STA130H1F

Class #10 - Influence of other variables

Prof. Nathalie Moon

2018-11-26

Today

Big idea:

Examining the effect 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

Predict median house prices

- Median house price for each census tract in Boston (1976)

neighbourhood

```r
library(MASS)
glimpse(Boston)
```

```
## Observations: 506
## Variables: 14
## $ crim    <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06905, 0.02985, ...
## $ zn      <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12.5, 12.5,...
## $ indus   <dbl> 2.31, 7.07, 7.07, 2.18, 2.18, 2.18, 7.87, 7.87, 7.87, ...
## $ chas    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ nox     <dbl> 0.538, 0.469, 0.469, 0.458, 0.458, 0.458, 0.524, 0.524...
## $ rm      <dbl> 6.575, 6.421, 7.185, 6.998, 7.147, 6.430, 6.012, 6.172...
## $ age     <dbl> 65.2, 78.9, 61.1, 45.8, 54.2, 58.7, 66.6, 96.1, 100.0,...
## $ dis     <dbl> 4.0900, 4.9671, 4.9671, 6.0622, 6.0622, 6.0622, 5.5605...
## $ rad     <int> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, ...
## $ tax     <dbl> 296, 242, 242, 222, 222, 222, 311, 311, 311, 311, 311,...
## $ ptratio <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15.2, ...
## $ black   <dbl> 396.90, 396.90, 392.83, 394.63, 396.90, 394.12, 395.60...
## $ lstat   <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19.15, 29.9...
## $ medv    <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.1, 16.5, ...
```

We want to predict the median house price in each census tract

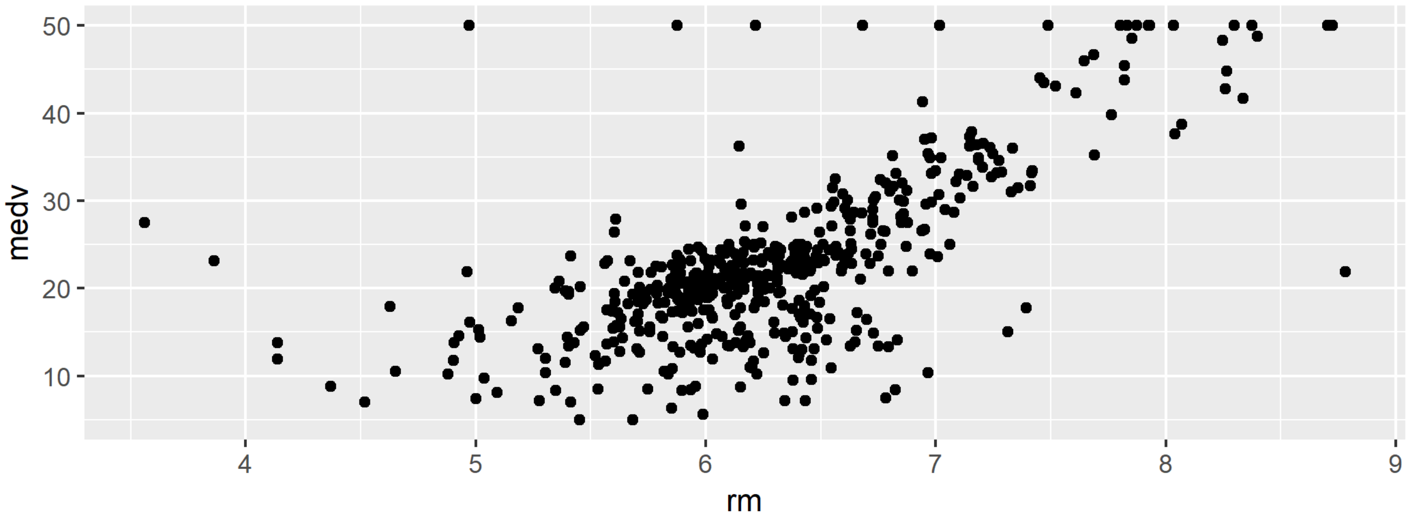
Predict median house prices

For each of 506 census tracts, we have:

- `crim`: per capita crime rate
- `indus`: proportion of non-retail business acres
- `chas`: Charles river dummy variable (=1 if tract bounds river, 0 otherwise)
- `rm`: average number of rooms per dwelling (predictor)
- `age`: proportion of owner-occupied units built prior to 1940
- `rad`: index of accessibility to radial highways
- `ptratio`: pupil-teacher by town
- `lstat`: percentage of low income residents
- `medv`: median value of owner-occupied homes (in $1000s) (outcome to predict)
- ...

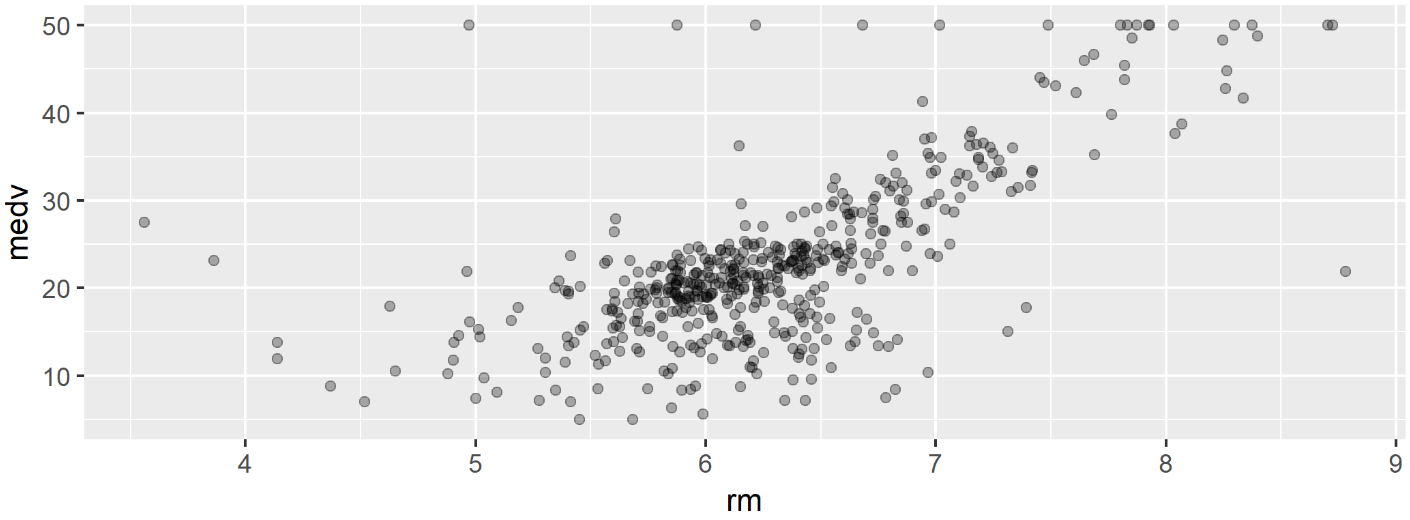
Relationship between median price and average number of rooms

```
Boston %>% ggplot(aes(x=rm, y=medv)) + geom_point()
```



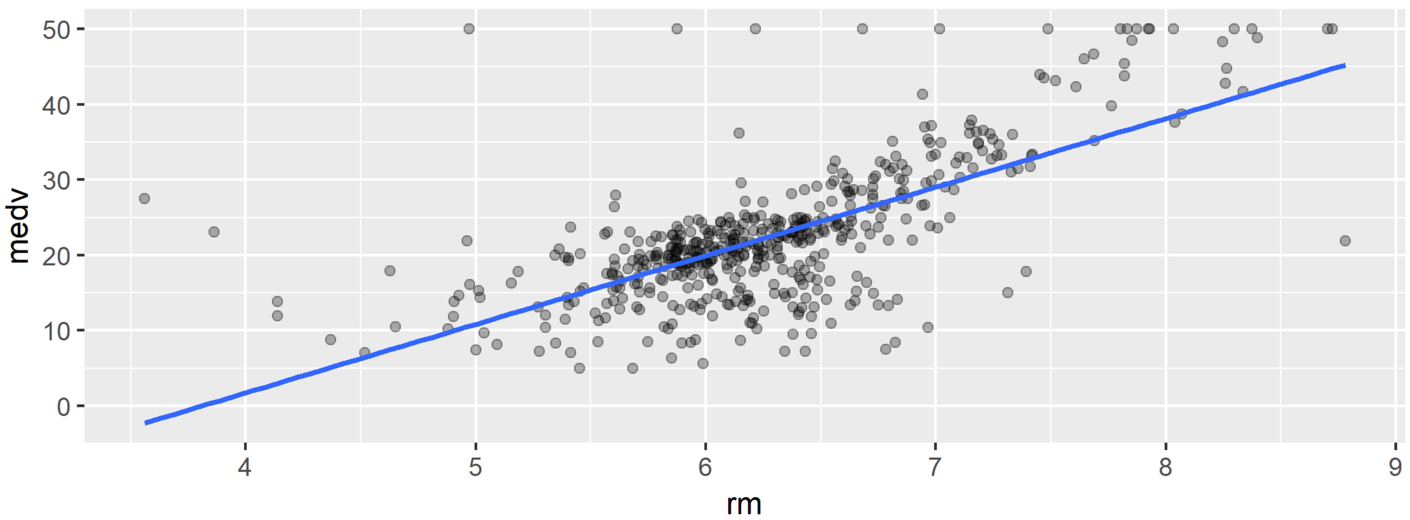
Relationship between median price and average number of rooms

```
Boston %>% ggplot(aes(x=rm, y=medv)) + geom_point(alpha=0.3)
```

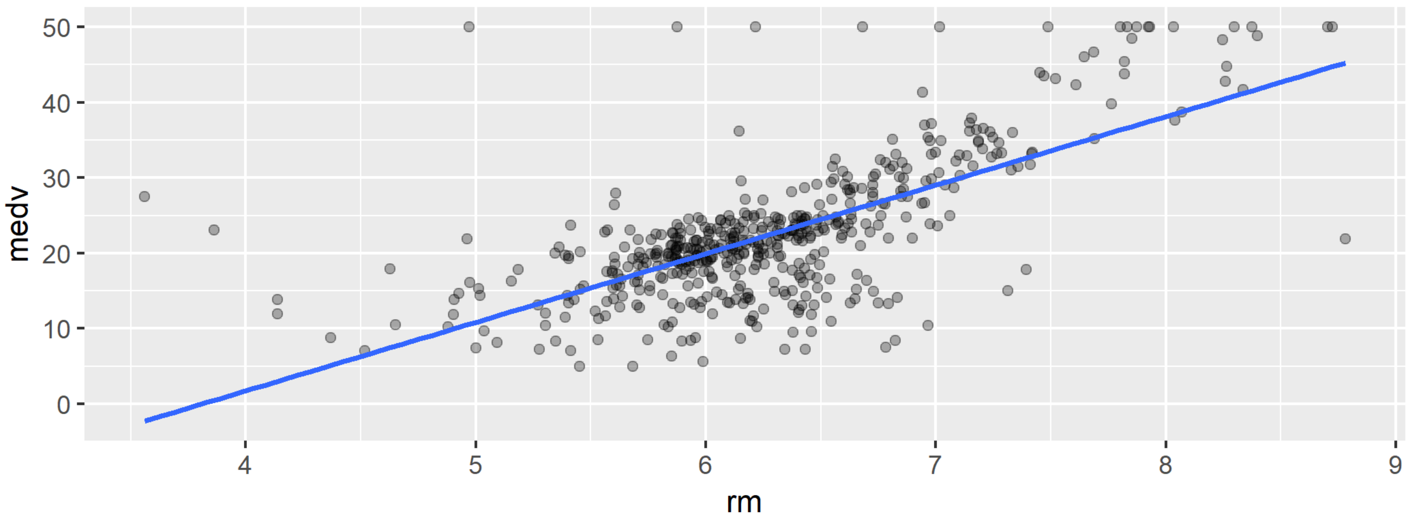


Relationship between median price and average number of rooms

```
Boston %>% ggplot(aes(x=rm, y=medv)) + geom_point(alpha=0.3) +
  geom_smooth(method="lm", se=FALSE)
```

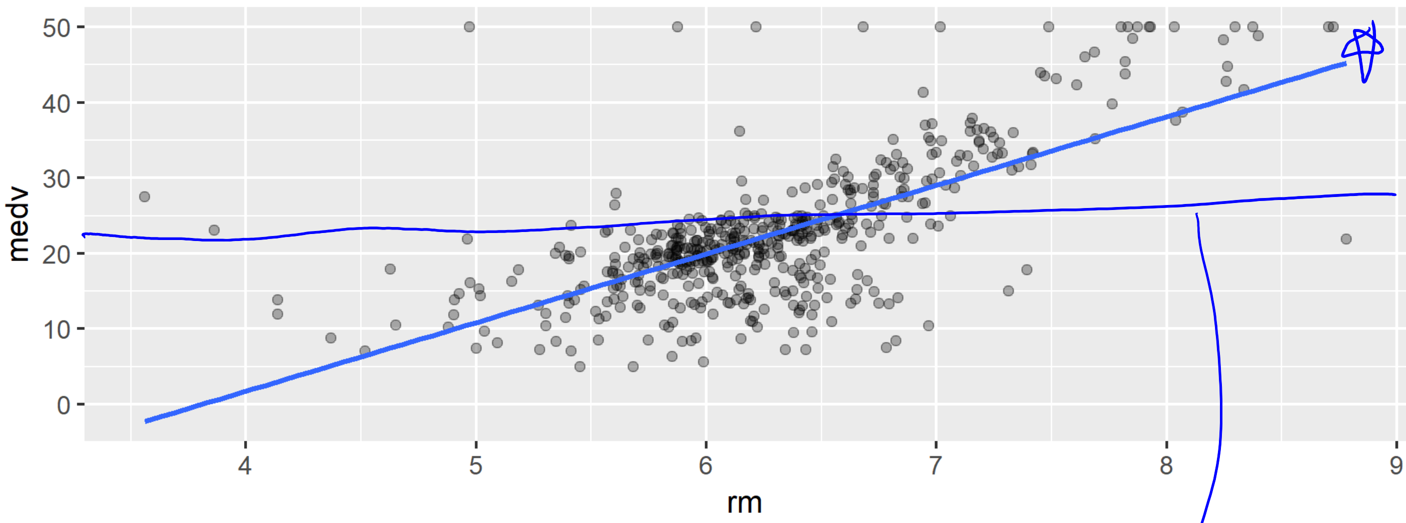


Is the association real or just due to chance?



If there is truly no association, what would the (true) line look like?

Is the association real or just due to chance?



If there is truly no association, what would the (true) line look like?

in $1000s

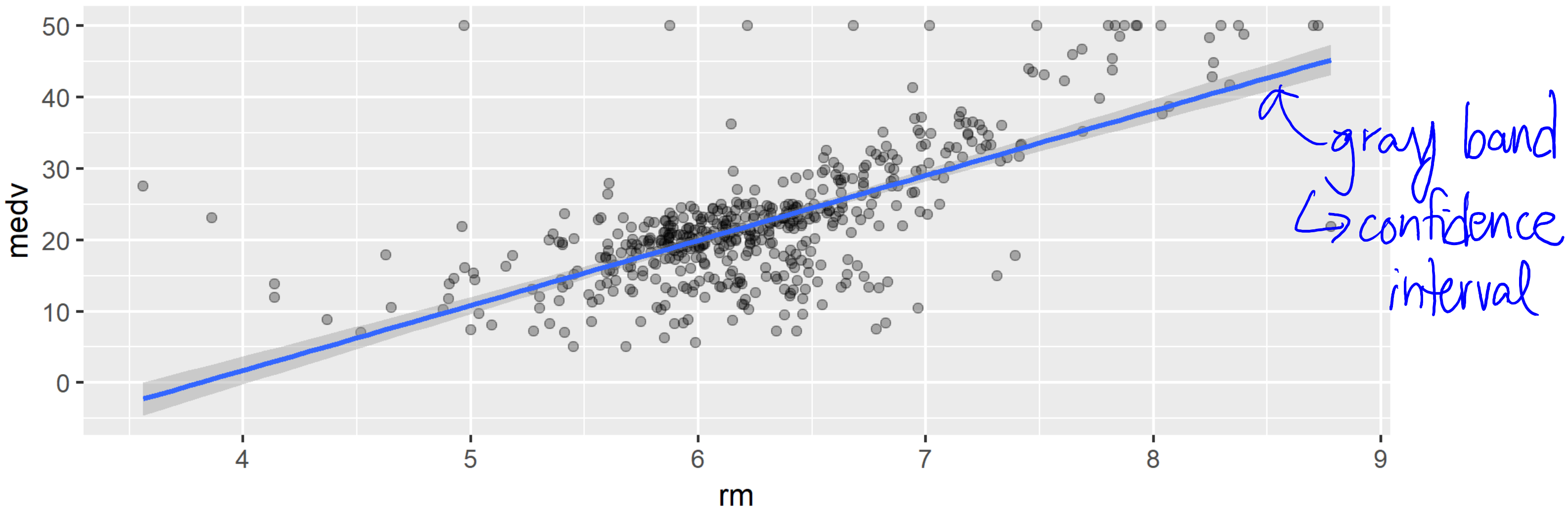
```
mean(Boston$medv)
```

```
## [1] 22.53281
```

horizontal line with y = average of median house prices

Confidence interval for the fitted line

```r
Boston %>% ggplot(aes(x=rm, y=medv)) + geom_point(alpha=0.3) +
  geom_smooth(method="lm") # by default, se=TRUE
```



Confidence interval for fitted line

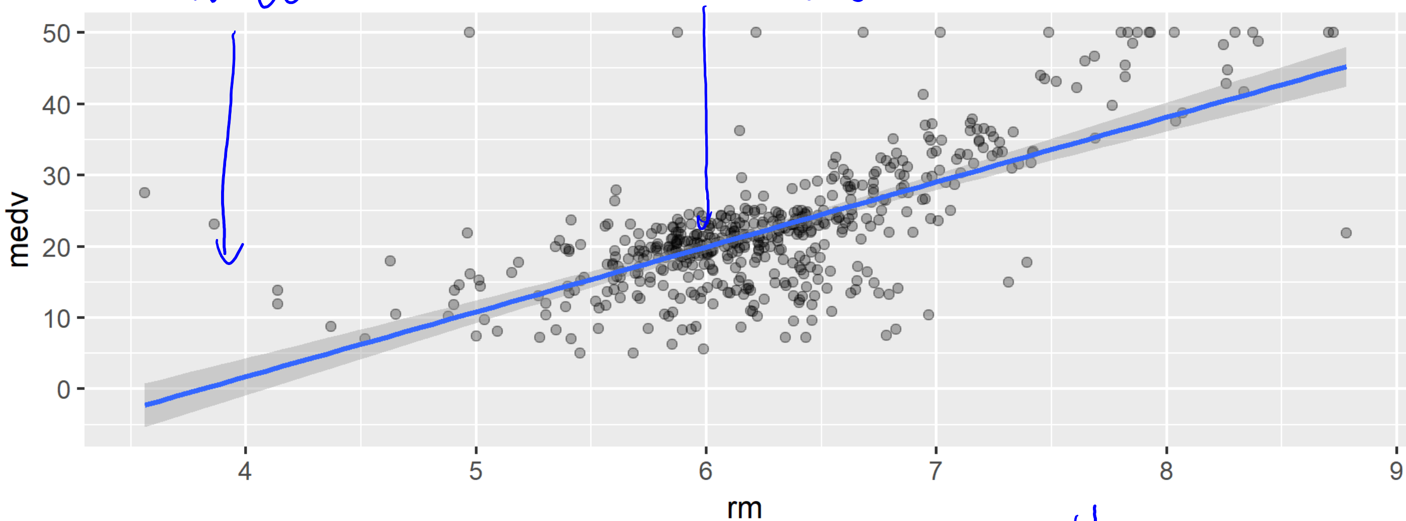
- Gray shaded area around the fitted regression line is a 95% confidence interval for the line
  - `geom_smooth(method="lm", level="0.90")` for 90% confidence interval, etc.
  - Default is `level=0.95`
- The confidence interval plotted is based on the following assumptions:
  - all observations are independent
  - error terms have a symmetric, bell-shaped distribution
- You can also get a confidence interval using the boostrap approach

Confidence interval for fitted line

```r
Boston %>% ggplot(aes(x=rm, y=medv)) + geom_point(alpha=0.3) +
  geom_smooth(method="lm", level=0.99) # by default, se=TRUE
```

wider

much narrower.



No.

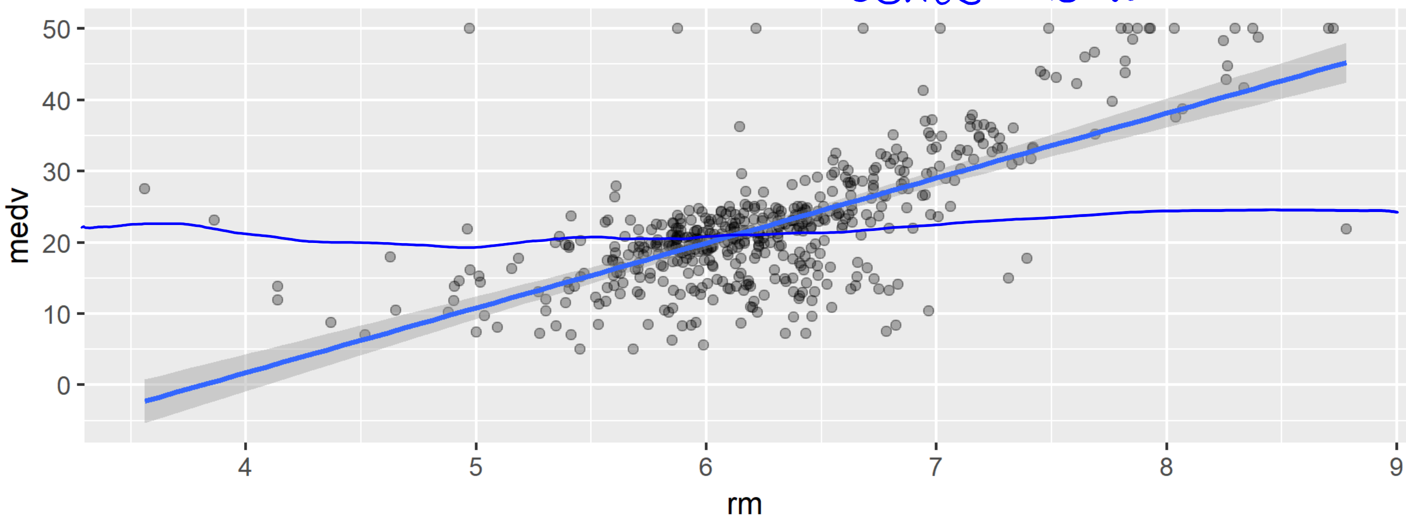
Is the confidence interval always the same width? Why?

It is easier to get more accurate estimates close to the mean, while it is harder to estimate the extremes due to leso data for these.

Confidence interval for fitted line

```
Boston %>% ggplot(aes(x=rm, y=medv)) + geom_point(alpha=0.3) +
  geom_smooth(method="lm", level=0.99) # by default, se=TRUE
```

99% confidence interval.



Does the confidence interval indicate that there is an association?

↳ Yes, because the horizontal line does not lie within the confidence band.

β_0: beta-naught (↳ zero).

Inference for simple linear regression

What is the equation for the linear regression model we've fit?

$$y = \beta_0 + \beta_1 x + \varepsilon$$

(median price) (avg # rooms) (error)

Inference for simple linear regression

What is the equation for the linear regression model we've fit?

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

where y_i is the median house price in census tract i and x_{i1} is the average number of rooms for houses in census tract i

How can we write H_0 and H_A to test if there is an association between the median house price and the average number of rooms?

$H_0: \beta_1 = 0$ (no association between x and y)

$H_A: \beta_1 \neq 0$ (there is an association between x and y)

Using R for hypothesis testing

```
summary(lm(medv ~ rm, data=Boston))$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -34.670621  2.6498030 -13.08423 6.950229e-34
## rm            9.102109  0.4190266  21.72203 2.487229e-74
```

R gives p-values for hypothesis test of the form:

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0$$

What is the p-value for this test?

$2.5 \times 10^{-74} \approx 0$.

p-value for testing $H_0: \beta_0 = 0$ vs $H_A: \beta_0 \neq 0$.

but we generally aren't interested in this

Using R for hypothesis testing

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -34.670621  2.6498030 -13.08423 6.950229e-34
## rm            9.102109  0.4190266  21.72203 2.487229e-74
```

The estimate of the slope $\hat{\beta}_1$ is:

The p-value for testing $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$ is:

To calculate the p-value, the lm function assumes that observations are independent and that the errors have a symmetric, bell-shaped distribution.

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

↳ Yes, because the pvalue is very close to 0.

Which statements are true?

```
summary(lm(medv ~ rm, data=Boston))$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -34.670621  2.6498030 -13.08423 6.950229e-34
## rm            9.102109  0.4190266  21.72203 2.487229e-74
```

```
summary(lm(medv ~ rm, data=Boston))$r.squared
```

```
## [1] 0.4835255
```

✓ (a) An increase of 1 in the average number of rooms is associated with an increase of $9,100 in the median price

✗ (b) Approximately 48% of the median house prices can be predicted by the average number of rooms per house in each census tract

→ x is the avg # rooms.

✓ (c) The linear model is $\hat{y} = -34.7 + 9.1x$, where y is price in $1000s

✓ (d) The median price of houses in census tracts with an average of 0 rooms per house is -$34,6706.

↳ but meaningless...

What other factors might affect house prices?

- Many other variables in our dataset
- Let's look to see if house prices are affected by the **number of crimes per capita** in each census tract

What other factors might affect house prices?

- Many other variables in our dataset
- Let's look to see if house prices are affected by the **number of crimes per capita** in each census tract

```
summary(Boston$crim)
```

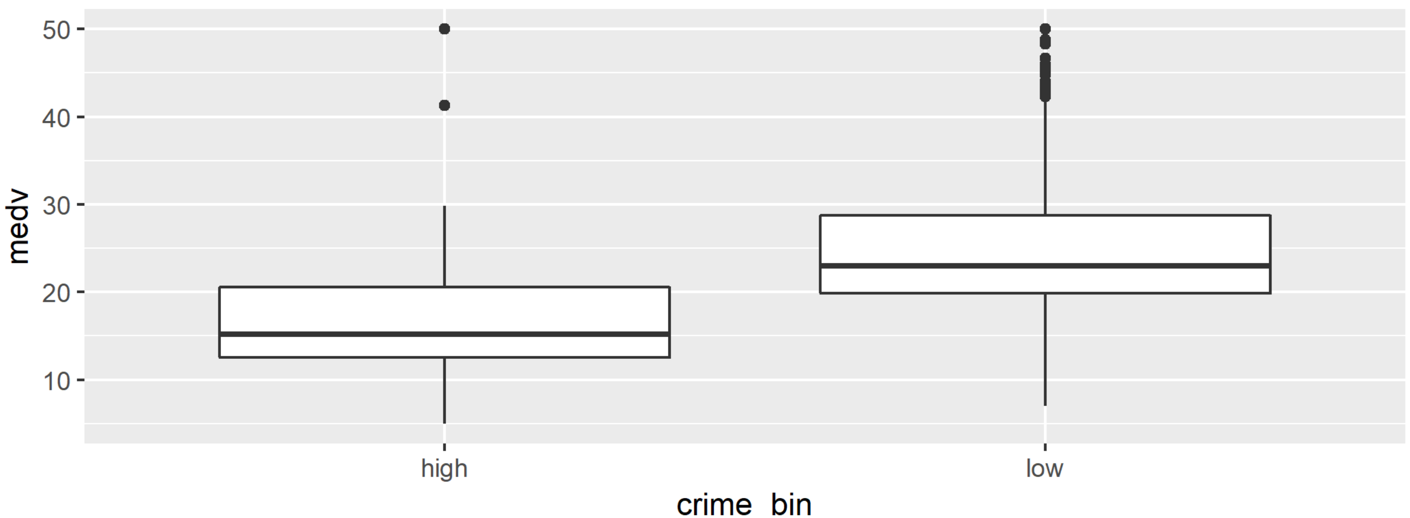
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

We'll define a new variable `crime_bin` to make it easier to visualize the relationship

```
Boston <- Boston %>%
  mutate(crime_bin = ifelse(crim < 1, "low", "high"))
```

Relationship between crime (low/high) and price

```
Boston %>% ggplot(aes(x=crime_bin, y=medv, group=crime_bin)) +
  geom_boxplot()
```



Regression with `crime_bin` as predictor

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  17.61379  0.6433695 27.377413 8.806068e-102
## crime_binlow  7.49705  0.7942673  9.438951  1.389388e-19
```

The fitted regression equation is:

$$\widehat{median\_price} = 17.6 + 7.5 \times low\_crime$$

where y_i is the median price in thousands for tract i and $I(low\_crime)$ is 1 if it is a low crime area and 0 otherwise

How should we interpret the slope β_1 ?

β_1 is the change in med. house price associated with going from a high crime area to a low crime area.

Regression with `crime_bin` as predictor

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  17.61379  0.6433695 27.377413 8.806068e-102
## crime_binlow  7.49705  0.7942673  9.438951  1.389388e-19
```

The fitted regression equation is:

$$\widehat{median\_price} = 17.6 + 7.5 \times low\_crime$$

where y_i is the median price in thousands for tract i and $I(low\_crime)$ is 1 if it is a low crime area and 0 otherwise

How should we interpret the slope β_1?

On average, the median house price in low crime census tracts is $7,500 higher than in high crime census tracts.

Regression with categorical predictors

$$\widehat{median\_price} = 17.6 + 7.5 \times low\_crime$$

- R encodes categorical predictors as **indicator variables** (also called **dummy variables**)
- R picks a baseline value. Here the baseline is 'high'
- For high crime areas:

$$\widehat{median\_price} = 17.6$$

- For low crime areas:

$$\widehat{median\_price} = 17.6 + 7.5$$

Inference for simple linear regression

Could the difference between the average median price for high and low crime census tracts just be due to chance?

The regression model is

$$median\_price = \beta_0 + \beta_1 \times low\_crime + \epsilon$$

where

$$low\_crime = \begin{cases} 1 & \text{if } crime\_bin \text{ is low} \\ 0 & \text{if } crime\_bin \text{ is high} \end{cases}$$

We can answer this question by testing:

H_0: $\beta_1 = 0$

vs

H_A: $\beta_1 \neq 0$

Inference for simple linear regression

```
summary(lm(medv ~ crime_bin, data=Boston))$coefficients
```

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)    17.61379  0.6433695 27.377413 8.806068e-102
## crime_binlow    7.49705  0.7942673  9.438951  1.389388e-19
```

↳ pvalue we focus on.

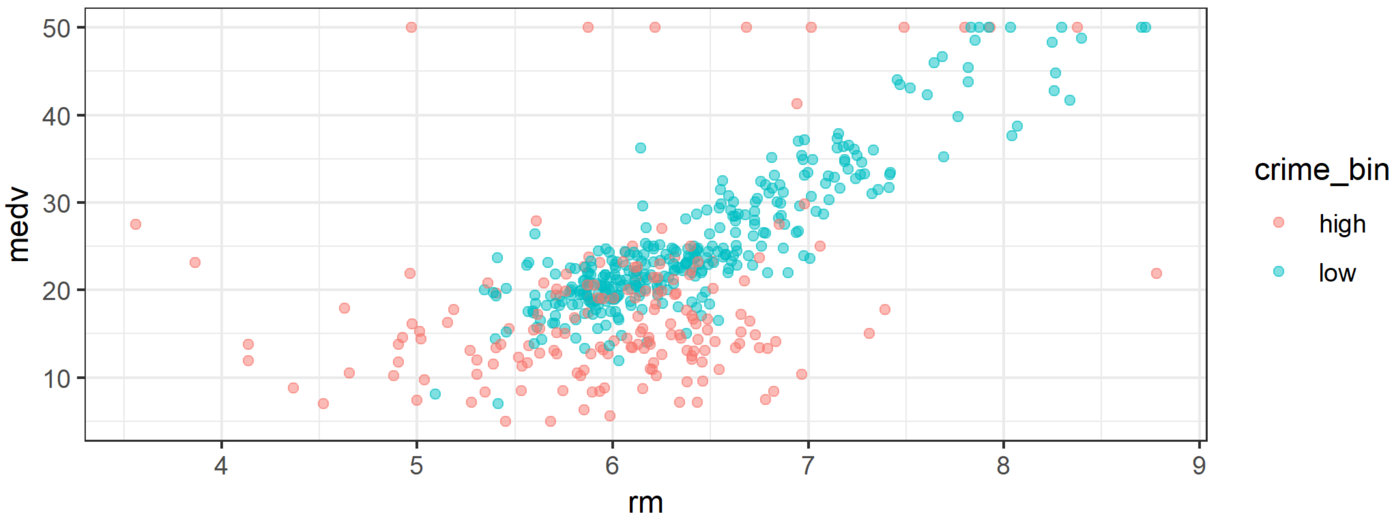
What conclusion would we make?

Since the pvalue is very close to 0,
there is very strong evidence against H_0.

Is the relationship between median price and average number of rooms the same in high and low crime areas?

Is the relationship between median price and average number of rooms the same in high and low crime areas?

```r
ggplot(Boston, aes(x=rm, y=medv, color=crime_bin)) +
  geom_point(alpha=0.5) + theme_bw()
```



Multiple linear regression

Regression equation (Model 1):

$$median\_price = \beta_0 + \beta_1 \times low\_crime + \beta_2 \times avg\_rooms + \epsilon$$

Model 1 for high crime areas

$$\hat{price} = \hat{\beta}_0 + \hat{\beta}_2 \times avg\ rooms$$

Model 1 for low crime areas

$$\hat{price} = (\hat{\beta}_0 + \hat{\beta}_1) + \hat{\beta}_2 \times avg\ rooms.$$

How would you describe these two lines?

parallel.

Fitted model

```
parallel_lines <- lm(medv ~ crime_bin + rm, data=Boston)
parallel_lines$coefficients
```

```
##   (Intercept) crime_binlow            rm
##    -32.607001     4.157368      8.339713
```

Regression equation:

$$median\_price = \beta_0 + \beta_1 \times low\_crime + \beta_2 \times avg\_rooms + \epsilon$$

Fitted regression equation:

$$\hat{price} = \hat{\beta}_0 + \hat{\beta}_1 \times low\ crime + \hat{\beta}_2 \times avg.rooms.$$

Fitted model

```
parallel_lines <- lm(medv ~ crime_bin + rm, data=Boston)
parallel_lines$coefficients
```

```
##  (Intercept) crime_binlow           rm
##   -32.607001     4.157368     8.339713
```

Regression equation:

$$median\_price = \beta_0 + \beta_1 \times low\_crime + \beta_2 \times avg\_rooms + \epsilon$$

Fitted regression equation:

$$median\_price = -32.6 + 4.2 \times low\_crime + 8.3 \times avg\_rooms + \epsilon$$

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

Plotting parallel lines

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

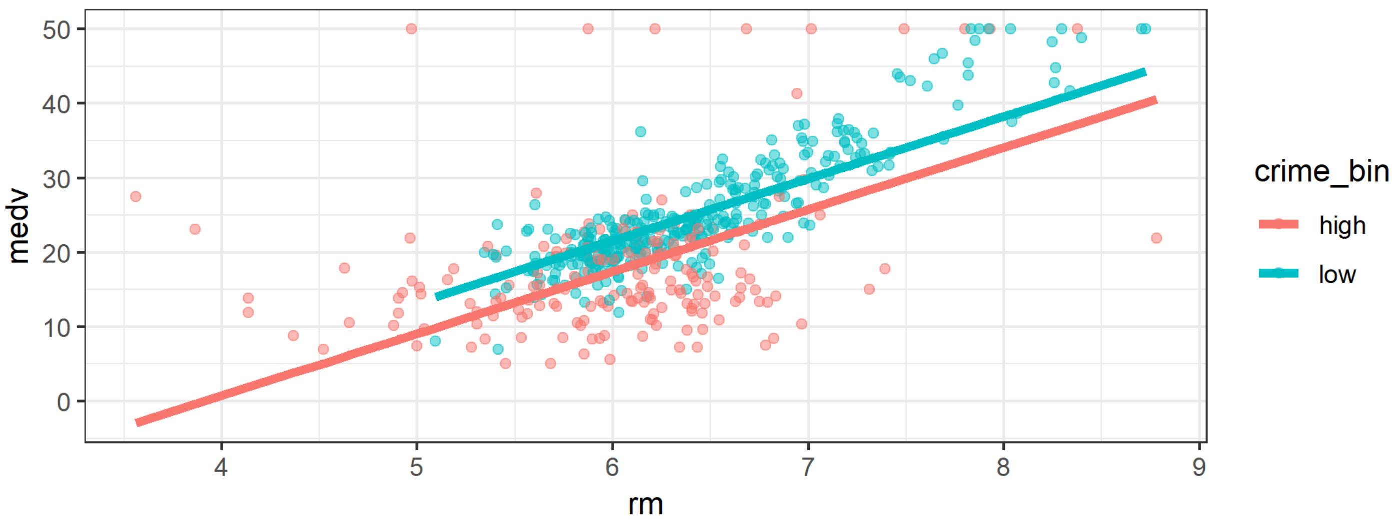
```
library(broom)
augment(parallel_lines)
```

```
## # A tibble: 506 x 10
##      medv crime_bin    rm .fitted .se.fit .resid    .hat .sigma .cooksd
##   * <dbl> <chr>     <dbl>   <dbl>   <dbl>  <dbl>   <dbl>  <dbl>   <dbl>
##  1  24    low        6.58    26.4   0.354  -2.38 0.00311   6.35 1.48e-4
##  2  21.6  low        6.42    25.1   0.348  -3.50 0.00301   6.35 3.08e-4
##  3  34.7  low        7.18    31.5   0.472   3.23 0.00553   6.35 4.83e-4
##  4  33.4  low        7.00    29.9   0.423   3.49 0.00445   6.35 4.52e-4
##  5  36.2  low        7.15    31.2   0.461   5.05 0.00529   6.34 1.13e-3
##  6  28.7  low        6.43    25.2   0.348   3.53 0.00301   6.35 3.12e-4
##  7  22.9  low        6.01    21.7   0.388   1.21 0.00374   6.35 4.58e-5
##  8  27.1  low        6.17    23.0   0.363   4.08 0.00328   6.35 4.55e-4
##  9  16.5  low        5.63    18.5   0.480  -2.01 0.00572   6.35 1.94e-4
## 10  18.9  low        6.00    21.6   0.389  -2.72 0.00377   6.35 2.33e-4
## # ... with 496 more rows, and 1 more variable: .std.resid <dbl>
```

Plotting the parallel lines

Join up the fitted values to plot the parallel lines model

```
ggplot(Boston, aes(x=rm, y=medv, color=crime_bin)) +
  geom_point(alpha=0.5) + theme_bw() +
  geom_line(data=augment(parallel_lines),
            aes(y=.fitted, colour=crime_bin), lwd=1.5)
```



Model with non-parallel lines

Add a new independent variable to the model, which is the product of `crime_bin` and `rm`. This is called an **interaction term**.

Model 2:

$$
\begin{aligned}
median\_price = \beta_0 + \beta_1 \times low\_crime + \beta_2 \times avg\_rooms \\
+ \beta_3 \times (low\_crime \times avg\_rooms) + \epsilon
\end{aligned}
$$

Model 2 for high crime areas

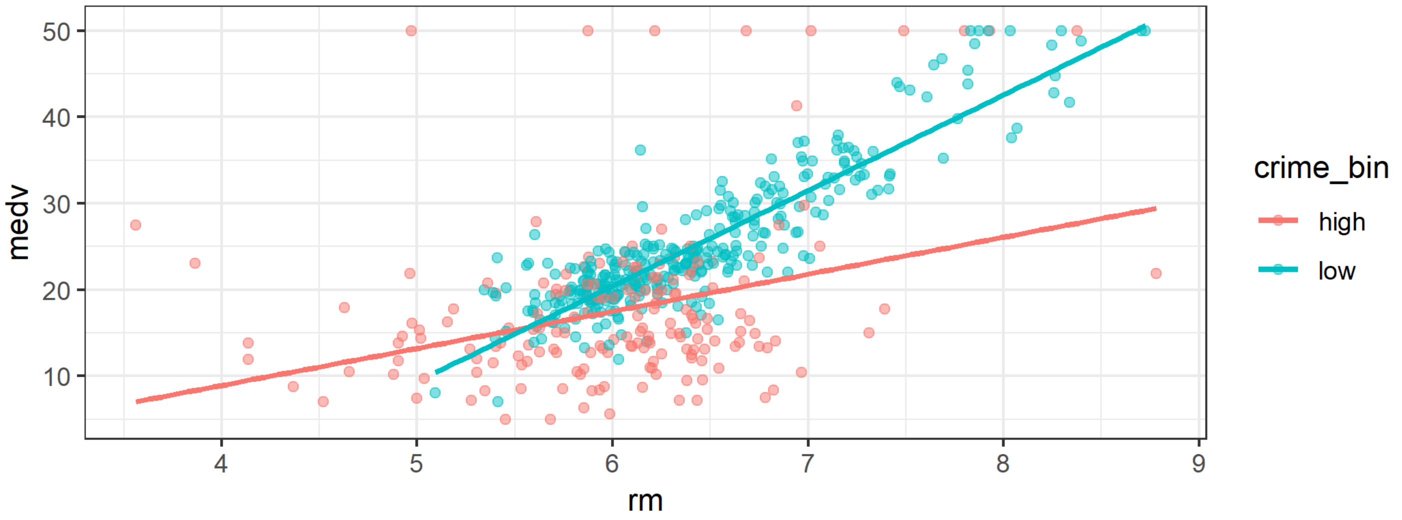
$$price = \beta_0 + \quad \beta_2 \times avg\ rooms + \varepsilon$$

Model 2 for low crime areas

$$price = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) \times avg\ rooms + \varepsilon.$$

Plot of non-parallel lines

```
ggplot(Boston, aes(x=rm, y=medv, color=crime_bin)) +
  geom_point(alpha=0.5) + theme_bw() +
  geom_smooth(method="lm", se=FALSE)
```



Fitted lines for high and low crime areas

Including the term `crime_bin * rm` on the right-side of the model in `lm` automatically includes both variables and their interaction in the model

```
summary(lm(medv ~ crime_bin * rm, data=Boston))$coefficients
```

```
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       -8.225571  3.7347764 -2.202427 2.808901e-02
## crime_binlow     -37.727624  4.9577523 -7.609824 1.367579e-13
## rm                 4.290910  0.6156830  6.969349 1.005587e-11
## crime_binlow:rm    6.774222  0.7963858  8.506206 2.078212e-16
```

β_0 β_1 β_2 β_3

based on last slide.

Fitted line for low income areas:

Fitted line for high income areas:

Could the difference in slopes for high and low income areas just be due to chance?

Model:

$$median\_price = \beta_0 + \beta_1 \times low\_crime + \beta_2 \times avg\_rooms + \beta_3 \times (low\_crime \times avg\_rooms) + \epsilon$$

What would be appropriate hypotheses to test this?

$H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$.

What do you conclude?

Reject H_0 bcs pvalue is very small

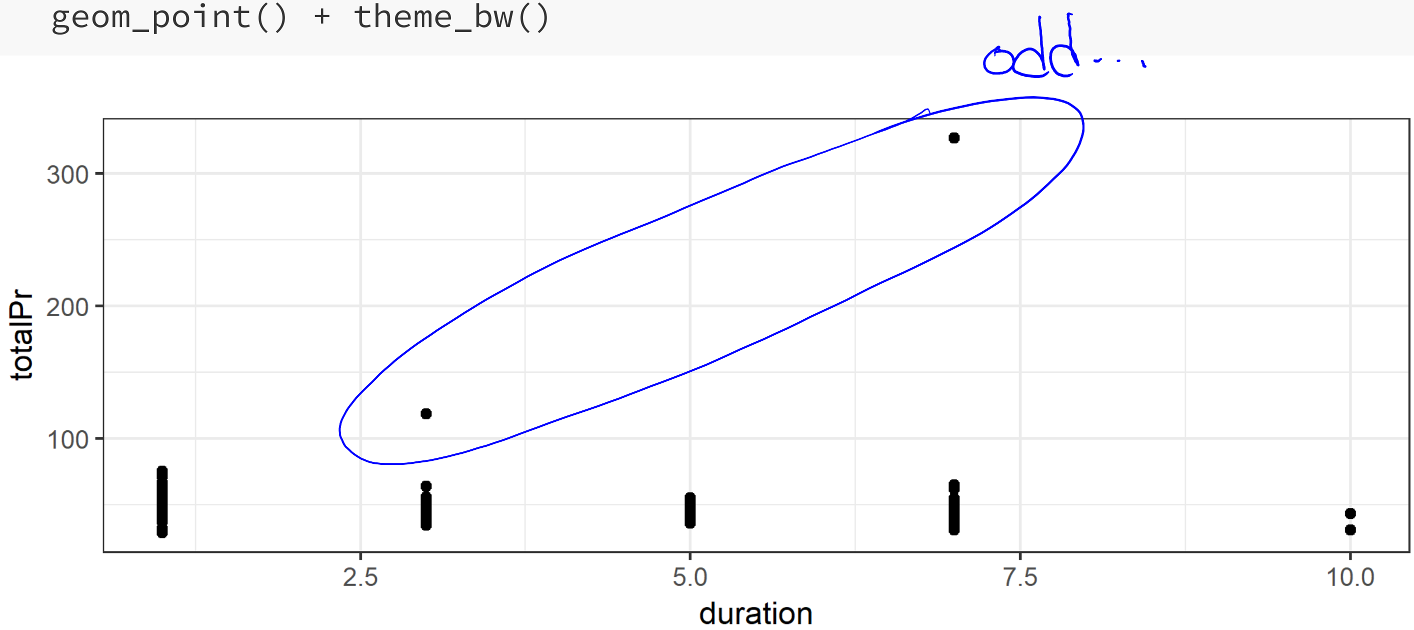
Example: eBay auctions of *Mario Kart*

- 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...
```

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

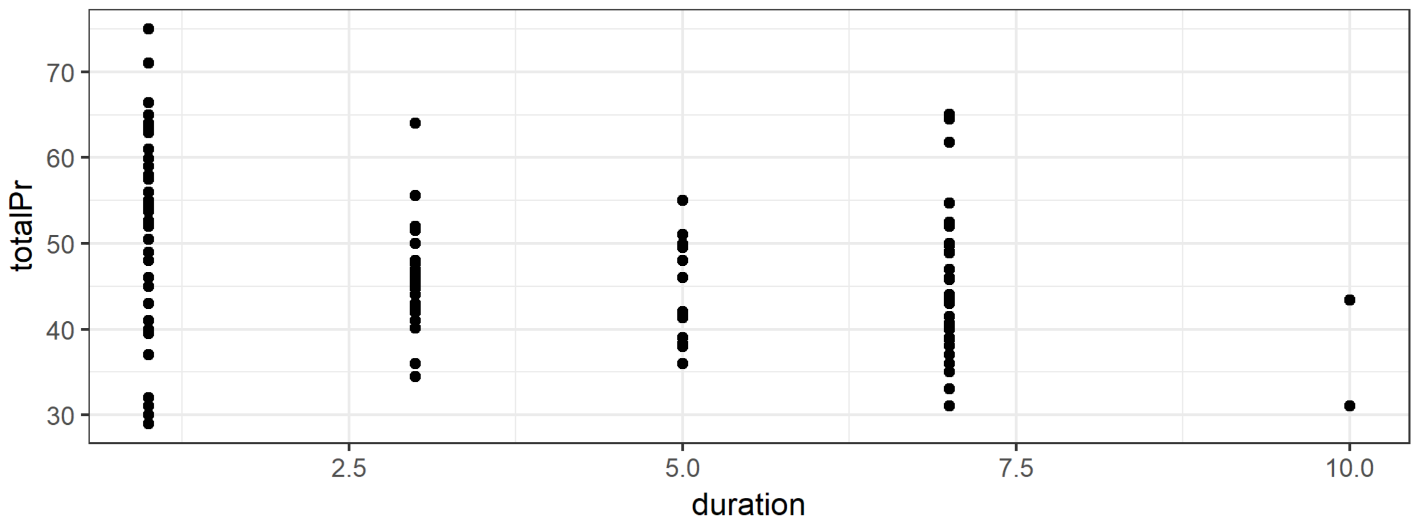


What should we do with the two outlying values of `totalPr`?

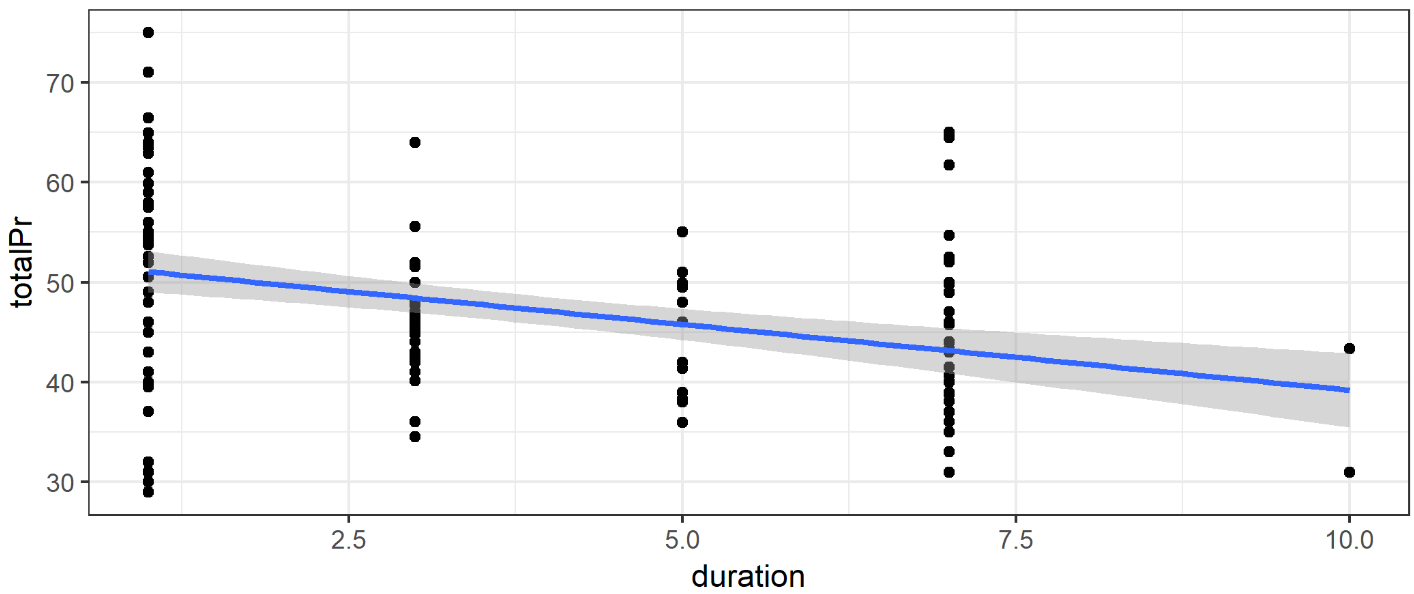
- Remove outliers only if there is a good reason.
- In these two auctions, and only these two auctions, the game was sold with other items.

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

```r
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?

Not really...

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
```

→ pvalue for intercept

→ pvalue for slope

→ pvalue = 4.9 x 10-6.

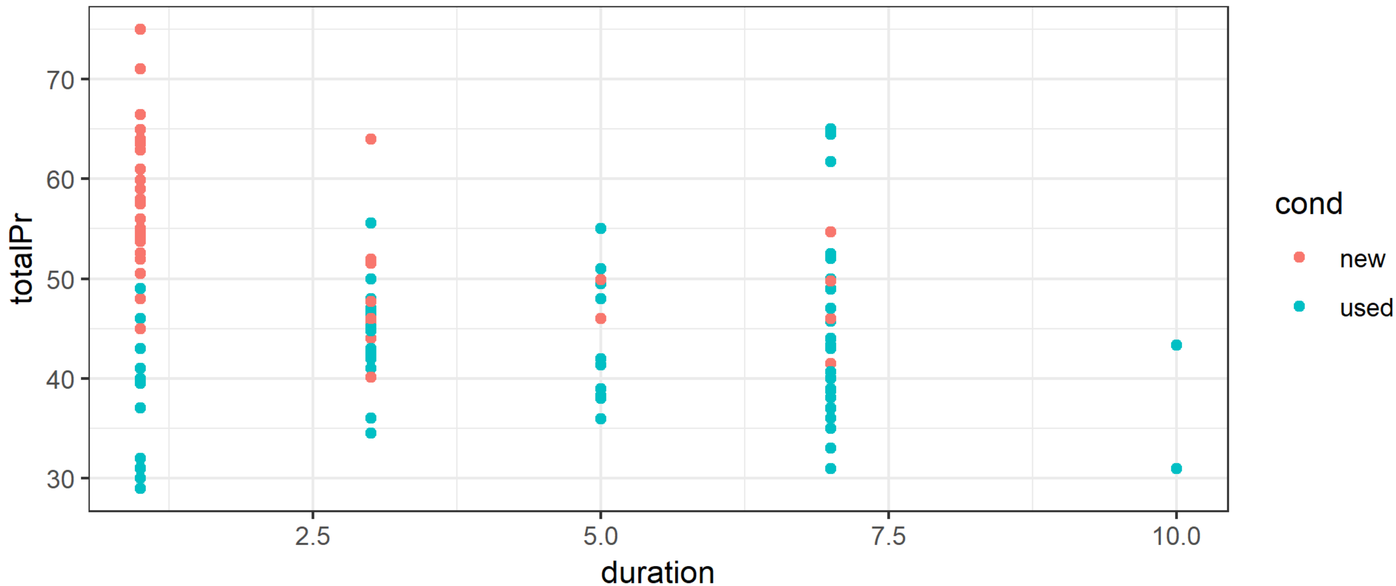
We have strong evidence that the slope is not 0.

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`.

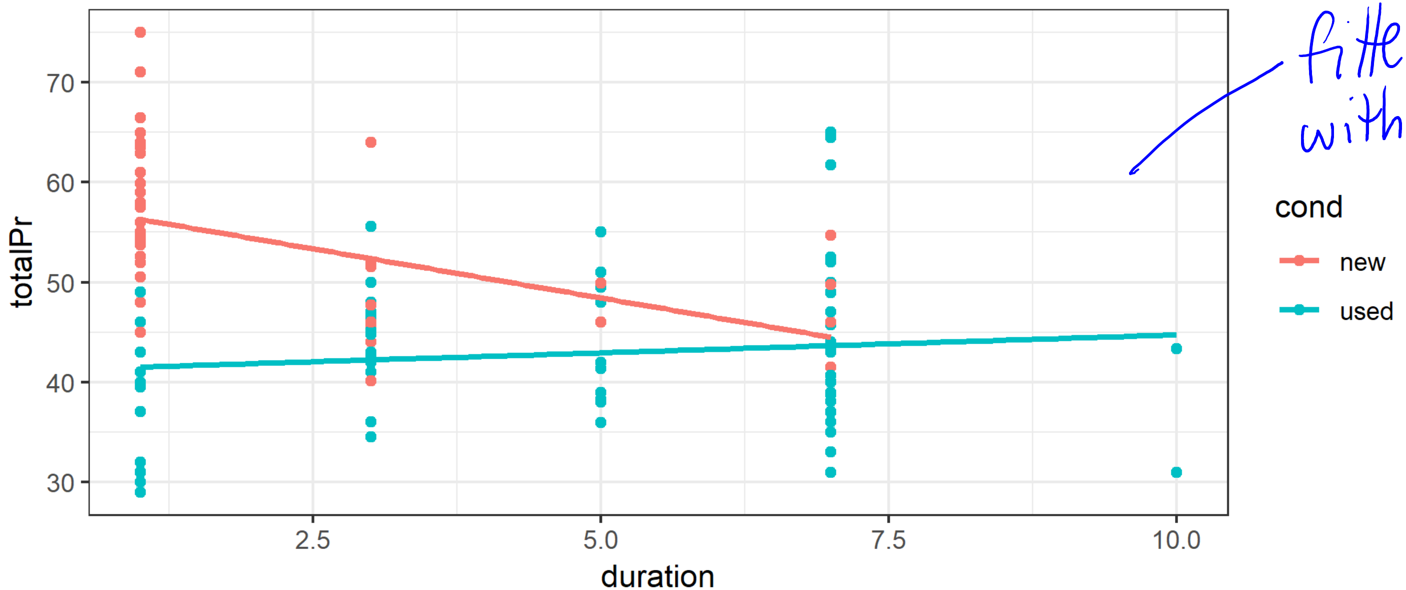
```
ggplot(marioKart2, aes(x=duration, y=totalPr, color=cond)) +
  geom_point() + theme_bw()
```



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

2 predictors: duration and condition.

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



fitted lines with interaction term

- 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
```

Based on the output above, which level of `cond` is the baseline (reference) level?

(a) new

(b) used

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.

```r
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
```


- Proportion of observations in each cell of contingency table.

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

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


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

```r
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
```

Some vocabulary

histograms

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.

→ prop.table( )

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.

addmargins( )

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

Practice questions

```
##                          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
```

What percentage of treatments were successfull?

80%

What percentage of individuals got treatment A?

50%

More vocabulary and notation:

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

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

$P(E_1 \mid E_2)$ 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?

The table below shows the joint distribution of `outcome` and `treatment`.

```
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
```

$$P(\text{success}) = 0.80285714$$

P(trt A and success)

$$P(\text{success} \mid \text{treatment A}) = 0.39/0.50 = 0.78$$

P(trt A)

$$P(\text{success} \mid \text{treatment B}) = 0.41285714/0.5 = 0.8257143$$

The table below shows the joint distribution of `outcome` and `treatment`.

```
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
```

$$P(\text{success}) = 0.80285714$$

$$P(\text{success} \mid \text{treatment A}) = 0.39/0.50 = 0.78$$

$$P(\text{success} \mid \text{treatment B}) = 0.41285714/0.5 = 0.8257143$$

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

↳ seems like trt B is better than trt A...

The table below shows the joint distribution of `outcome` and `treatment`.

```
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
```

$$P(\text{success}) = 0.80285714$$

$$P(\text{success} \mid \text{treatment A}) = 0.39/0.50 = 0.78$$

$$P(\text{success} \mid \text{treatment B}) = 0.41285714/0.5 = 0.8257143$$

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

Yes! Success is more likely with treatment B.

Independence

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?

↳ how bad their symptoms are.

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
```

Column percentages (conditional distribution of success given treatment):

```
prop.table(table(kidney_stones$outcome, kidney_stones$treatment),
           margin = 2) # columns sum to 1
```

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

overall

```
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
```

better

large stones ↳ bad.

```
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
```

small stones ↳ better.

Which treatment is better?

A is better...

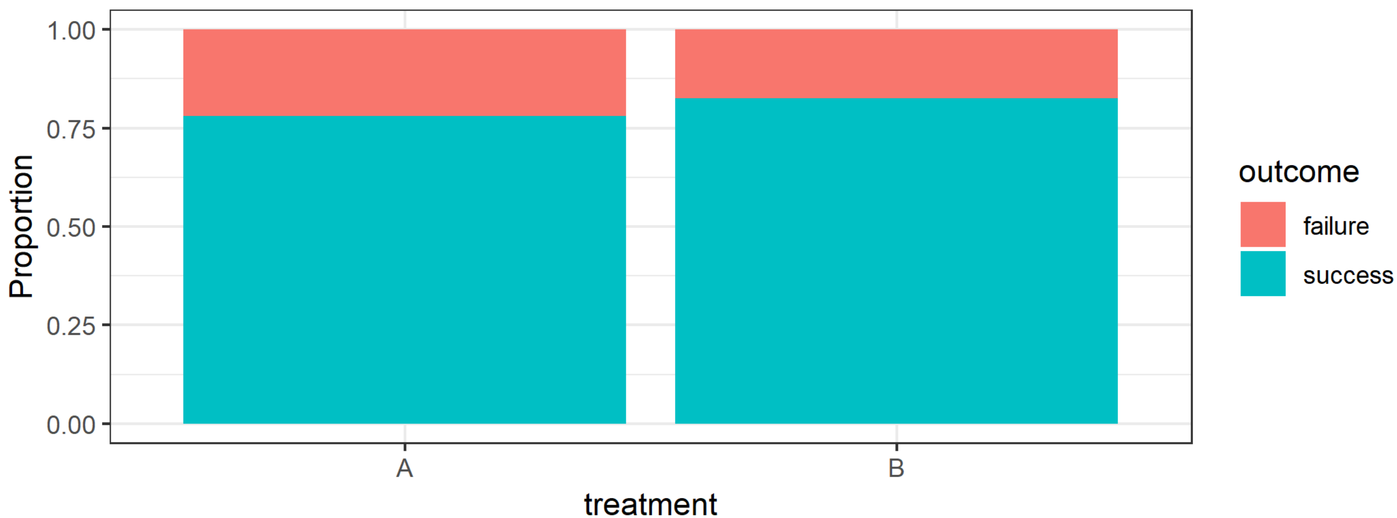
This example is another case of **Simpson's paradox**.

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.

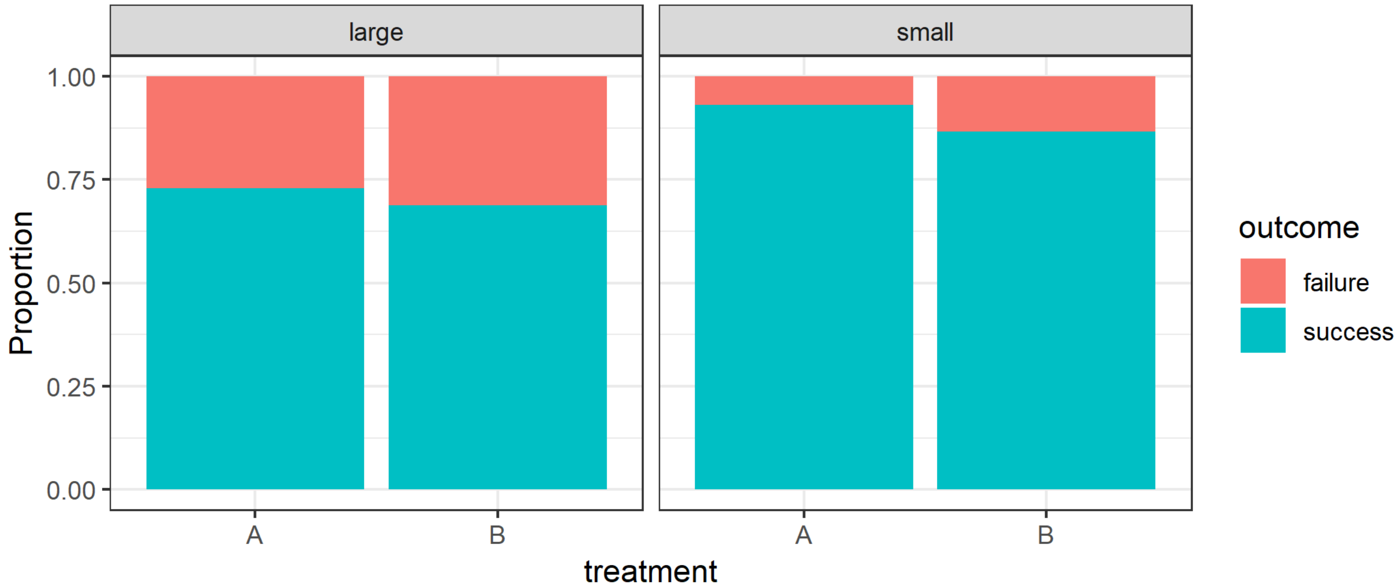
Visualizing the kidney stone data: treatment and outcome

```
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.

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.

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