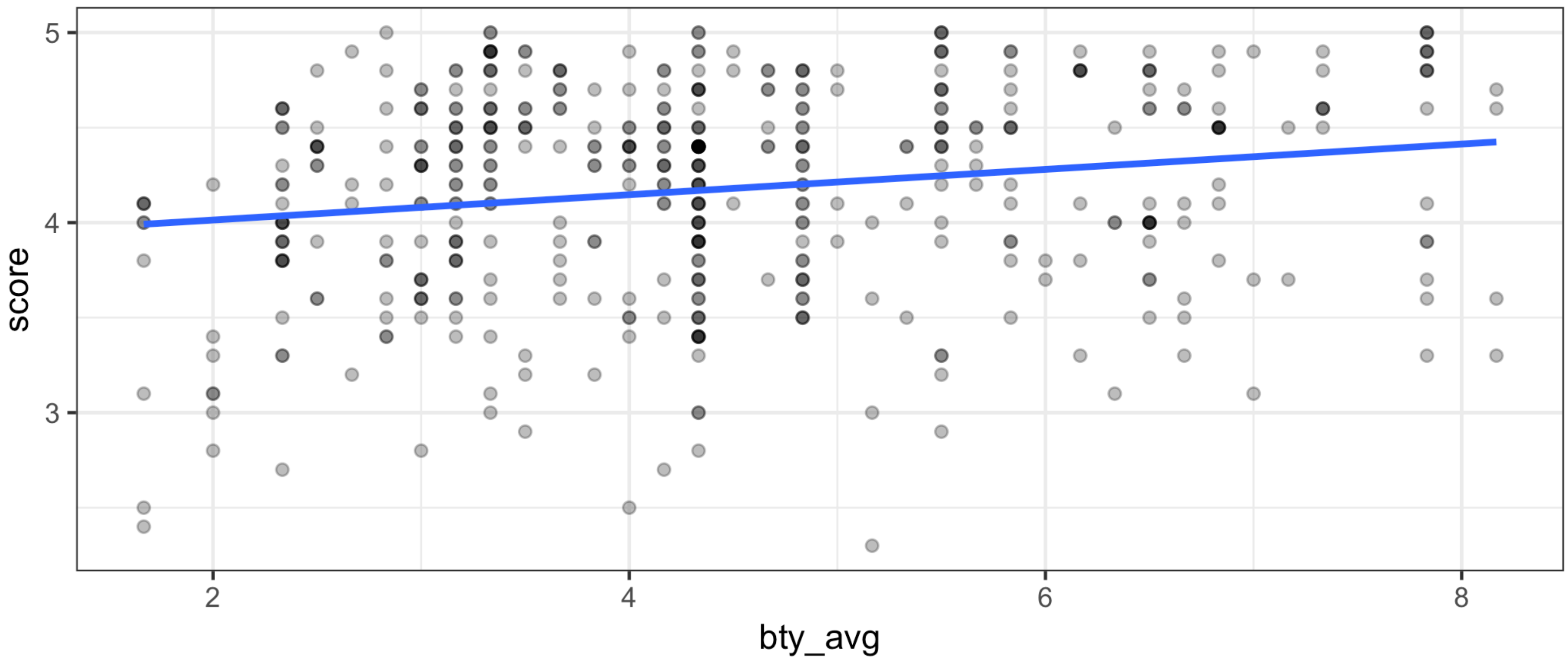
Use some transparency so we can see where there are overlapping points

```
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)
```

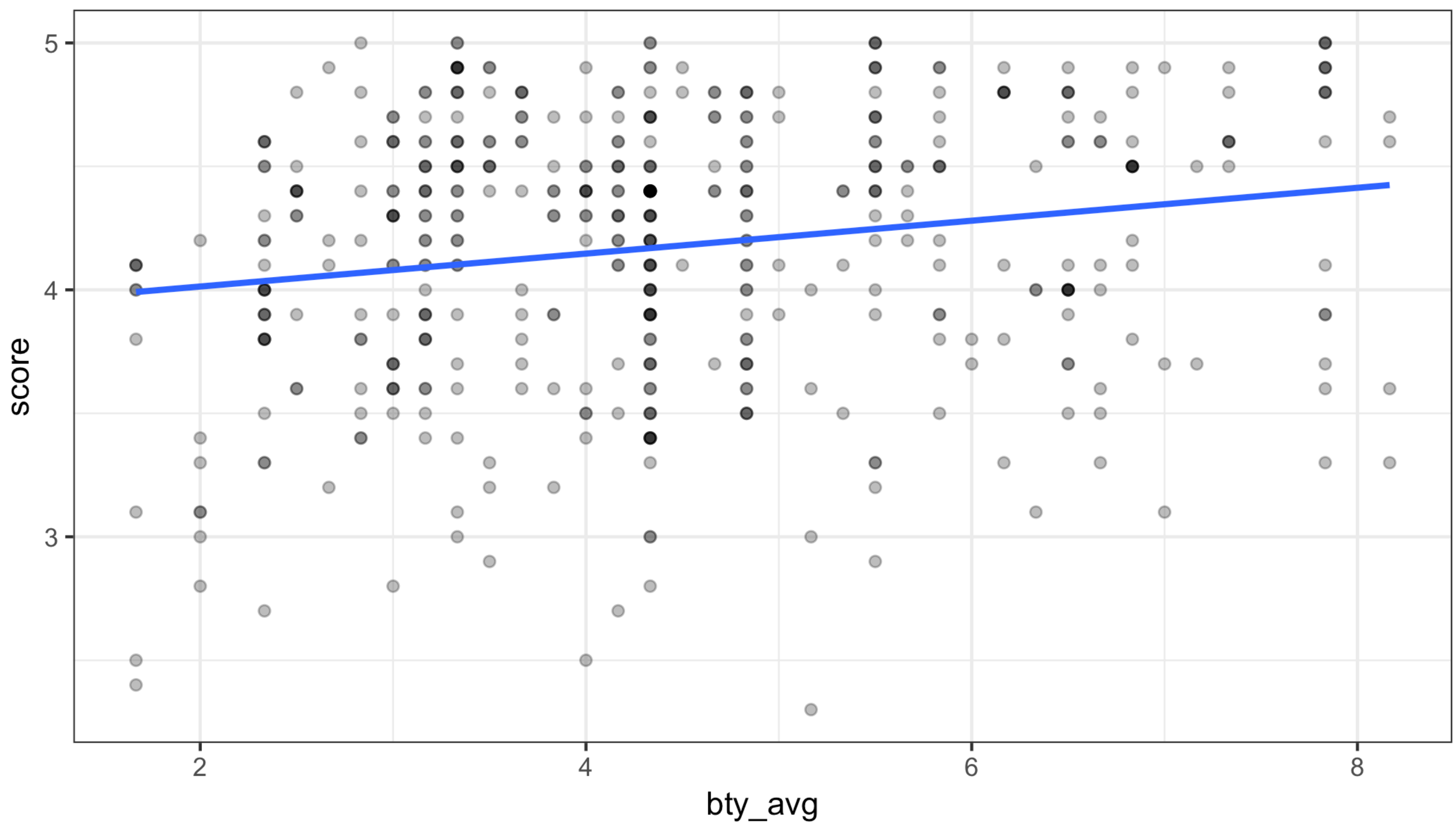


What would the slope be if there was no relationship?

line would be horizontal => Slope = 0

$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$ $H_0: \beta_1 = 0$, $H_a: \beta_1 \neq 0$

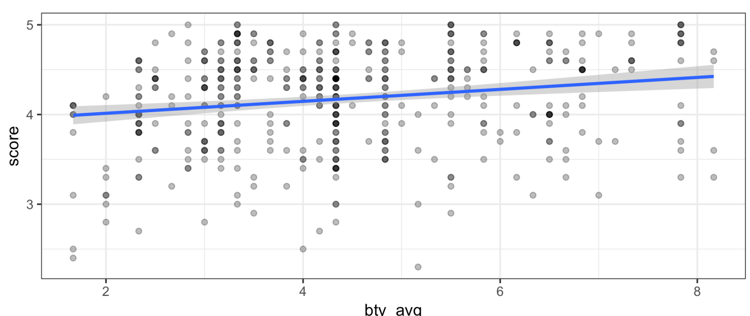
$Score_i = \beta_0 + \beta_1 bty\_Avg_i + \varepsilon_i$



# Confidence interval for the slope

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

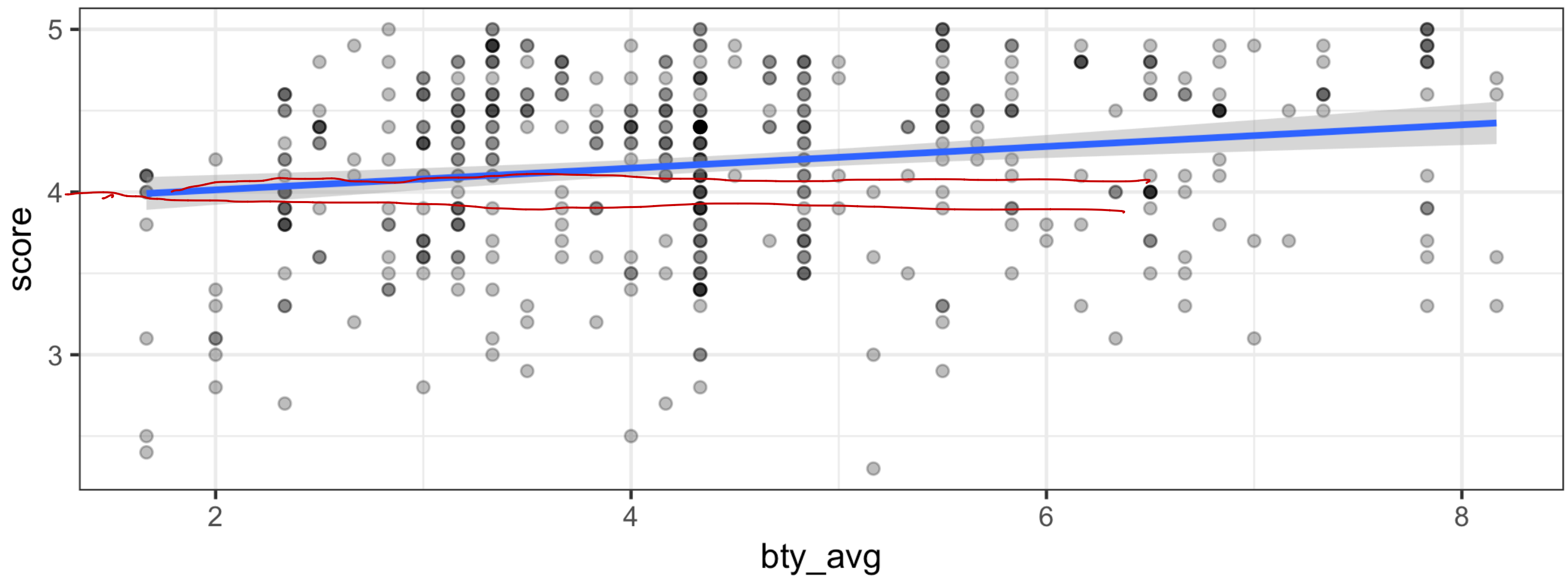
```
ggplot(evals,
       aes(x = bty_avg,
           y = score)) +
  geom_point(alpha = 0.3)  +
  theme_bw() +
  geom_smooth(method = "lm")
```




adds a 95% CI for the regression line by default.

se = FALSE then no CI will be produced.

- 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, but can also be calculated using the bootstrap.




*Does the confidence interval indicate that 0 is a possible value for $\beta_1$ (the parameter for the slope)?*

No. It's not possible to draw a horizontal line at any point of bty_avg and still remain in the shaded area.

# Inference for regression part 2: Hypothesis test for the slope

- Output from the summary command for the estimated regression coefficients gives results for an hypothesis test with hypotheses:

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

produces linear regression.

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

Linear regression of Score on bty_Avg.

```
##               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
```

$\hat{Score}_i = 3.88 + 0.067\, bty\_Avg.$

- The estimate of the slope is 0.06664.
- The `lm()` function, by default, calculates the P-value for regression coefficients 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?*

Since the p-value is very small there is strong evidence against $H_0: \beta_1 = 0$. ∴ It's likely that $\beta_1 \neq 0$

# What other factors might affect course evaluations?

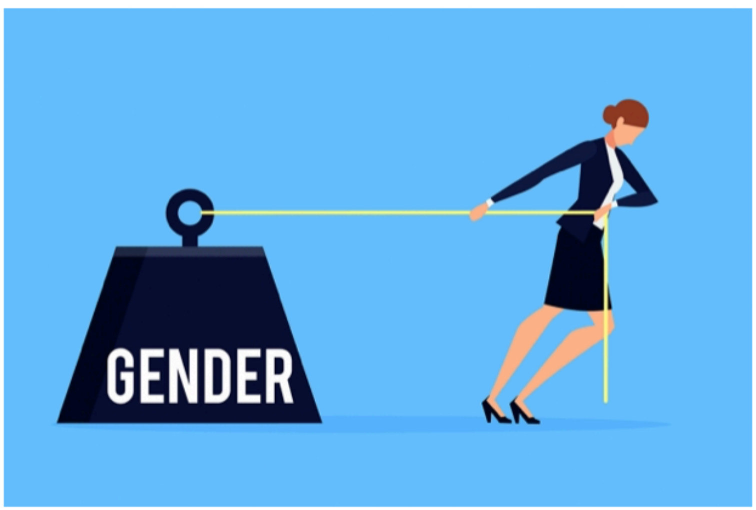
THE CHRONICLE OF HIGHER EDUCATION

☰ SECTIONS | FEATURED: Stories of Student Hunger and Homelessness | Strategies for Gen-Ed Reform



ADVICE

## Why We Must Stop Relying on Student Ratings of Teaching



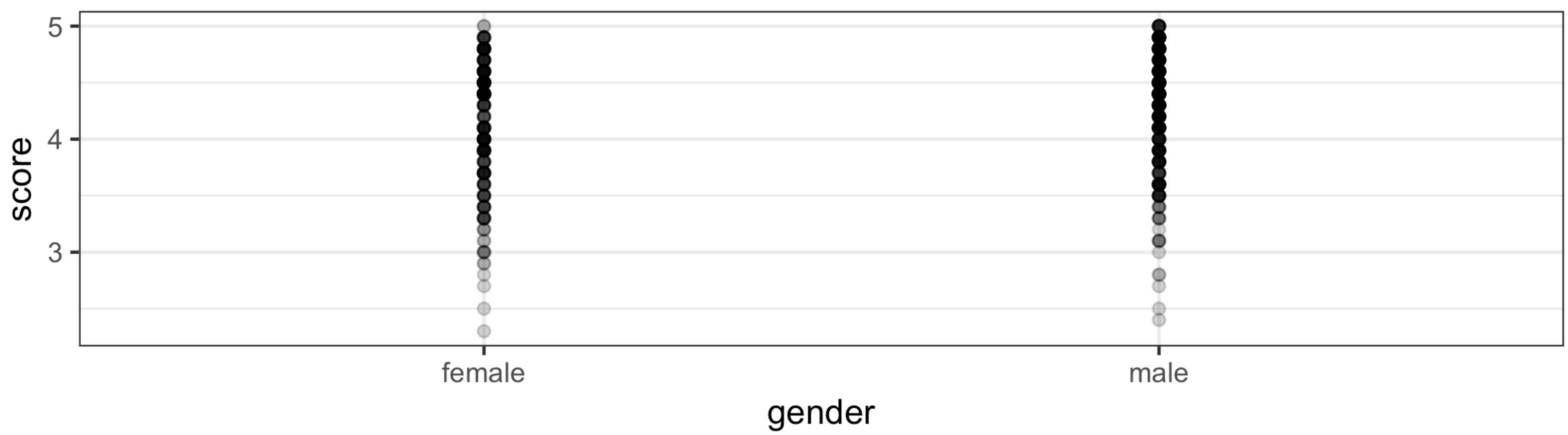

iStock

*By Michelle Falkoff* | APRIL 25, 2018

# Regression when the independent variable is a categorical variable

# Relationship between `score` and `gender`?

```
ggplot(evals, aes(x = gender, y = score)) +
  geom_point(alpha = 1/5) +
  theme_bw()
```




```
evals %>%
  group_by(gender) %>%
  summarise(n = n(), mean = mean(score))
```

```
## # A tibble: 2 x 3
##   gender     n  mean
##   <fct>  <int> <dbl>
## 1 female   195  4.09
## 2 male     268  4.23
```

# Regression with `gender` as the independent variable

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

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

prediction eqn.

$$\widehat{score} = 4.09 + 0.14\, male$$

Interpretation: On average, course evaluation scores for male professors are $0.14$ higher than for female professors.

4.23 − 4.09 = 0.14

males: $\hat{Score} = 4.09 + 0.14 \cdot 1 = 4.09 + 0.14$

females: $\hat{Score} = 4.09 + 0.14 \times 0 = 4.09$

# Regression with `gender` as the independent variable

$$\widehat{score} = 4.09 + 0.14\, male$$

- 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 `male` is 1 for observations for which `gender` is male and 0 otherwise.

$male = \begin{cases} 1 & \text{if prof is male} \\ 0 & \text{o.w} \end{cases}$

- For females,

$$\widehat{score} = 4.09$$

- For males,

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

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

The regression model is

$$score_i = \beta_0 + \beta_1\, male_i + \epsilon_i, i = 1, \ldots, 463$$

where,

$$male_i = \begin{cases} 1 & \text{if } i^{th} gender \text{ is } male \\ 0 & \text{if } i^{th} gender \text{ is } female. \end{cases}$$

We can answer the question with an hypothesis test with hypotheses

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

males: $Score_i = \beta_0 + \beta_1 + \varepsilon_i$

females: $Score_i = \beta_0 + \beta_1 \times 0 + \varepsilon_i = \beta_0 + \varepsilon_i$

$Score_i(\text{males}) - Score_i(\text{females}) = \beta_1$

Even if a difference is statistically significant it does not always follow that the difference is practically significant.

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

```
##                 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?

Average score for males is 4.23

" " females is 4.09

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

diff = 0.14

Is the difference due to chance?

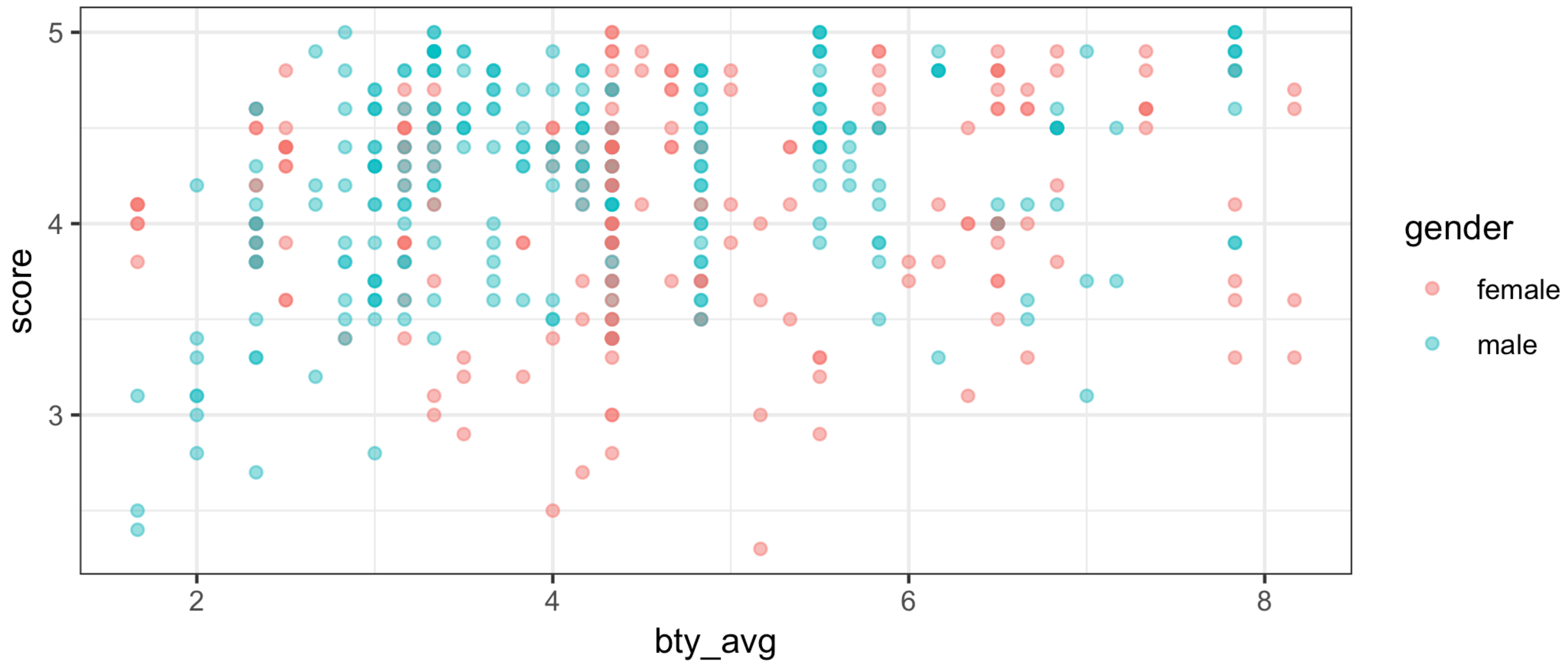
Is the difference stat. significant?

- P-value is very small
- ∴ evidence against $H_0: \beta_1 = 0$
- ∴ $\beta_1$ corresponds to the mean diff. between males and females we have evidence that course rating diff. is not due to chance.

# 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()
```



$male_i = \begin{cases} 1 & \text{if prof is male} \\ 0 & \text{o.w} \end{cases}$

**Model 1:**

$$score_i = \beta_0 + \beta_1\, male_i + \beta_2\, bty\_avg_i + \epsilon_i, i = 1, \ldots, 463$$

Model 1 for male professors:

$male_i = 1$

$$score_i = \beta_0 + \beta_1 + \beta_2\, bty\_avg_i + \epsilon_i, i = 1, \ldots, 463$$

Model 1 for female professors:

$male_i = 0$

$$score_i = \beta_0 + \beta_2\, bty\_avg_i + \epsilon_i, i = 1, \ldots, 463$$

male prof. $Score_i = (\beta_0 + \beta_1) + \beta_2\, bty\_Avg_i + \epsilon_i$

females prof: $Score_i = \beta_0 + \beta_2\, bty\_Avg + \epsilon_i$.

Regression equations have the Same slope but 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
```

$\hat{Score}_i = 3.747 + 0.1724\,gender + 0.0742\,bty\_avg$

prediction equation. or estimated regression line.

For males:

$\hat{Score}_i = 3.747 + 0.1724 + 0.0742\,bty\_Avg.$

For females:

$\hat{Score}_i = 3.747 + 0.0742\,bty\_Avg.$

# Plotting the parallel lines

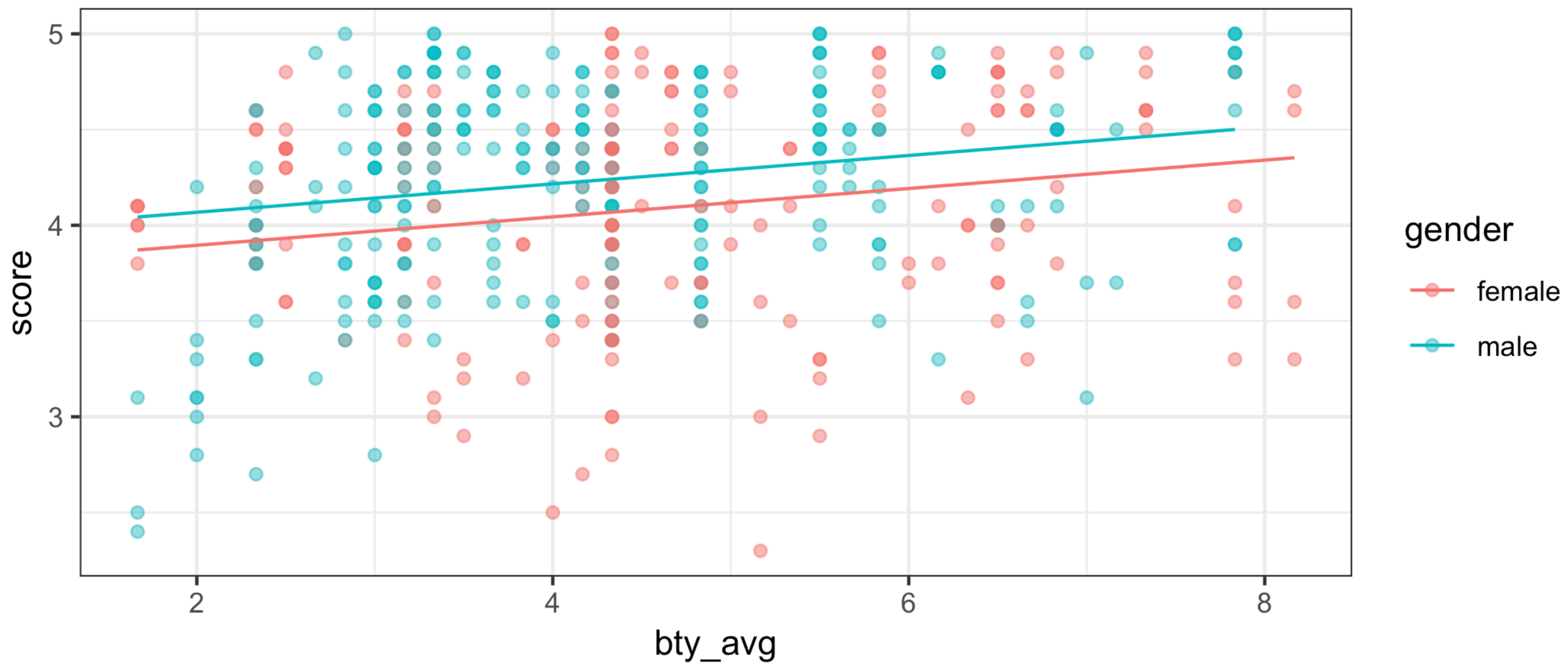
The `augment` function (in the library `broom`) creates a data frame with predicted values (`.fitted`), residuals, etc. for linear model output.

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

```
## # A tibble: 463 x 10
##    score gender bty_avg .fitted .se.fit  .resid    .hat .sigma .cooksd
##  * <dbl> <fct>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>
##  1   4.7 female    5       4.12  0.0383  0.582  0.00524  0.529 2.14e-3
##  2   4.1 female    5       4.12  0.0383 -0.0181 0.00524  0.529 2.07e-6
##  3   3.9 female    5       4.12  0.0383 -0.218  0.00524  0.529 3.00e-4
##  4   4.8 female    5       4.12  0.0383  0.682  0.00524  0.528 2.94e-3
##  5   4.6 male      3       4.14  0.0381  0.458  0.00519  0.529 1.31e-3
##  6   4.3 male      3       4.14  0.0381  0.158  0.00519  0.529 1.56e-4
##  7   2.8 male      3       4.14  0.0381 -1.34   0.00519  0.526 1.13e-2
##  8   4.1 male      3.33    4.17  0.0355 -0.0669 0.00451  0.529 2.43e-5
##  9   3.4 male      3.33    4.17  0.0355 -0.767  0.00451  0.528 3.19e-3
## 10   4.5 female    3.17    3.98  0.0450  0.518  0.00723  0.529 2.35e-3
## # ... with 453 more rows, and 1 more variable: .std.resid <dbl>
```

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))
```




we fit a model where we assumed that slopes are the same. But, are they really the same?

# Lines for each gender that aren't parallel

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

interaction term.

**Model 2:**

$$score_i = \beta_0 + \beta_1\, male + \beta_2\, bty\_avg_i + \beta_3\, (male \times bty\_avg)_i + \epsilon_i$$

male = 1

Model 2 for male professors:

$$score_i = \beta_0 + \beta_1 + \beta_2\, bty\_avg_i + \beta_3\, bty\_avg_i + \epsilon_i$$

$$score_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)\, bty\_avg_i + \epsilon_i$$

male = 0

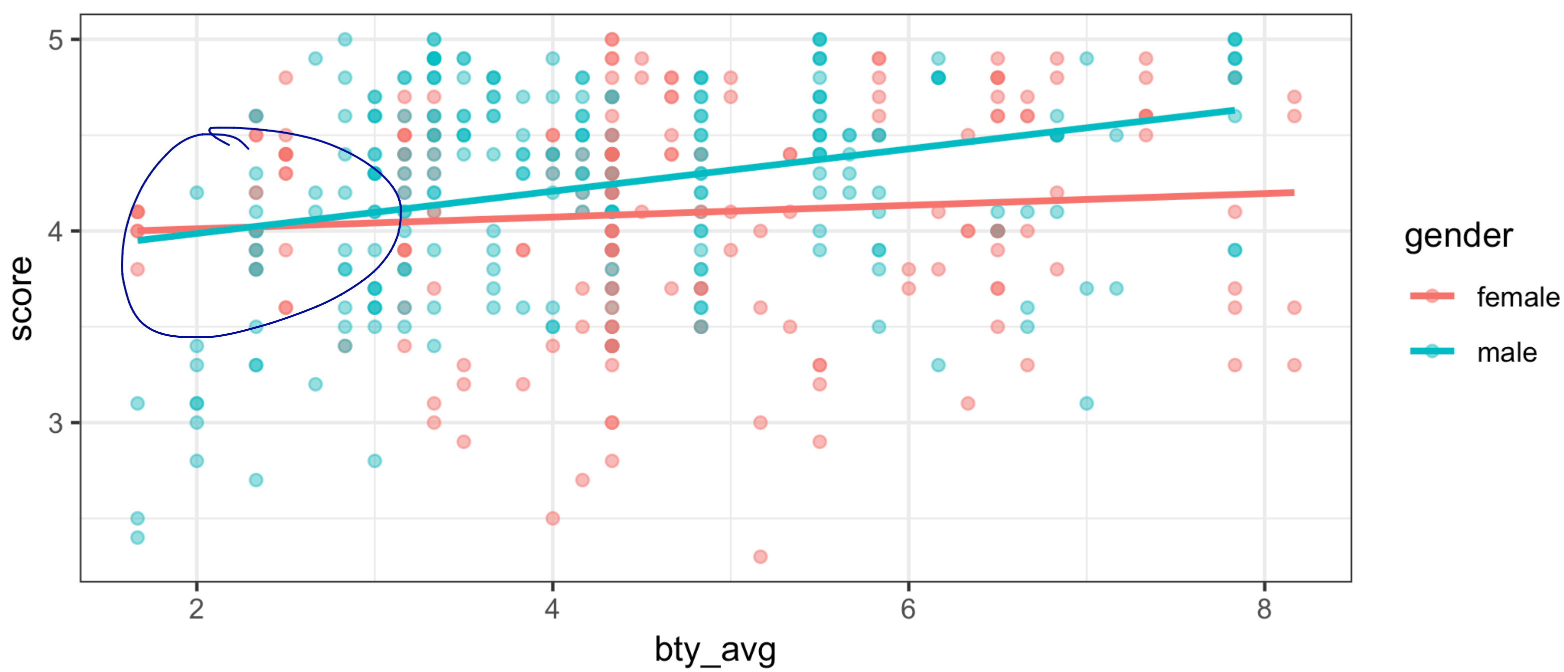
Model 2 for female professors:

$$score_i = \beta_0 + \beta_2\, bty\_avg_i + \epsilon_i$$

when does ⊛ and ⊛ have the same slope? When $\beta_3 = 0$.

# Plot of non-parallel lines

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




When separate models are fit then lines are no longer parallel!

# 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
```

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

$\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$

bty * gendermale.

$\widehat{Score}_i = 3.95 + 0.03 \times$ bty_Avg $- 0.184$ gendermale $+ 0.0796$ bty_Avg $\times$ gendermale.

For male profs: gendermale = 1

$\widehat{Score}_i = 3.95 + 0.03$ bty_Avg $- 0.184 + 0.079$ bty_Avg

$\widehat{Score}_i = 3.95 + 0.03 \times$ bty_Avg. (gendermale = 0).

the p-value for $H_0: \beta_3 = 0$ is small $\therefore$ evidence against $H_0$.

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

Model:

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

*What would be appropriate hypotheses to test?*

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

Caution: just because plot shows non-parallel lines does not imply test will reject $H_0: \beta_3 = 0$

*What do you conclude?*

∵ the p-value is very small we have enough evidence to reject $H_0$. ∴ the relationship between score and bty Avg is different for male and female professors.

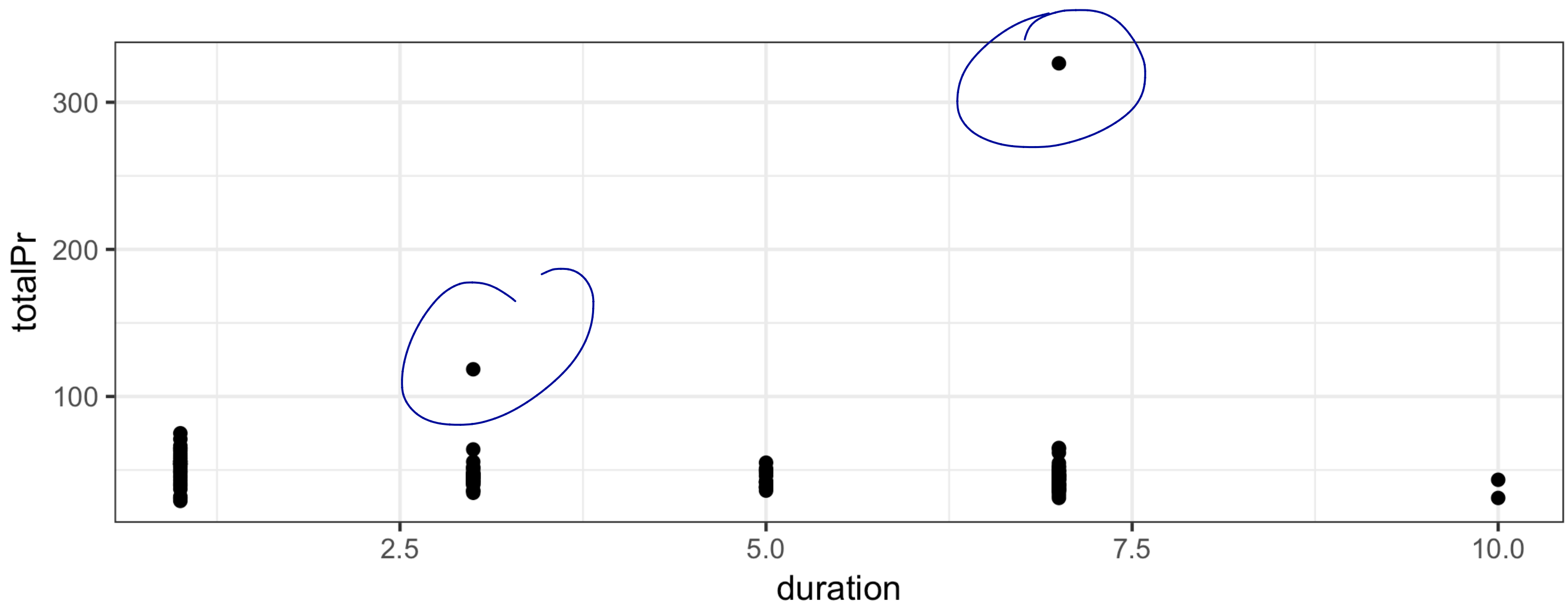
# Example: eBay auctions of *Mario Kart*

- 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        <fct> new, used, new, new, new, new, used, new, used, use...
## $ 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      <fct> standard, firstClass, firstClass, standard, media, ...
## $ sellerRate  <int> 1580, 365, 998, 7, 820, 270144, 7284, 4858, 27, 201...
## $ stockPhoto  <fct> yes, yes, no, yes, yes, yes, yes, yes, yes, no, yes...
## $ wheels      <int> 1, 1, 1, 1, 2, 0, 0, 2, 1, 1, 2, 2, 2, 2, 1, 0, 1, ...
## $ title       <fct> ~~ Wii MARIO KART &amp; WHEEL ~ NINTENDO Wii ~ BRAN...
```

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

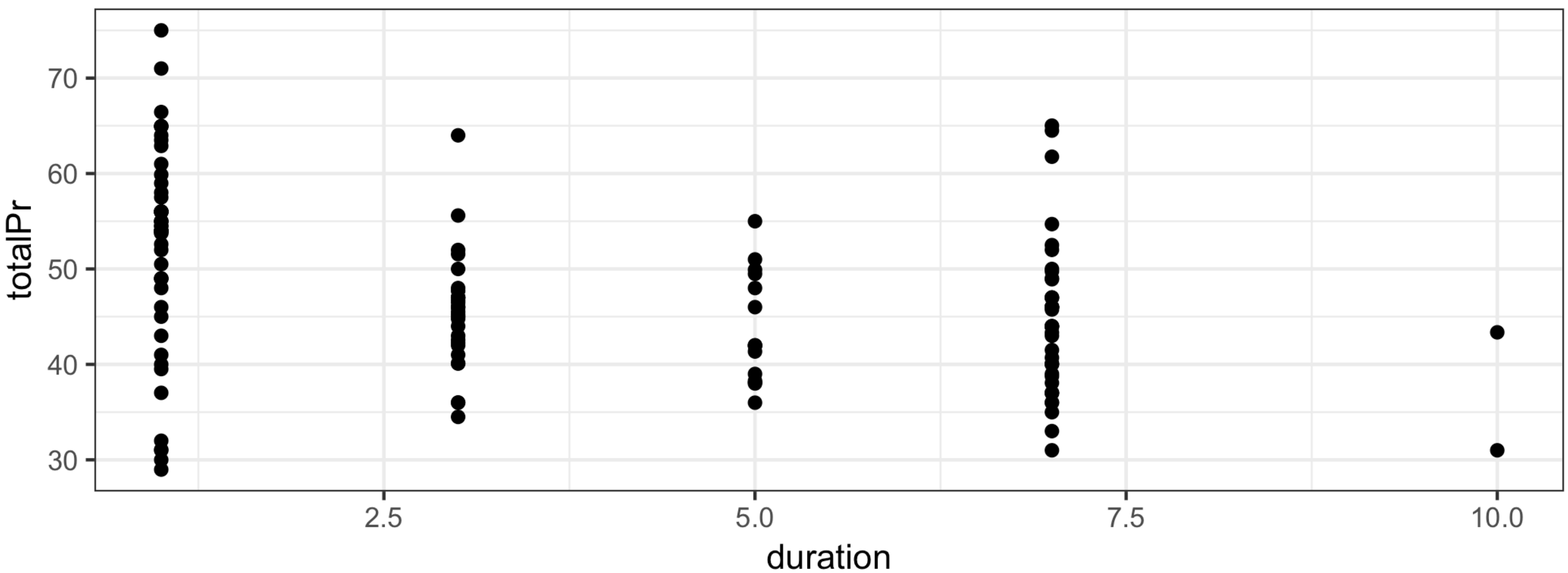


# What should we do with the two outlying values of `totalPr`?

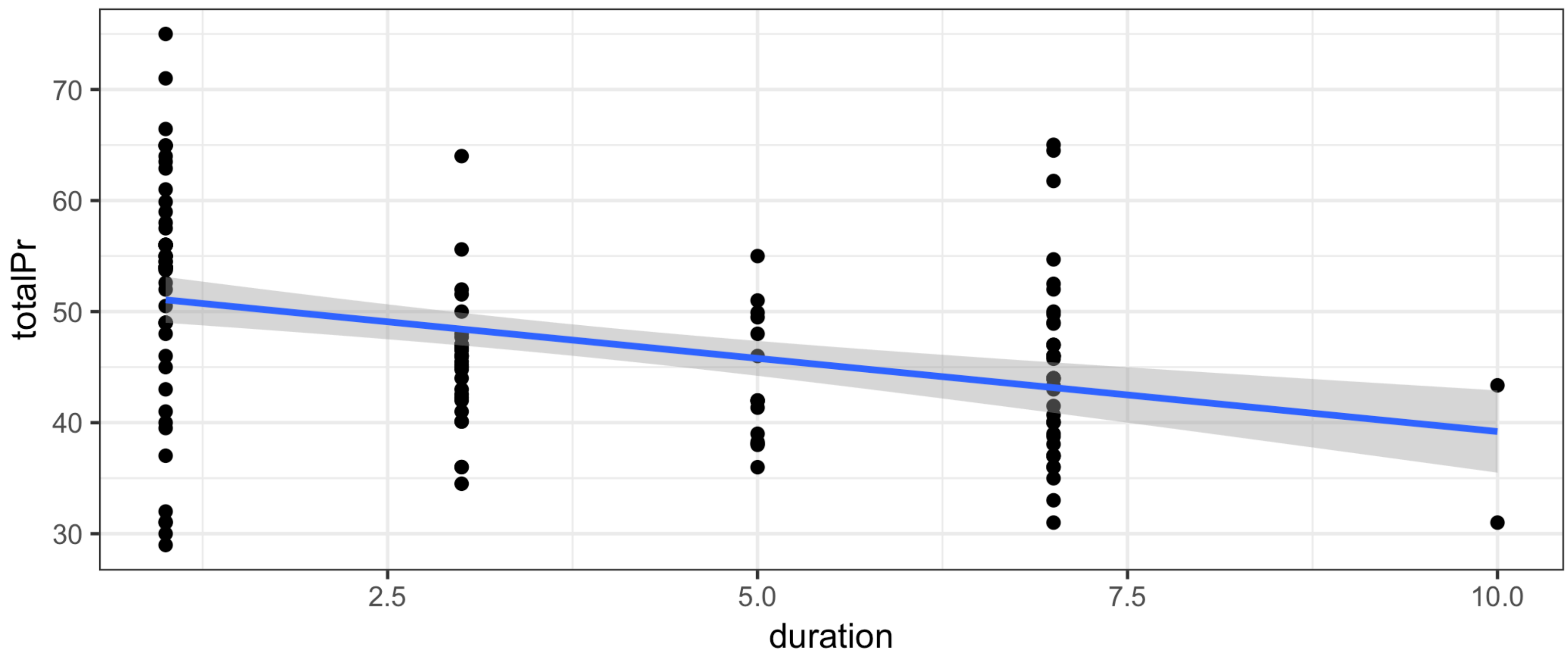
- Remove outliers only if there is a good reason.
- In these two auctions, and only these two auctions, the game was sold with other items.

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

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



```r
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?* No.

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
```

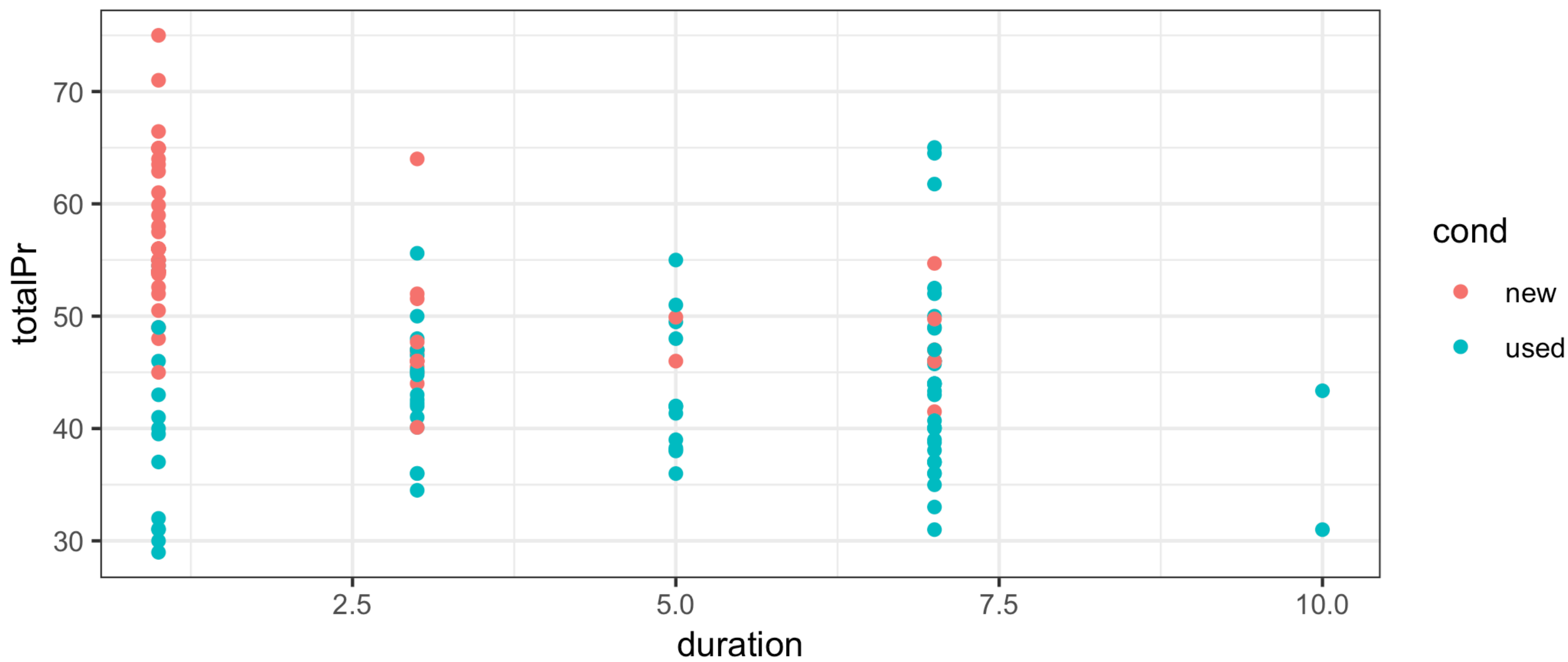
```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 52.373584  1.2607560 41.541411 3.010309e-80
## duration    -1.317156  0.2769021 -4.756756 4.866701e-06
```

We have strong evidence that the slope is not 0.

There must be something else affecting the relationship ...

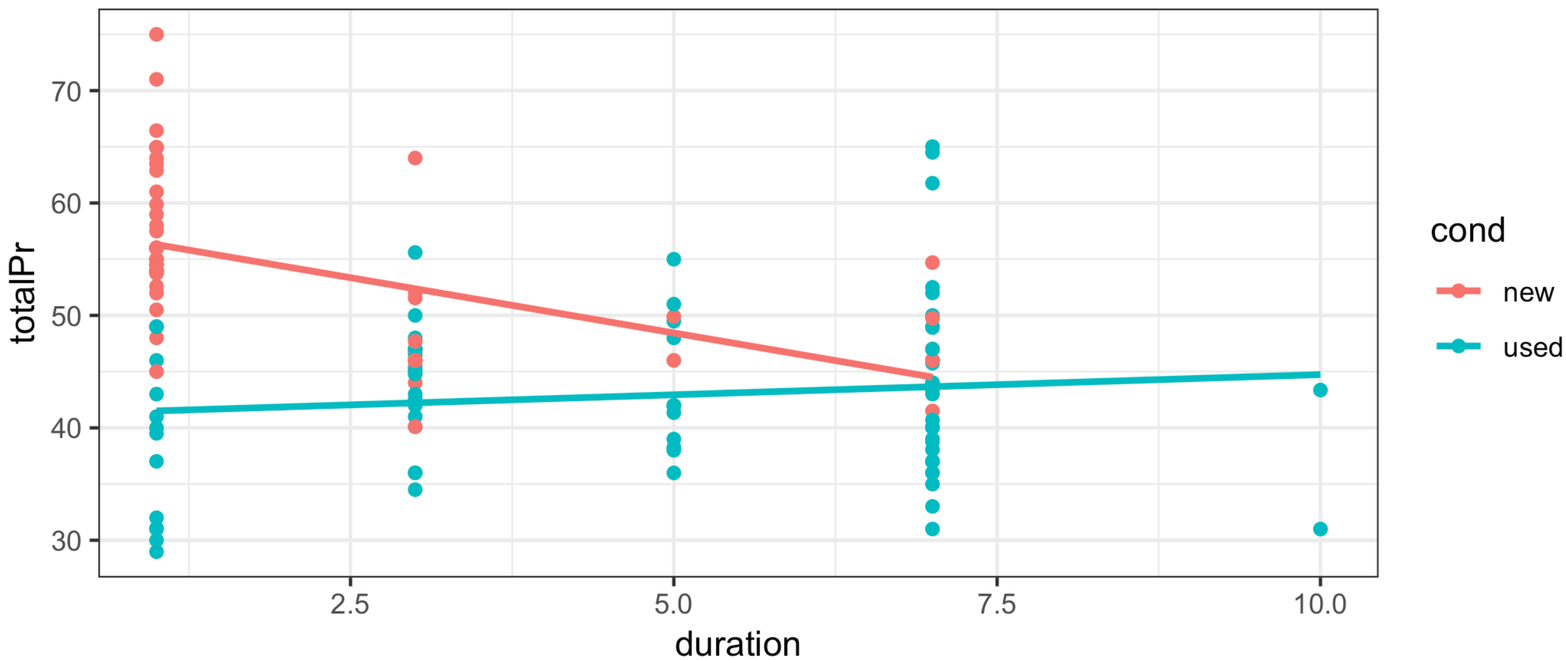
Consider the role of `cond`.
`cond` is a categorical variable for the game's condition, either `new` or `used`.

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



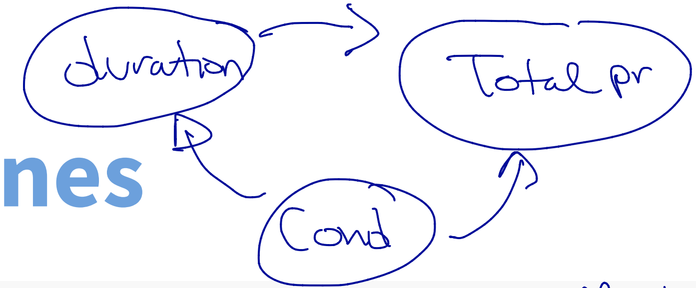

New games, which are more desirable, were mostly sold in one-day auctions.

```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




Example of Confounding.

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

all the data

```
##              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
```

```
marioKart2_used <- marioKart2 %>% filter(cond == "used")
summary(lm(totalPr ~ duration, data = marioKart2_used))$coefficients
```

Used games.

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 41.1463022  1.7924487 22.955358 5.976630e-37
## duration     0.3589676  0.3329894  1.078015 2.842669e-01
```

```
marioKart2_new <- marioKart2 %>% filter(cond == "new")
summary(lm(totalPr ~ duration, data = marioKart2_new))$coefficients
```

new games

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 58.268226  1.2497467 46.624029 4.353419e-47
## duration    -1.965595  0.4104444 -4.788944 1.233340e-05
```

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

model with interaction term.

```
##                       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
```

plug in.

Used = { 1 if game used; 0 if game new }

then you will get lines above.

leave as exercise.

# 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
```

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

# Describing Two-Way Tables

- The (2x2) *contingency table* below shows counts of patients being treated for kidney stones.

```
tab <- table(kidney_stones$outcome,
             kidney_stones$treatment, deparse.level = 2)
addmargins(tab)
```

```
##                    kidney_stones$treatment
## kidney_stones$outcome   A   B Sum
##               failure  77  61 138
##               success 273 289 562
##               Sum     350 350 700
```

adds Sum of rows and column).

Contingency table.

- Proportion of observations in each cell of contingency table.

```
prop.table(tab)
```

```
##                    kidney_stones$treatment
## kidney_stones$outcome          A          B
##               failure 0.11000000 0.08714286
##               success 0.39000000 0.41285714
```

77/700 61/700 289/700 273/700

- **Joint, marginal, and conditional distributions**.

```
addmargins(prop.table(tab))
```

```
##                    kidney_stones$treatment
## kidney_stones$outcome          A          B        Sum
##               failure 0.11000000 0.08714286 0.19714286
##               success 0.39000000 0.41285714 0.80285714
##               Sum     0.50000000 0.50000000 1.00000000
```

marginal distribution of outcome.

marginal distribution of treatment.

The conditional distribution of failure given treatment is:

Among subjects that recieved treatment A what proportion failed?

77/350 = 0.11/0.50

what prop. recieveing trt. B failed

0.087/0.50 = 61/350.

# 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.

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

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

What percentage of successes were Treatment A? previous slide.

# 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)) #%>%
```

```
## # A tibble: 8 x 5
## # Groups:   size, treatment [4]
##   size  treatment outcome     n per_success
##   <chr> <chr>     <chr>   <int>       <dbl>
## 1 large A         failure    71      0.270
## 2 large A         success   192      0.730
## 3 large B         failure    25      0.312
## 4 large B         success    55      0.688
## 5 small A         failure     6      0.0690
## 6 small A         success    81      0.931
## 7 small B         failure    36      0.133
## 8 small B         success   234      0.867
```

```
  #filter(outcome=="success")
```

Column percentages (conditional distribution of success given treatment):

```
prop.table(table(kidney_stones$outcome, kidney_stones$treatment), margin = 2)
```

```
##
##                   A         B
##   failure 0.2200000 0.1742857
##   success 0.7800000 0.8257143
```

overall. Success of each treatment (i.e., Conditional distribution)

```
large <- kidney_stones %>% filter(size == "large")
prop.table(table(large$outcome, large$treatment),margin = 2)
```

```
##
##                  A        B
##   failure 0.269962 0.312500
##   success 0.730038 0.687500
```

when we take size of kidney stone into account then treatment A is better.

```
small <- kidney_stones %>% filter(size == "small")
prop.table(table( small$outcome, small$treatment), margin = 2)
```

```
##
##                    A          B
##   failure 0.06896552 0.13333333
##   success 0.93103448 0.86666667
```

This is another example of Simpson's paradox.

*Which treatment is better?*

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

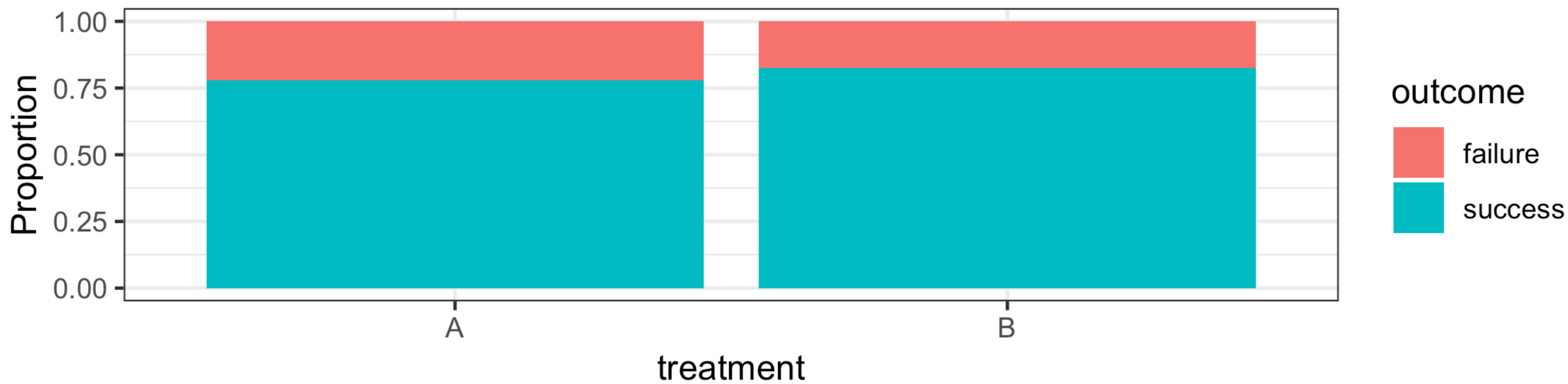
## Moral of the story:

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

Slides 46-52 will be covered next class ...

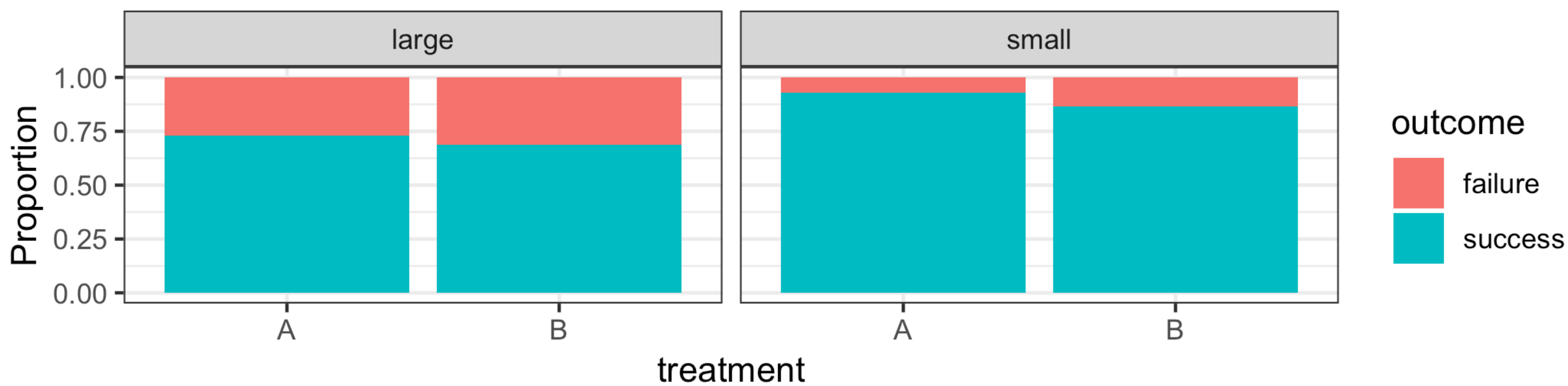
Visualizing the kidney stone data: treatment and outcome

```
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.

# 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.

# 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 confounding?

- 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?*

# Examples of confounding?

- 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?*

# Examples of confounding?

- 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*.

# 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.

# 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*).

# 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.

# 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.

# How can confounding be avoided?

- In a randomized experiment, if there is a difference in our measurement of interest, we *may* be able to 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.

# How can confounding be avoided?

- In a randomized experiment, if there is a difference in our measurement of interest, we *may* be able to 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.

# How can confounding be avoided?

- In a randomized experiment, if there is a difference in our measurement of interest, we *may* be able to 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, it would be considered unethical to randomly assign people to smoke marijuana.