Social Science Inquiry II

Week 6: Linear models, part I

Molly Offer-Westort

Department of Political Science,
University of Chicago

Winter 2022

Loading packages for this class

```
> library(ggplot2)
```

- We often want to discuss the relationship between an *independent variable* and a *dependent variable*.

- We often want to discuss the relationship between an *independent variable* and a *dependent variable*.
- One way to do this is to talk about a conditional mean; for example, if $X \in \{0, 1\}$, we may be interested in $\mathrm{E}[Y|X = 0]$ and $\mathrm{E}[Y|X = 1]$.

- We often want to discuss the relationship between an *independent variable* and a *dependent variable*.
- One way to do this is to talk about a conditional mean; for example, if $X \in \{0, 1\}$, we may be interested in $\mathrm{E}[Y|X = 0]$ and $\mathrm{E}[Y|X = 1]$.
- (Which one is the *independent variable* and which is the *dependent variable*?)

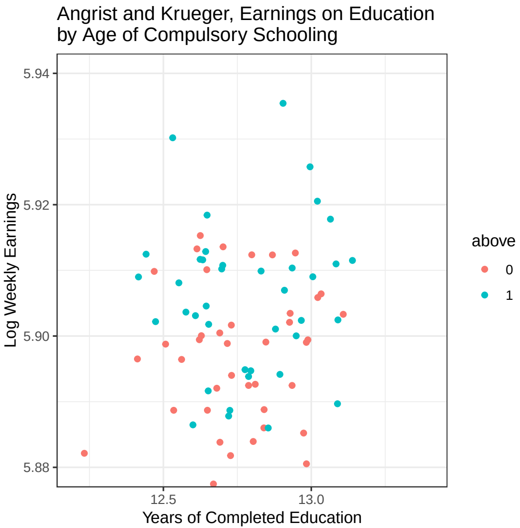
- We often want to discuss the relationship between an *independent variable* and a *dependent variable*.
- One way to do this is to talk about a conditional mean; for example, if $X \in \{0, 1\}$, we may be interested in $\mathrm{E}[Y|X = 0]$ and $\mathrm{E}[Y|X = 1]$.
- (Which one is the *independent variable* and which is the *dependent variable*?)
- What if X takes on more than a few values?

Recall:

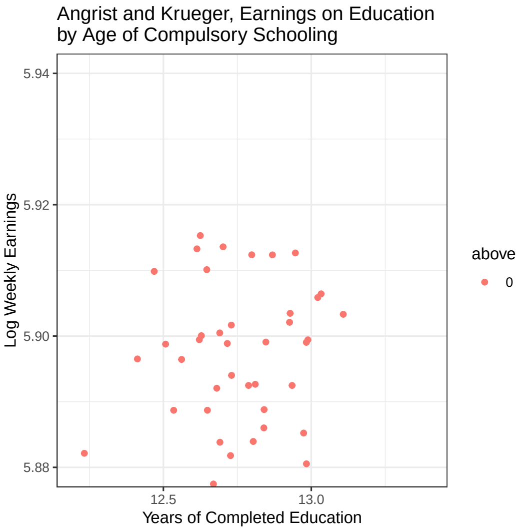
Angrist, Joshua D., and Alan B. Krueger. ''Does compulsory school attendance affect schooling and earnings?'' *The Quarterly Journal of Economics* 106.4 (1991): 979-1014.

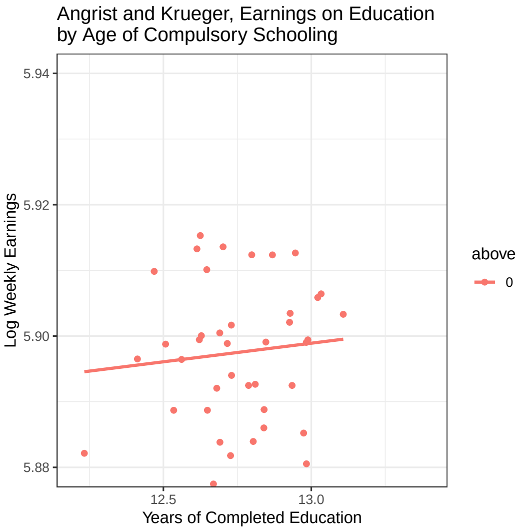
```
> file <- "https://raw.githubusercontent.com/UChicago-pol-methods/SOSC13200-W22/mai
> dat <- read.csv(file, as.is = TRUE)
> dat$year_of_birth_adj <- dat$year_of_birth +
+   0.25 * (dat$quarter_of_birth-1)
> states_above_16 <- c(15, 23, 32, 35, 39, 40, 41, 42, 48, 49, 51, 53)
> dat$states_above_16 <- 1 * (dat$place_of_birth %in% states_above_16)
> dat_agg <- aggregate(x = dat[, c('log_weekly_wage', 'education')],
+                      by = list(`year_of_birth_adj` = dat$year_of_birth_adj,
+                                `quarter_of_birth` = dat$quarter_of_birth,
+                                `above` = as.factor(dat$states_above_16)),
+                      FUN = mean)
```

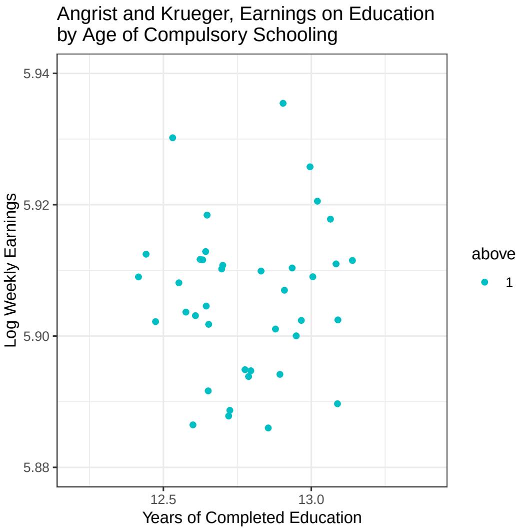
```
> head(dat_agg)
  year_of_birth_adj quarter_of_birth above log_weekly_wage education
1                30                1     0        5.882141  12.23273
2                31                1     0        5.898764  12.50745
3                32                1     0        5.888690  12.53485
4                33                1     0        5.892047  12.68044
5                34                1     0        5.888694  12.64883
6                35                1     0        5.877465  12.66922
```

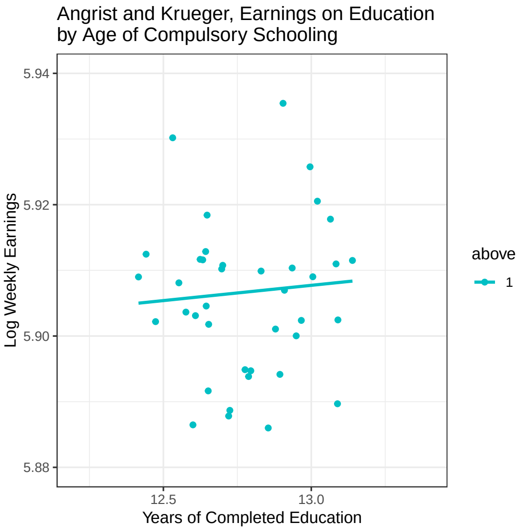

Angrist and Krueger, Earnings on Education
by Age of Compulsory Schooling
Log Weekly Earnings
5.94
5.92
5.90
5.88
12.5
13.0
Years of Completed Education
above
0
1

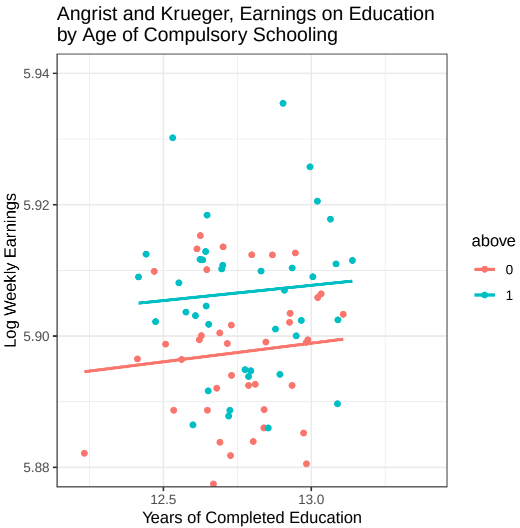
- Suppose we want to draw a line through these points.
- What is the best way to pick the line?


Angrist and Krueger, Earnings on Education
by Age of Compulsory Schooling
Log Weekly Earnings
5.94
5.92
5.90
5.88
12.5
13.0
Years of Completed Education
above
0


Angrist and Krueger, Earnings on Education
by Age of Compulsory Schooling
5.94
5.92
5.90
5.88
Log Weekly Earnings
12.5
13.0
Years of Completed Education
above
0


Angrist and Krueger, Earnings on Education
by Age of Compulsory Schooling
Log Weekly Earnings
5.94
5.92
5.90
5.88
12.5
13.0
Years of Completed Education
above
1


Angrist and Krueger, Earnings on Education
by Age of Compulsory Schooling
Log Weekly Earnings
5.94
5.92
5.90
5.88
12.5
13.0
Years of Completed Education
above
1


Angrist and Krueger, Earnings on Education
by Age of Compulsory Schooling
5.94
5.92
5.90
5.88
Log Weekly Earnings
12.5
13.0
Years of Completed Education
above
0
1

- We would like to describe a conditional relationship in the data

$$\mathrm{E}\left[Y|X=x\right]=g(x)$$

where the simplest version of $g(x)$ is

$$g(x)=\beta_0+\beta_1 x$$

- We would like to describe a conditional relationship in the data

$$\mathrm{E}\left[Y|X=x\right]=g(x)$$

where the simplest version of $g(x)$ is

$$g(x)=\beta_0+\beta_1 x$$

- In other words,

$$Y_i=\beta_0+\beta_1 X_i+\epsilon_i$$

where

$$\mathrm{E}\left[\epsilon_i|X_i\right]=0$$

and

$$\mathrm{Var}\left[\epsilon_i|X_i\right]=\sigma^2$$

- In practice, we describe Y_i as a function of X_i in the data we *observe*.

- In practice, we describe Y_i as a function of X_i in the data we *observe*.
- We refer to this as "regressing Y *on* X."

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- In practice, we describe Y_i as a function of X_i in the data we *observe*.
- We refer to this as "regressing Y *on* X."

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- We can then define *residuals*

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right).$$

- We calculate estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ as the values that minimize the residual sums of squares

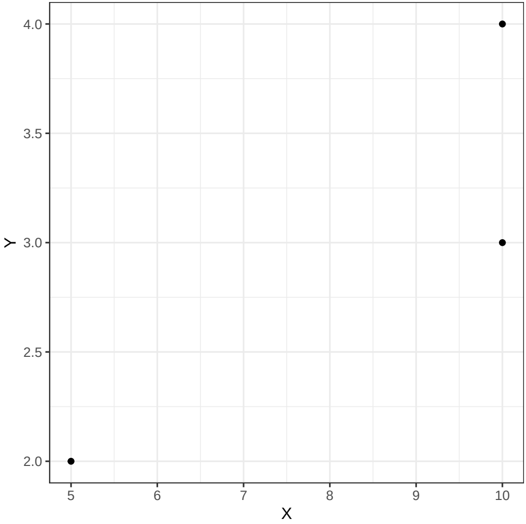
- We calculate estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ as the values that minimize the residual sums of squares

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

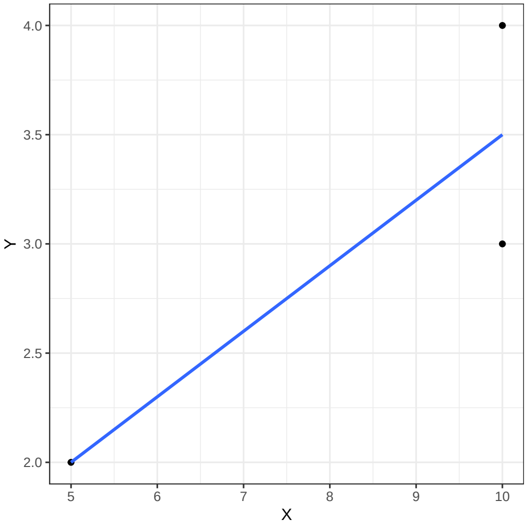
Suppose we had the following data points:

```
> toy_dat <- data.frame(Y = c(2, 3, 4),
+                       X = c(5, 10, 10))
> toy_dat
  Y  X
1 2  5
2 3 10
3 4 10
```

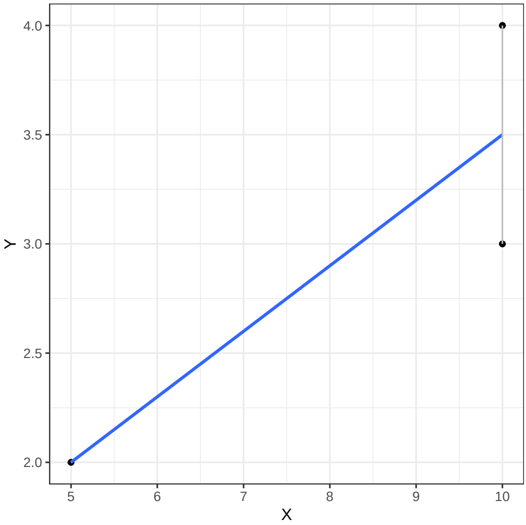
What values for $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the residual sum of squares?



What values for $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the residual sum of squares?



What values for $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the residual sum of squares?

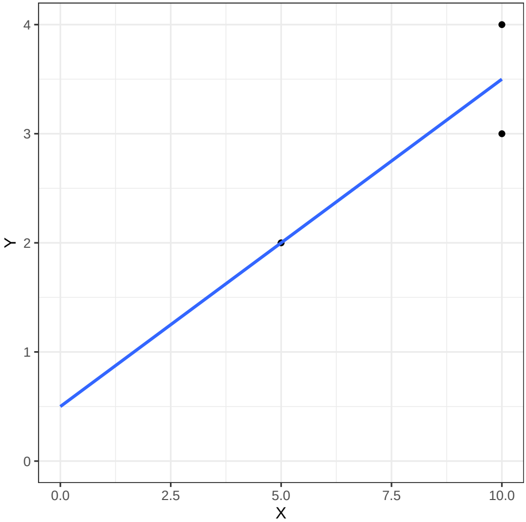


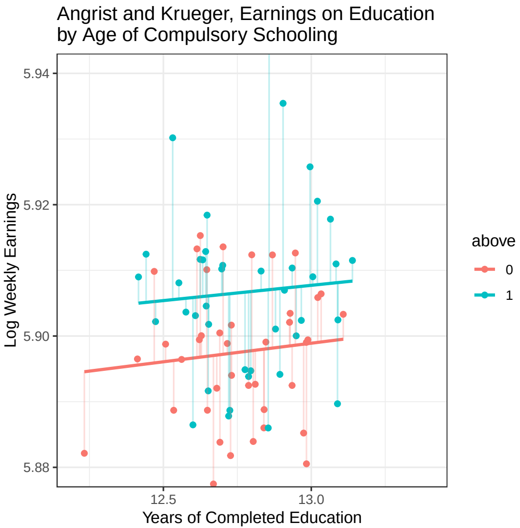
We can also think about $\hat{\beta}_0$ and $\hat{\beta}_1$ as the *y-intercept*, i.e., where the line crosses the y-axis, and the *slope*, respectively.

```
> lm(Y~X, data = toy_dat)
Call:
lm(formula = Y ~ X, data = toy_dat)

Coefficients:
(Intercept)            X
        0.5          0.3
```

We can also think about $\hat{\beta}_0$ and $\hat{\beta}_1$ as the *y-intercept*, i.e., where the line crosses the y-axis, and the *slope*, respectively.





Angrist and Krueger, Earnings on Education
by Age of Compulsory Schooling
Log Weekly Earnings
5.94
5.92
5.90
5.88
12.5
13.0
Years of Completed Education
above
0
1

- Why do we minimize the sum of *squared* distances? (rather than…absolute distances? Or cubed distances?)

- Why do we minimize the sum of *squared* distances? (rather than... absolute distances? Or cubed distances?)
- According to IMS:
  - *It is the most commonly used method.*
  - *Computing the least squares line is widely supported in statistical software.*
  - *In many applications, a residual twice as large as another residual is more than twice as bad. For example, being off by 4 is usually more than twice as bad as being off by 2. Squaring the residuals accounts for this discrepancy.*
  - *The analyses which link the model to inference about a population are most straightforward when the line is fit through least squares.*

- Why do we minimize the sum of *squared* distances? (rather than…absolute distances? Or cubed distances?)
- According to IMS:
  - *It is the most commonly used method.*
  - *Computing the least squares line is widely supported in statistical software.*
  - *In many applications, a residual twice as large as another residual is more than twice as bad. For example, being off by 4 is usually more than twice as bad as being off by 2. Squaring the residuals accounts for this discrepancy.*
  - *The analyses which link the model to inference about a population are most straightforward when the line is fit through least squares.*(What does this mean??)

- Why do we minimize the sum of *squared* distances? (rather than...absolute distances? Or cubed distances?)
- According to IMS:
  - *It is the most commonly used method.*
  - *Computing the least squares line is widely supported in statistical software.*
  - *In many applications, a residual twice as large as another residual is more than twice as bad. For example, being off by 4 is usually more than twice as bad as being off by 2. Squaring the residuals accounts for this discrepancy.*
  - *The analyses which link the model to inference about a population are most straightforward when the line is fit through least squares.*(What does this mean??)
- Other potential reasons...
  - Squared distances will always be positive (so will absolute distances)
  - But absolute distances don't provide a unique solution to the minimization problem, squared distances do
  - It's easier to take the derivative of the square, rather than absolute.
  - **Minimizing RSS gives a linear approximation to the conditional expectation function.**

- Why do we minimize the sum of *squared* distances? (rather than...absolute distances? Or cubed distances?)
- According to IMS:
  - *It is the most commonly used method.*
  - *Computing the least squares line is widely supported in statistical software.*
  - *In many applications, a residual twice as large as another residual is more than twice as bad. For example, being off by 4 is usually more than twice as bad as being off by 2. Squaring the residuals accounts for this discrepancy.*
  - *The analyses which link the model to inference about a population are most straightforward when the line is fit through least squares.*(What does this mean??)
- Other potential reasons...
  - Squared distances will always be positive (so will absolute distances)
  - But absolute distances don't provide a unique solution to the minimization problem, squared distances do
  - It's easier to take the derivative of the square, rather than absolute.
  - **Minimizing RSS gives a linear approximation to the conditional expectation function.** (Why is it only an approximation? When is it not approximate?)

Frametitle

- Why do we take the squared distance in terms of Y–instead of in terms of X?

Frametitle

- Why do we take the squared distance in terms of Y–instead of in terms of X?
- What if we regressed X Y instead of Y X?

Returning to Butler and Broockman...

Butler, Daniel M., & Broockman, David E. (2011). Do politicians racially discriminate against constituents? A field experiment on state legislators.

Data is available at the Yale ISPS data archive:
`isps.yale.edu/research/data`

Loading the data

```
> df <- read.csv('../data/butler-broockman.csv', as.is = TRUE)
> head(df)
  leg_party leg_republican leg_black leg_latino reply_atall treat_deshawn
1         R              1         0          0           1             0
2         D              0         0          0           1             1
3         R              1         0          0           0             1
4         R              1         0          0           0             1
5         D              0         0          0           0             0
6         D              0         0          0           1             1
  treat_demprimary treat_repprimary treat_noprimary treat_group treat_jake
1                1                0               0           5          1
2                0                1               0           2          0
3                0                0               1           6          0
4                0                0               1           6          0
5                0                0               1           0          1
6                0                1               0           2          0
  leg_notwhite leg_white leg_notblackotherminority treat_primary
1            0         1                         0             1
2            0         1                         0             1
3            0         1                         0             0
4            0         1                         0             0
5            0         1                         0             0
6            0         1                         0             1
```

Recall that treatment is 1 if the sender was DeShawn Jackson, and 0 if Jake Mueller.

```
> table(df$treat_deshawn)

   0    1
2431 2428
```

Recall that treatment is 1 if the sender was DeShawn Jackson, and 0 if Jake Mueller.

```
> table(df$treat_deshawn)

   0    1
2431 2428
```

The primary outcome is whether legislators replied at all.

```
> table(df$reply_atall)

   0    1
2112 2747
```

We're going to manipulate our data so it takes the format $Y \sim D$.

```
> df$D <- df$treat_deshawn
> df$Y <- df$reply_atall
```

We're going to manipulate our data so it takes the format $Y \sim D$.

```
> df$D <- df$treat_deshawn
> df$Y <- df$reply_atall
```

To get the difference-in-means estimate of the ATE,

```
> Y1 <- df$Y[which(df$D == 1)]
> Y0 <- df$Y[which(df$D == 0)]
```

We're going to manipulate our data so it takes the format $Y \sim D$.

```
> df$D <- df$treat_deshawn
> df$Y <- df$reply_atall
```

To get the difference-in-means estimate of the ATE,

```
> Y1 <- df$Y[which(df$D == 1)]
> Y0 <- df$Y[which(df$D == 0)]
> (dm_hat <- mean(Y1) - mean(Y0))
[1] -0.01782424
```

We're going to manipulate our data so it takes the format $Y \sim D$.

```
> df$D <- df$treat_deshawn
> df$Y <- df$reply_atall
```

To get the difference-in-means estimate of the ATE,

```
> Y1 <- df$Y[which(df$D == 1)]
> Y0 <- df$Y[which(df$D == 0)]
> (dm_hat <- mean(Y1) - mean(Y0))
[1] -0.01782424
```

Legislators were 1.7 percentage points less likely to reply to an email if the sender was identified as DeShawn Jackson as compared to Jake Mueller.

What is the relationship with the conditional means?

```
> lm(Y~D, data = df)

Call:
lm(formula = Y ~ D, data = df)

Coefficients:
(Intercept)            D
    0.57425     -0.01782
```

- How do we interpret the coefficient on D?

- How do we interpret the coefficient on D?
- The intercept?

Credit to Andy Eggers...

- The *fitted* regression can be written as

$$\hat{Y}_i = .574 - .018 D_i$$

Credit to Andy Eggers...

- The *fitted* regression can be written as

$$\hat{Y}_i = .574 - .018 D_i$$

- We can express the conditional means as:

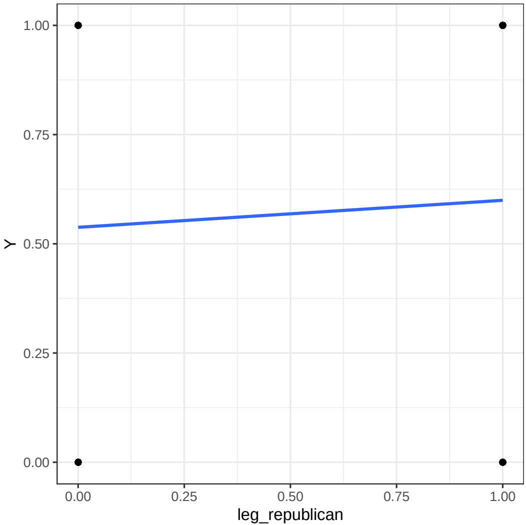
$$\hat{Y} = \begin{cases} .574 & D_i = 0 \\ .574 - .018 & D_i = 1 \end{cases}$$

Credit to Andy Eggers…

```
> lm(reply_atall ~ leg_party, data = df)
Call:
lm(formula = reply_atall ~ leg_party, data = df)

Coefficients:
(Intercept)   leg_partyR
    0.53775      0.06179
> lm(reply_atall ~ leg_party - 1, data = df)
Call:
lm(formula = reply_atall ~ leg_party - 1, data = df)

Coefficients:
leg_partyD  leg_partyR
    0.5377      0.5995
```


1.00
0.75
0.50
0.25
0.00
Y
0.00
0.25
0.50
0.75
1.00
leg_republican

Minimizing the sum of squared distances ...

```
> table(df$Y[which(df$leg_republican == 0)])

   0    1
1243 1446

> mean(df$Y[which(df$leg_republican == 0)])

[1] 0.5377464

> table(df$Y[which(df$leg_republican == 1)])

   0    1
 869 1301

> mean(df$Y[which(df$leg_republican == 1)])

[1] 0.5995392

>
```

Minimizing the sum of squared distances . . .

```
> table(df$Y[which(df$leg_republican == 0)])

   0    1
1243 1446

> mean(df$Y[which(df$leg_republican == 0)])

[1] 0.5377464

> table(df$Y[which(df$leg_republican == 1)])

   0    1
 869 1301

> mean(df$Y[which(df$leg_republican == 1)])

[1] 0.5995392

>
```

. . . reproduces exactly the conditional means with a binary independent variable.

Extracting components from an lm object

```
> lm1 <- lm(Y ~ D, data = df)
> names(lm1)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.resi
 [9] "xlevels"       "call"          "terms"         "model"
> summary(lm1)
Call:
lm(formula = Y ~ D, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5743 -0.5564  0.4258  0.4436  0.4436

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.57425    0.01005  57.114   <2e-16 ***
D           -0.01782    0.01422  -1.253     0.21
```

References I

Butler, D. M. and Broockman, D. E. (2011). Do politicians racially discriminate against constituents? a field experiment on state legislators. American Journal of Political Science, 55(3):463–477.