Session 2: Regressions

Welcome!

TA: Jennifer Lin

MMSS 211: Institutions, Rules, & Models in Social Science

2022-04-05

Goals

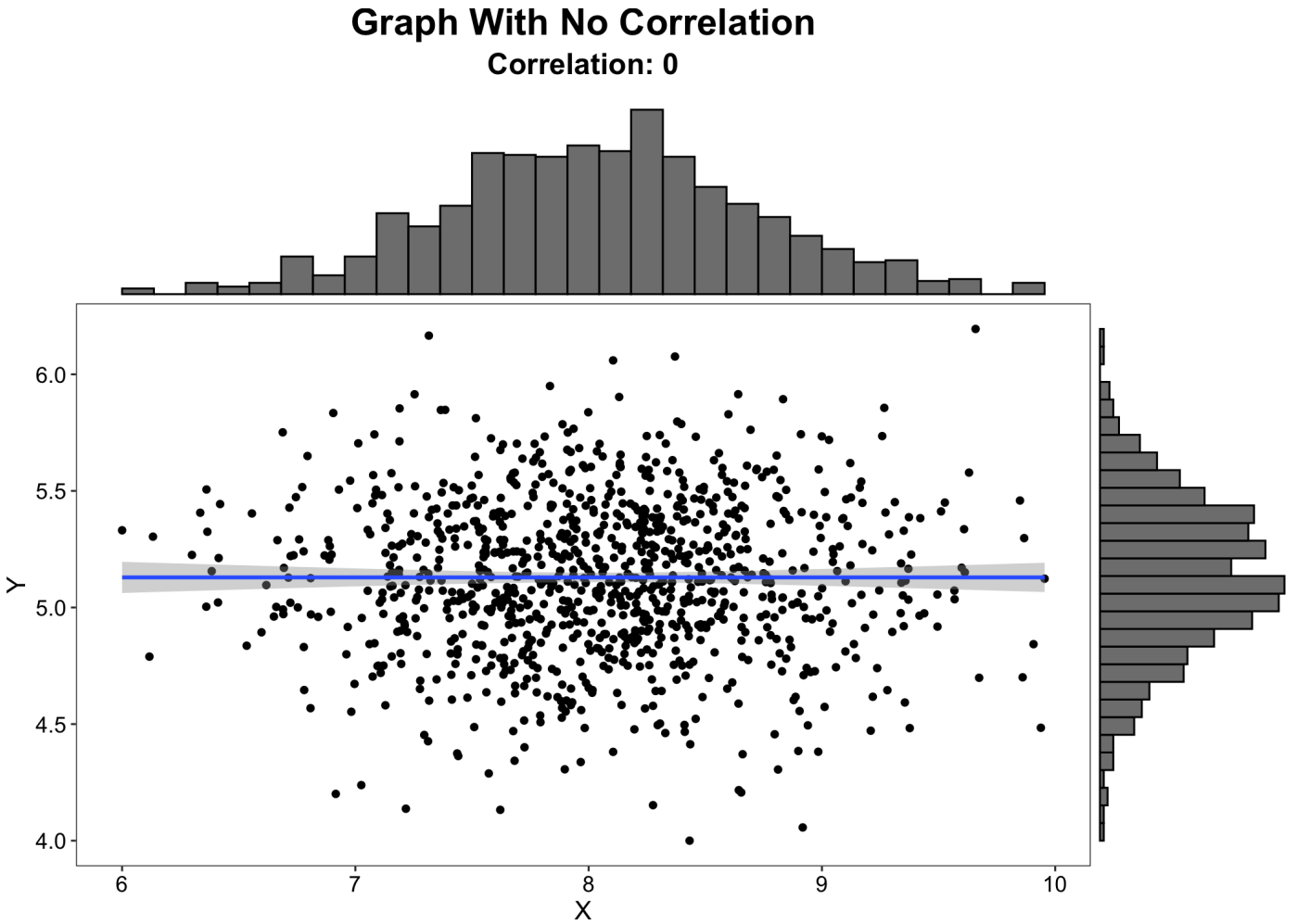
1. Reinforce understanding of the basics of linear regression
2. Learn to conduct regressions in R
3. Learn ways to effectively display regression results

The Fundimentals of Regression

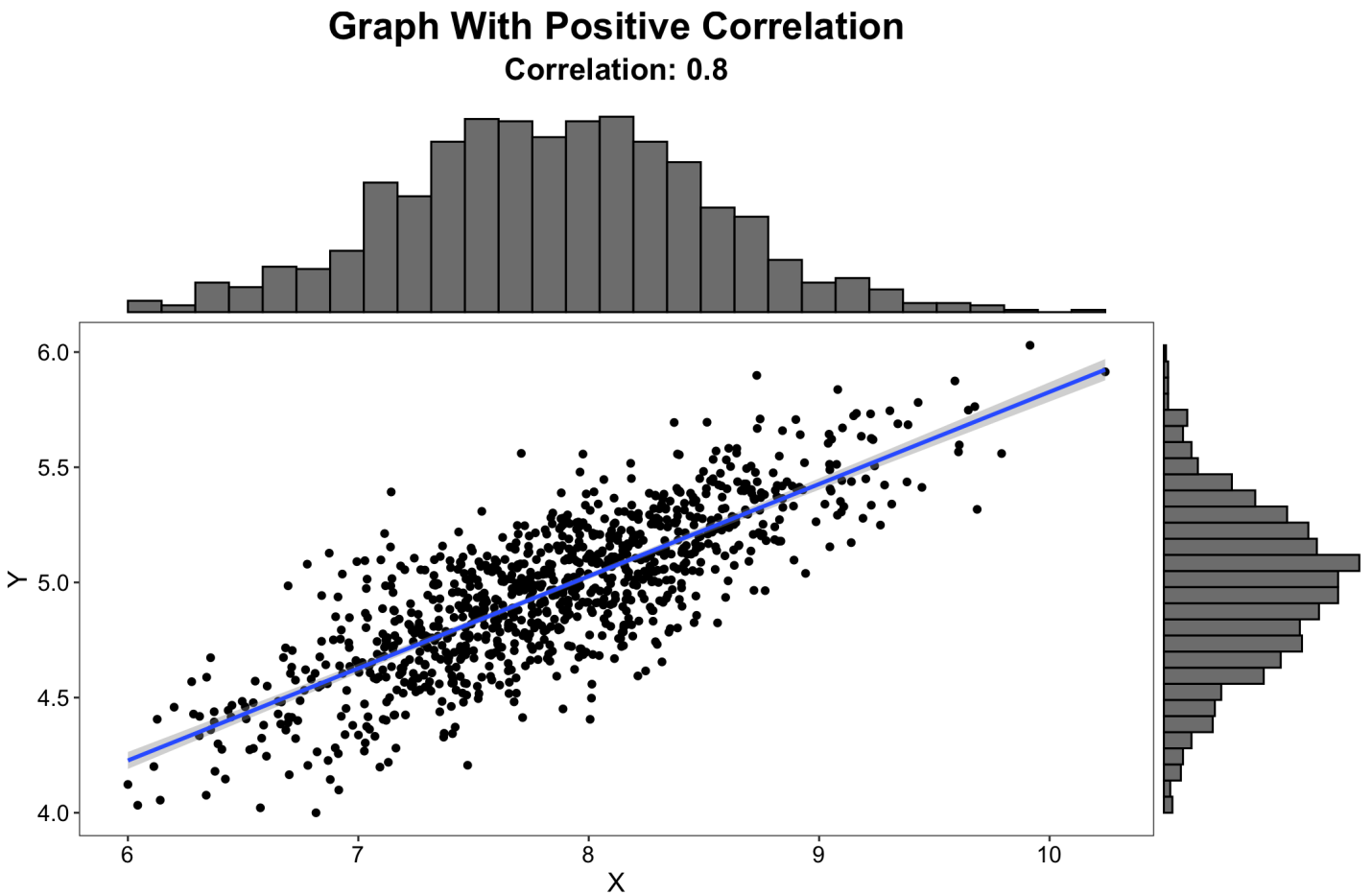
What is a Regression

- **O**rdinary **L**east **S**quares (OLS) is the **B**est **L**inear **U**nbiased **E**stimator
- Regression is a line that sums up the relationship between X and Y

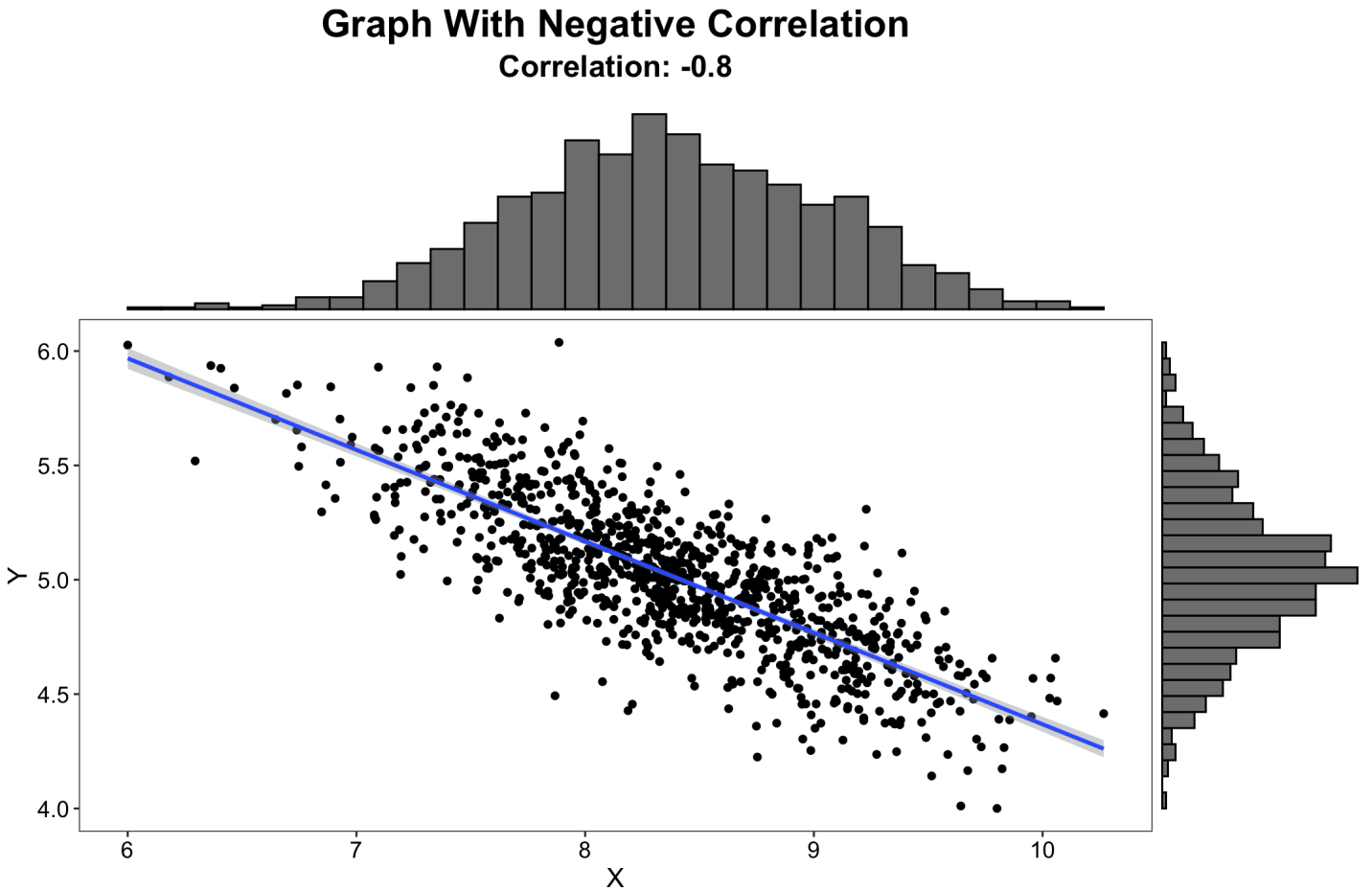
No Correlation



Positive Correlation



Negative Correlation



The Math Behind the Regression

Recall:

$$Y = mX + b$$

The Regression:

$$y = \alpha + \beta X + \epsilon$$

or

$$y = a + bX$$

Multiple Regressions

We can control for any number of other variables that we think might have an effect on the Y variable of interest

It is "subtracting out" the effect of X_2 on Y.

In Math:

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

Purpose of Regressions

1. For **summarizing** data: Calculates one line that describes the relationship between two variables the best and gives us a statistic that summarizes this relationship. This is the *coefficient*.
2. For making **inferences**: The line provides an average for the relationship for X and Y and we can use it to predict Y values from other X values that might not already be in the dataset.

Interpreting Regressions

Unit	Meaning
Y	Outcome for any X (input)
β	Slope (rise/run)
α	Y-intercept, where X is 0
ε	Error term

β is also known as the **coefficient**.

A 1-unit increase in X is associated with a coefficient-sized change in Y

Regressions in Practice

Data

The American National Elections Study is fielded every four years to assess American attitudes towards political matters during Presidential Election years

- The data include items on demographics (race, gender, age) along with an assortment of feeling thermometer variables.
- Feeling Thermometer variables start with `FT_` and all range from 0-100.

```
ANES <- read.csv("ANES_2020.csv")
```

A Preview of Examples

1. Continuous IV and Continuous DV
2. Continuous IV and Continuous DV with continuous controls
3. Continuous DV with categorical IV
4. Continuous DV with categorical IV and categorical controls

Continuous IV and Continuous DV

Unit	Variable	Meaning
X_1	FT_Fauci	Feelings towards Anthony Fauci
Y	FT_CDC	Feelings towards the CDC

$$Y_{CDC} = \alpha + \beta_{Fauci} X_{Fauci} + \epsilon$$

```
model1 <- lm(
  FT_CDC ~ FT_Fauci,
  data = ANES)

model1
```

```
##
## Call:
## lm(formula = FT_CDC ~ FT_Fauci, data = ANES)
##
## Coefficients:
## (Intercept)     FT_Fauci
##     38.5365       0.4642
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = FT_CDC ~ FT_Fauci, data = ANES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.953 -11.745   0.047  13.972  61.464
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.536458   0.556954   69.19   <2e-16 ***
## FT_Fauci     0.464165   0.007464   62.19   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.05 on 7151 degrees of freedom
##   (1127 observations deleted due to missingness)
## Multiple R-squared:  0.351,    Adjusted R-squared:  0.3509
## F-statistic:  3868 on 1 and 7151 DF,  p-value: < 2.2e-16
```

Continuous IV and Continuous DV with continuous controls

Unit	Variable	Meaning
X_1	`FT_Fauci`	Feelings towards Anthony Fauci
X_2	`age`	Age
X_3	`pid7`	Party ID (7-Category)
Y	`FT_CDC`	Feelings towards the CDC

$$Y_{CDC} = \alpha + \beta_{Fauci} X_{Fauci} + \beta_{age} X_{age} + \beta_{pid7} X_{pid7} + \epsilon$$

```
model2 <- lm(
  FT_CDC ~ FT_Fauci + age + pid7,
  data = ANES)
```

```
##
## Call:
## lm(formula = FT_CDC ~ FT_Fauci + age + pid7, data = ANES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.978 -11.419   0.388  13.914  61.848
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.160563   1.101122  38.289  < 2e-16 ***
## FT_Fauci     0.440590   0.009236  47.703  < 2e-16 ***
## age          0.006668   0.013817   0.483    0.629
## pid7        -0.594624   0.121897  -4.878  1.1e-06 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.03 on 6870 degrees of freedom
##   (1406 observations deleted due to missingness)
## Multiple R-squared:  0.3539,    Adjusted R-squared:  0.3536
## F-statistic:  1254 on 3 and 6870 DF,  p-value: < 2.2e-16
```

Interpreting a Regression Table

Components of a Table

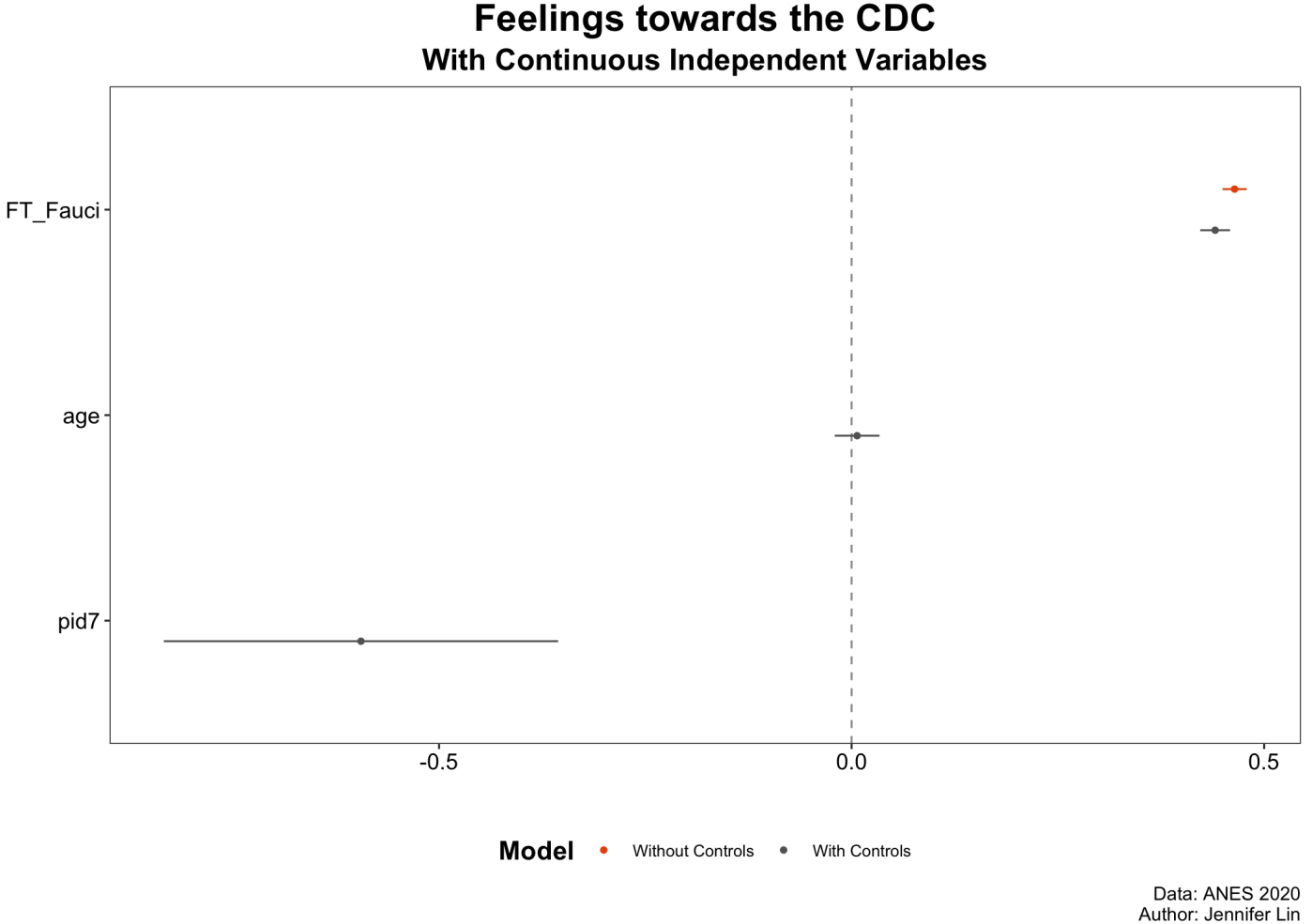
Unit	Meaning
Coefficients	β values for the regression
(Intercept)	Y-intercept, where X is 0
Estimate	Slope
Std. Error	Standard Error
t value	T Statistic -- standard score
p-value	Probability of getting result by chance
Signif. Codes	Asterisks that symbolize "rarity"
R^2	Variation in Y explained by X
F	Model performance measure

```
modelsummary(
  list(model1, model2),
  estimate = "{estimate} ({std.error}){stars}",
  statistic = NULL,
  output = "kableExtra") %>%
  kable_styling(font_size = 14)
```

	Model 1	Model 2
(Intercept)	38.536 (0.557)***	42.161 (1.101)***
FT_Fauci	0.464 (0.007)***	0.441 (0.009)***
age		0.007 (0.014)
pid7		−0.595 (0.122)***
Num.Obs.	7153	6874
R2	0.351	0.354
R2 Adj.	0.351	0.354
AIC	62463.2	60018.2
BIC	62483.8	60052.4
Log.Lik.	−31228.612	−30004.090
F	3867.526	1254.197
RMSE	19.05	19.03

Plotting the Results

```
dwplot(list(model1, model2),
       vline = geom_vline(
         xintercept = 0, colour = "grey60", linetype = 2))+
  scale_color_manual(
    values = c("#636363", "#e6550d"),
    name = "Model",
    labels = c("With Controls", "Without Controls")
  )+
  labs(
    title = "Feelings towards the CDC",
    subtitle = "With Continuous Independent Variables",
    caption = "Data: ANES 2020
    Author: Jennifer Lin"
  )+
  theme_bw()+
  theme(
    title      = element_text(colour="black"),
    plot.title = element_text(size = 20, hjust = 0.5, face = 'bold
    plot.subtitle = element_text(size = 18, hjust = 0.5),
    legend.position = 'bottom'
  )
```


Feelings towards the CDC
With Continuous Independent Variables
FT_Fauci
age
pid7
-0.5
0.0
0.5
Model
Without Controls
With Controls
Data: ANES 2020
Author: Jennifer Lin

Continuous DV with categorical IV

Unit	Variable	Meaning
X_1	`PARTY`	Party ID (3-Category)
Y	`FT_Feminists`	Feelings towards Feminists

$$Y_{Feminists} = \alpha + \beta_{PARTY} X_{PARTY} + \epsilon$$

```
model3 <- lm(
  FT_Feminists ~ PARTY,
  data = ANES)
```

```
##
## Call:
## lm(formula = FT_Feminists ~ PARTY, data = ANES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.040 -13.729   3.972  13.972  56.271
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        73.0397     0.3922  186.24   <2e-16 ***
## PARTYIndependent -17.0114     0.8796  -19.34   <2e-16 ***
## PARTYRepublican  -29.3109     0.5730  -51.15   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.97 on 7301 degrees of freedom
##   (976 observations deleted due to missingness)
## Multiple R-squared:  0.265,    Adjusted R-squared:  0.2648
## F-statistic:  1316 on 2 and 7301 DF,  p-value: < 2.2e-16
```

Continuous DV with categorical IV and categroical controls

Unit	Variable	Meaning
X_1	PARTY	Party ID (3-Category)
X_2	FEMALE	Gender: Female
X_3	MINORITY	Race: Minority
X_4	VOTED_2020	Voted in 2020
Y	FT_Feminists	Feelings towards the Feminists

$$Y_{Feminists} = \alpha + \beta_{PARTY}X_{PARTY} + \beta_{FEMALE}X_{FEMALE} + \beta_{MINORITY}X_{MINORITY} + \beta_{voted}X_{voted} + \epsilon$$

```
model4 <- lm(
  FT_Feminists ~ PARTY + FEMALE + MINORITY + Voted_2020,
  data = ANES)
```

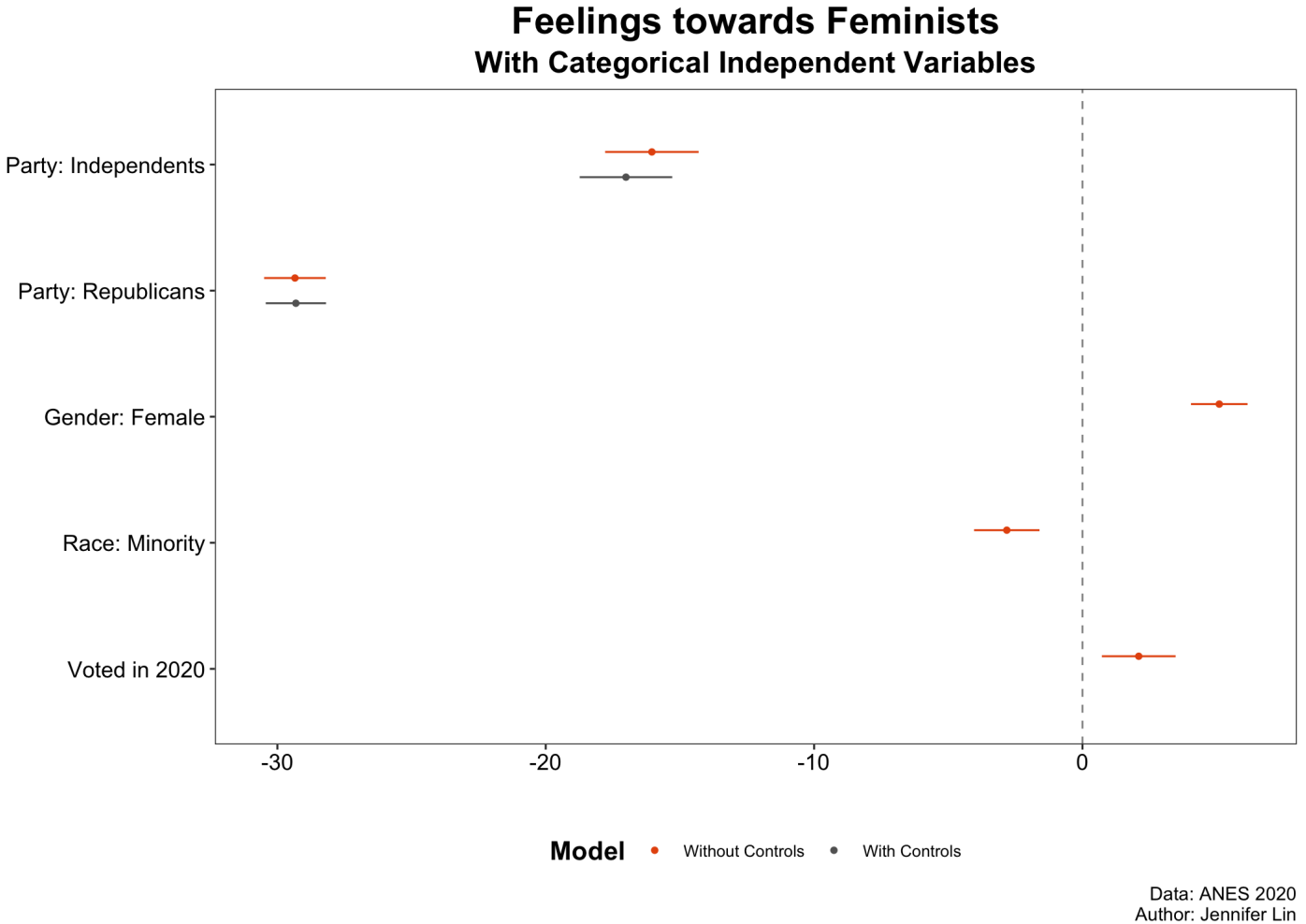
```
##
## Call:
## lm(formula = FT_Feminists ~ PARTY + FEMALE + MINORITY + Voted_2020,
##     data = ANES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.437 -14.271   2.909  13.659  62.922
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        69.2445     0.8187  84.574  < 2e-16 ***
## PARTYIndependent -16.0458     0.8885 -18.059  < 2e-16 ***
## PARTYRepublican  -29.3458     0.5861 -50.065  < 2e-16 ***
## FEMALETRUE         5.0966     0.5378   9.476  < 2e-16 ***
## MINORITYTRUE      -2.8207     0.6223  -4.533 5.92e-06 ***
## Voted_2020TRUE     2.0960     0.7009   2.991  0.00279 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.78 on 7298 degrees of freedom
##   (976 observations deleted due to missingness)
## Multiple R-squared:  0.2773,    Adjusted R-squared:  0.2768
## F-statistic:   560 on 5 and 7298 DF,  p-value: < 2.2e-16
```

```
modelsummary(
  list(model3, model4),
  estimate = "{estimate} ({std.error}){stars}",
  statistic = NULL,
  output = "kableExtra") %>%
  kable_styling(font_size = 14)
```

	Model 1	Model 2
(Intercept)	73.040 (0.392)***	69.245 (0.819)***
PARTYIndependent	−17.011 (0.880)***	−16.046 (0.889)***
PARTYRepublican	−29.311 (0.573)***	−29.346 (0.586)***
FEMALETRUE		5.097 (0.538)***
MINORITYTRUE		−2.821 (0.622)***
Voted_2020TRUE		2.096 (0.701)**
Num.Obs.	7304	7304
R2	0.265	0.277
R2 Adj.	0.265	0.277
AIC	66516.1	66398.9
BIC	66543.7	66447.1
Log.Lik.	−33254.074	−33192.435
F	1316.045	560.003
RMSE	22.97	22.78

Plotting the Results

```
dwplot(list(model3, model4),
       vline = geom_vline(
         xintercept = 0, colour = "grey60", linetype = 2)) %>%
  relabel_predictors(
    PARTYIndependent = "Party: Independents",
    PARTYRepublican = "Party: Republicans",
    FEMALETRUE = "Gender: Female",
    MINORITYTRUE = "Race: Minority",
    Voted_2020TRUE = "Voted in 2020"
  )+
  scale_color_manual(
    values = c("#636363", "#e6550d"),
    name = "Model",
    labels = c("With Controls", "Without Controls")
  )+
  labs(
    title = "Feelings towards Feminists",
    subtitle = "With Categorical Independent Variables",
    caption = "Data: ANES 2020
    Author: Jennifer Lin"
  )+
  theme_bw()+
```


Feelings towards Feminists
With Categorical Independent Variables
Party: Independents
Party: Republicans
Gender: Female
Race: Minority
Voted in 2020
-30
-20
-10
0
Model
Without Controls
With Controls
Data: ANES 2020
Author: Jennifer Lin

Exercise

1. Pick a continuous variable to serve as your dependent variable and pick a handful of variables that might serve as reasonable independent variables. Write their names down.
2. Write down a reasonable regression model you might run.
3. Conduct the regression.
4. Interpret the results. What does the coefficient for your main independent variable tell us about the relationship between that variable and the dependent variable?