SOC245: Visualizing Data
Precept 9: Characterizing Associations

Chris Felton and Vikram Ramaswamy

Freshman Scholars Institute
Princeton University

July 28, 2022

Outline

1. Linear Regression using OLS
    - Reviewing OLS
    - Computing OLS
2. Probability
    - Probability review
3. Computing probability of events in `R`

Outline

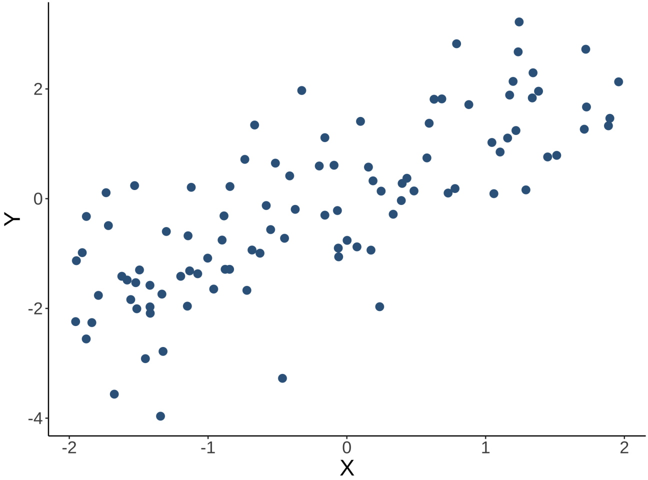
1. Linear Regression using OLS
   - Reviewing OLS
   - Computing OLS

2. Probability
   - Probability review

3. Computing probability of events in `R`

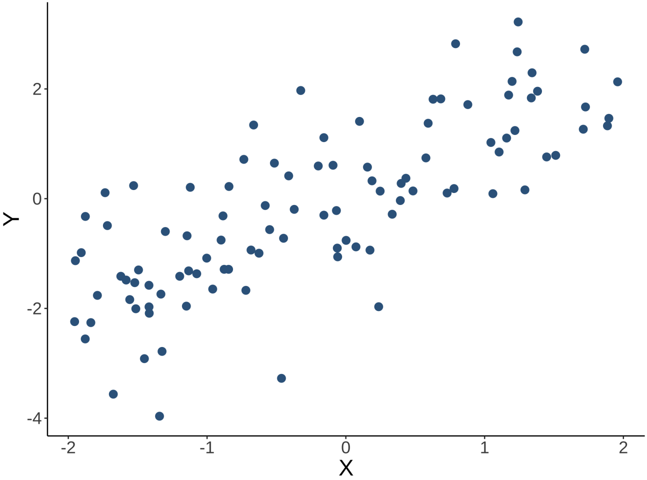
Reviewing OLS

Let's say we have a data that has a linear association.



Reviewing OLS

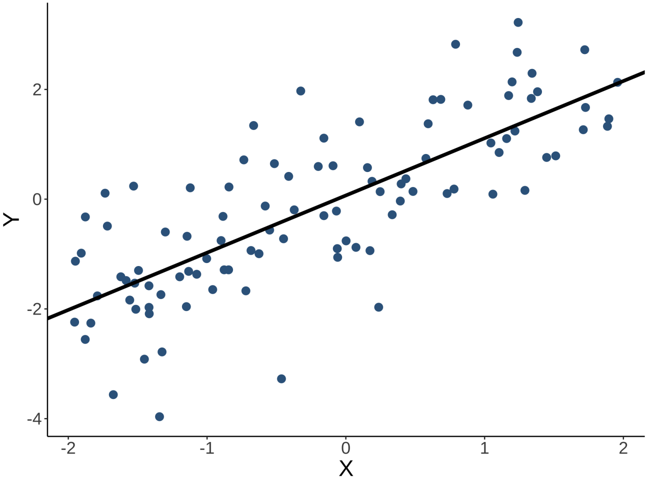
Let's say we have a data that has a linear association.



How do we find a line that best describes this data?

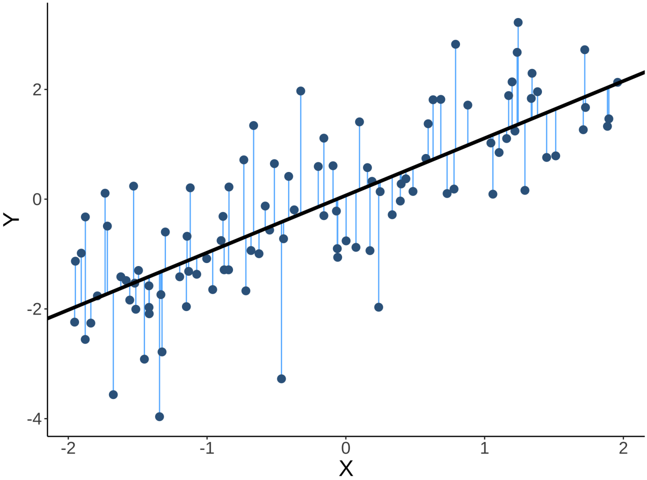
Reviewing OLS

Want to minimize the **sum of squared residuals (SSR)**.



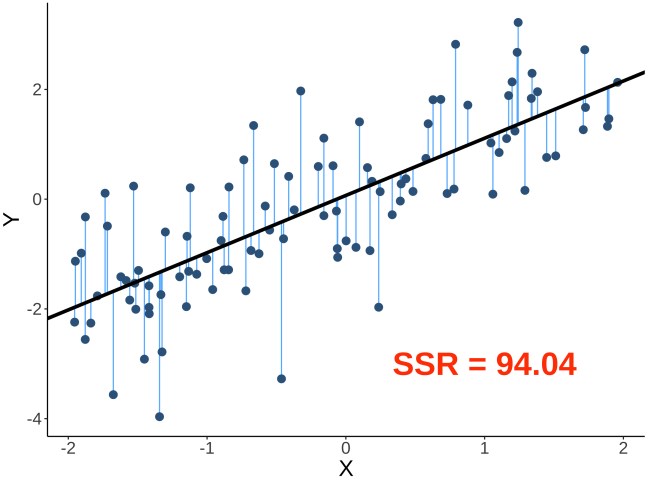
Reviewing OLS

Want to minimize the **sum of squared residuals (SSR)**.



Reviewing OLS

Want to minimize the **sum of squared residuals (SSR)**.



Reviewing OLS

- Our line is of the form $\widehat{Y} = a + bX$.
- Values for a and b that minimize the SSR are:

$$b = \frac{\widehat{\text{Cov}}(X, Y)}{\hat{\sigma}_X^2}$$

$$a = \overline{Y} - \frac{\widehat{\text{Cov}}(X, Y)}{\hat{\sigma}_X^2}\overline{X}$$

Let's work with the `oppotunity2.csv` dataset, that contains colleges along with the median income of children who graduate from it and the median income of their parents.

Let's work with the `oppotunity2.csv` dataset, that contains colleges along with the median income of children who graduate from it and the median income of their parents.

Load `tidyverse` and read in this dataset.

Computing OLS in `R`

What's the "manual" way to compute the linear regression?

Computing OLS in `R`

What's the "manual" way to compute the linear regression?

- We can use the equations we have to compute the slope and intercept!

$$b = \frac{\widehat{\text{Cov}}(X, Y)}{\hat{\sigma}_X^2}$$

$$a = \overline{Y} - \frac{\widehat{\text{Cov}}(X, Y)}{\hat{\sigma}_X^2}\overline{X}$$

Computing OLS in `R`

What's the "manual" way to compute the linear regression?

- We can use the equations we have to compute the slope and intercept!

$$b = \frac{\widehat{\text{Cov}}(X, Y)}{\hat{\sigma}_X^2}$$

$$a = \overline{Y} - \frac{\widehat{\text{Cov}}(X, Y)}{\hat{\sigma}_X^2}\overline{X}$$

- Note: `cov(X, Y)` finds the covariance between variables `X` and `Y`.

Computing OLS in `R`

What's the "manual" way to compute the linear regression?

- We can use the equations we have to compute the slope and intercept!

$$b = \frac{\widehat{\text{Cov}}(X, Y)}{\hat{\sigma}_X^2}$$

$$a = \overline{Y} - \frac{\widehat{\text{Cov}}(X, Y)}{\hat{\sigma}_X^2}\overline{X}$$

- Note: `cov(X, Y)` finds the covariance between variables `X` and `Y`.
- What's the formula for variance?

Computing OLS in R

```
b <- cov(col$par_median, col$k_median )/
        (var(col$par_median))
a <- mean(col$k_median) - b*mean(col$par_median)
```

What does this line look like?

Let's plot this association along with the linear regression

What does this line look like?

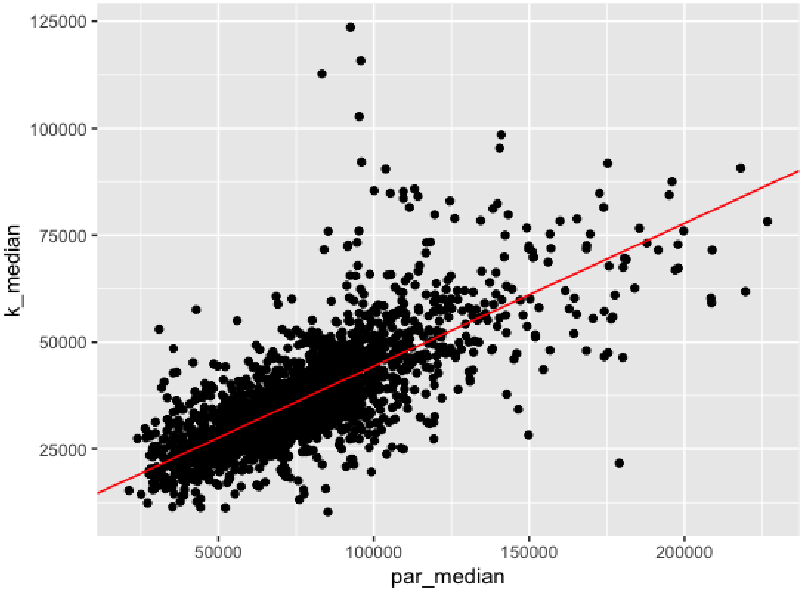
Let's plot this association along with the linear regression

```
ggplot(col) +
  geom_point(mapping =aes(x=par_median, y=k_median))
```

What does this line look like?

Let's plot this association along with the linear regression

```
ggplot(col) +
  geom_point(mapping =aes(x=par_median, y=k_median)) +
  geom_abline(intercept=a, slope=b, color="red")
```


125000
100000
75000
50000
25000
k_median
50000
100000
150000
200000
par_median

Computing OLS in `R`

`R` also has a built-in function to compute linear regression using OLS!

Computing OLS in `R`

`R` also has a built-in function to compute linear regression using OLS!

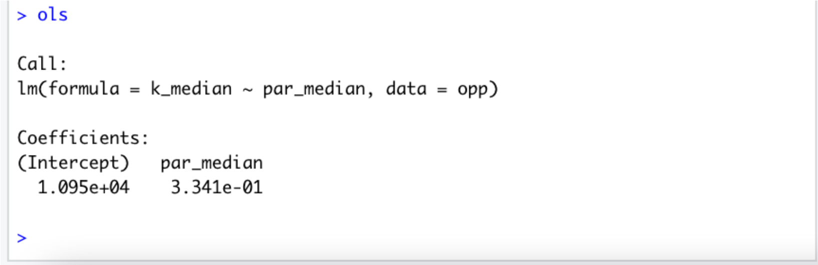
```
ols <- lm(k_median ~ par_median, data=col)
```

Computing OLS in `R`

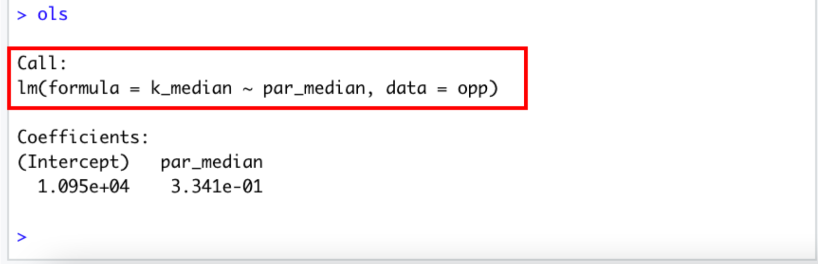
`R` also has a built-in function to compute linear regression using OLS!

```
ols <- lm(k_median ~ par_median, data=col)
```

Look at `ols`. What does it contain?

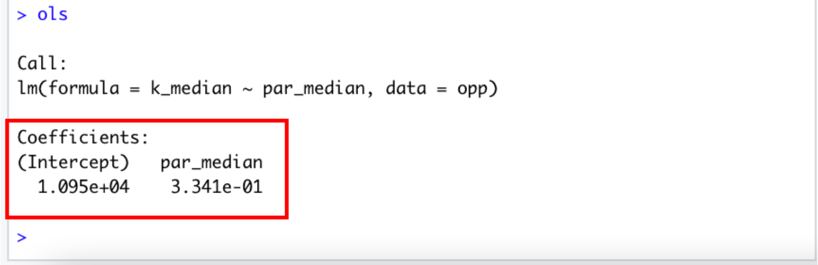


```
> ols

Call:
lm(formula = k_median ~ par_median, data = opp)

Coefficients:
(Intercept)   par_median
  1.095e+04    3.341e-01

>
```



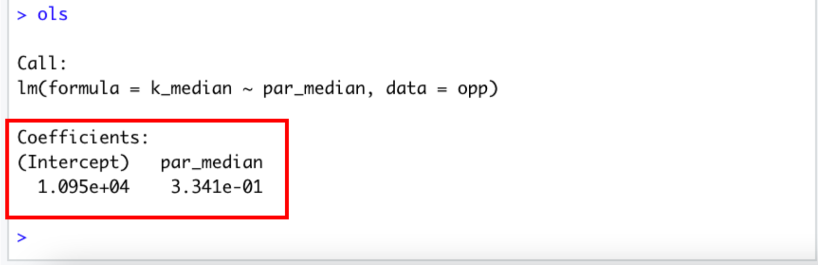
```
> ols

Call:
lm(formula = k_median ~ par_median, data = opp)

Coefficients:
(Intercept)   par_median
  1.095e+04    3.341e-01

>
```

We have what the regression model was trained on . . .



```
> ols

Call:
lm(formula = k_median ~ par_median, data = opp)

Coefficients:
(Intercept)   par_median
  1.095e+04    3.341e-01

>
```

. . . and the coefficients.

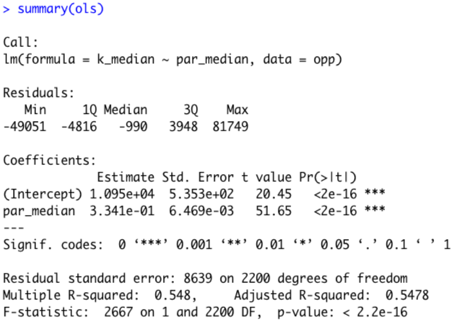


```
> ols

Call:
lm(formula = k_median ~ par_median, data = opp)

Coefficients:
(Intercept)   par_median
  1.095e+04    3.341e-01

>
```

. . . and the coefficients. (Are these what you found before?)

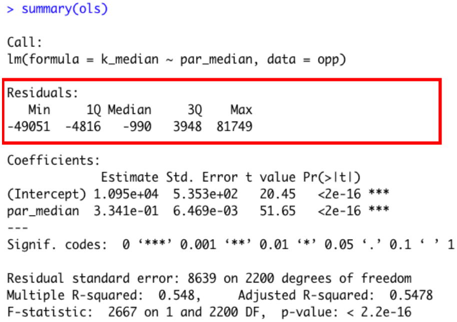
Now, try running `summary(ols)`

Now, try running `summary(ols)`



```
> summary(ols)

Call:
lm(formula = k_median ~ par_median, data = opp)

Residuals:
   Min     1Q Median     3Q    Max 
-49051  -4816   -990   3948  81749 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.095e+04  5.353e+02   20.45   <2e-16 ***
par_median  3.341e-01  6.469e-03   51.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8639 on 2200 degrees of freedom
Multiple R-squared:  0.548,     Adjusted R-squared:  0.5478 
F-statistic:  2667 on 1 and 2200 DF,  p-value: < 2.2e-16
```

Now, try running `summary(ols)`



```
> summary(ols)

Call:
lm(formula = k_median ~ par_median, data = opp)

Residuals:
   Min     1Q Median     3Q    Max 
-49051  -4816   -990   3948  81749 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.095e+04  5.353e+02   20.45   <2e-16 ***
par_median  3.341e-01  6.469e-03   51.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8639 on 2200 degrees of freedom
Multiple R-squared:  0.548,	Adjusted R-squared:  0.5478 
F-statistic:  2667 on 1 and 2200 DF,  p-value: < 2.2e-16
```

We also have some statistics about the residuals.

Accessing the residuals

You can access the residuals using `ols$residuals`.

Accessing the residuals

You can access the residuals using `ols$residuals`.

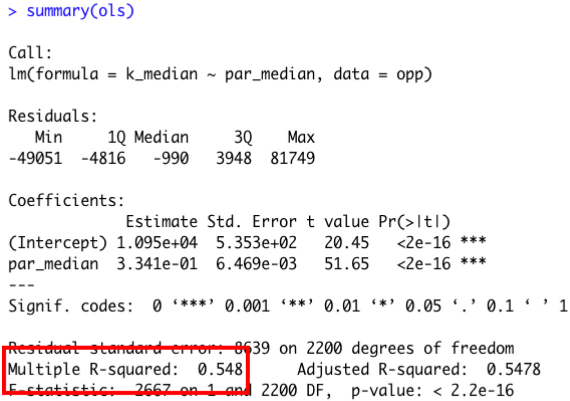
How would you compute the SSR of this line?

Computing R^2

Finally, the summary also contains R^2 value.

Computing R^2

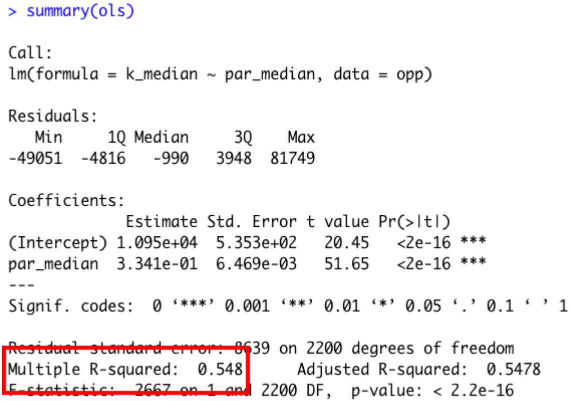
Finally, the summary also contains R^2 value.



```
> summary(ols)

Call:
lm(formula = k_median ~ par_median, data = opp)

Residuals:
   Min     1Q Median     3Q    Max 
-49051  -4816   -990   3948  81749 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.095e+04  5.353e+02   20.45   <2e-16 ***
par_median  3.341e-01  6.469e-03   51.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8639 on 2200 degrees of freedom
Multiple R-squared:  0.548	Adjusted R-squared:  0.5478 
F-statistic:  2667 on 1 and 2200 DF,  p-value: < 2.2e-16
```

Computing R^2

Finally, the summary also contains R^2 value.



```
> summary(ols)

Call:
lm(formula = k_median ~ par_median, data = opp)

Residuals:
   Min     1Q Median     3Q    Max 
-49051  -4816   -990   3948  81749 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.095e+04  5.353e+02   20.45   <2e-16 ***
par_median  3.341e-01  6.469e-03   51.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8639 on 2200 degrees of freedom
Multiple R-squared:  0.548	Adjusted R-squared:  0.5478 
F-statistic:  2667 on 1 and 2200 DF,  p-value: < 2.2e-16
```

We can access it using `summary(ols)$r.squared`.

Outline

Recall

- Probabilities indicate our uncertainity about events.

Recall

- Probabilities indicate our uncertainity about events.
- Intuitively, we measure how likely an event is to occur.

Examples

Let's say I'm rolling a 6-sided die. What is the probability I get an even number?

Examples

Suppose I have a bag containing 5 white marbles, 6 red marbles, and 4 black marbles, and I pick a marble without looking. What is the probability I pick a white marble?

Outline

1. Linear Regression using OLS
    - Reviewing OLS
    - Computing OLS
2. Probability
    - Probability review
3. Computing probability of events in `R`

Read in the dataset `masc_raw-responses.csv`. This is a dataset containing responses from roughly 6000 men about what they think it means to be a man.

Read in the dataset `masc_raw-responses.csv`. This is a dataset containing responses from roughly 6000 men about what they think it means to be a man.

The full list of questions is in the `masculinity-survey.pdf`

We'll consider question 7(b) in particular to analyze today.

We'll consider question 7(b) in particular to analyze today.

This is the question: How often would you say you ask a friend for personal advice?

Answer options:

1. Often
2. Sometimes
3. Rarely
4. Never, but open to it
5. Never, and not open to it

What is the probability that the survey respondant said “Rarely”?

Computing probabilities in R

- Let R denote the event that the response is “Rarely”

Computing probabilities in `R`

- Let R denote the event that the response is “Rarely”
- We want to compute $\Pr(R)$

Computing probabilities in `R`

- Let R denote the event that the response is “Rarely”
- We want to compute $\Pr(R)$
- How can we do this?

$\Pr(R)$

- What should our denominator be?

$\Pr(R)$

- What should our denominator be?
  - Total number of observations in the dataset

Pr(R)

- What should our denominator be?
  - Total number of observations in the dataset

```
denom <- nrow(masc)
```

$\Pr(R)$

- What should our denominator be?
  - Total number of observations in the dataset

```
denom <- nrow(masc)
```

- What should our numerator be?

$\Pr(R)$

- What should our denominator be?
    - Total number of observations in the dataset

```
denom <- nrow(masc)
```

- What should our numerator be?
    - Number of respondants who replied “Rarely”

Pr(R)

- What should our denominator be?
  - Total number of observations in the dataset

```
denom <- nrow(masc)
```

- What should our numerator be?
  - Number of respondants who replied "Rarely"

```
num <- nrow(filter(masc, q0007_0002 == "Rarely"))
```

You try!

What is the probability that the survey respondant said either “Never, but open to it” or “Never, and not open to it”?

You try!

What is the probability that the survey respondant said either “Never, but open to it” or “Never, and not open to it”?

```
nrow(filter(masc,
      q0007_0002 == "Never, but open to it" |
      q0007_0002 == "Never, and not open to it")) /
  nrow(masc)
```

How can we find

$$\Pr(R|\text{Respondant is over 65})$$