Social Statistics

Regression With Categorical Variables

November 17, 2021

Warm Up

Using same data from Monday (`hdi.csv`), what would you expect the relationship to be between these two variables:

- Adolescent birth rate (`adolescent_birth_rate`) is the number of births per 1,000 women ages 15-19
- Female secondary education rate (`female_secondary_educ`) is the proportion of females in a country (ages 25 and older) with at least some secondary education

Regress the adolescent birth rate on the female secondary education rate

Warm Up

```
birthrate_seceduc_model <-
      lm(adolescent_birth_rate ~ female_secondary_educ,
            data = hdi)

summary(birthrate_seceduc_model)
```

Warm Up - Model

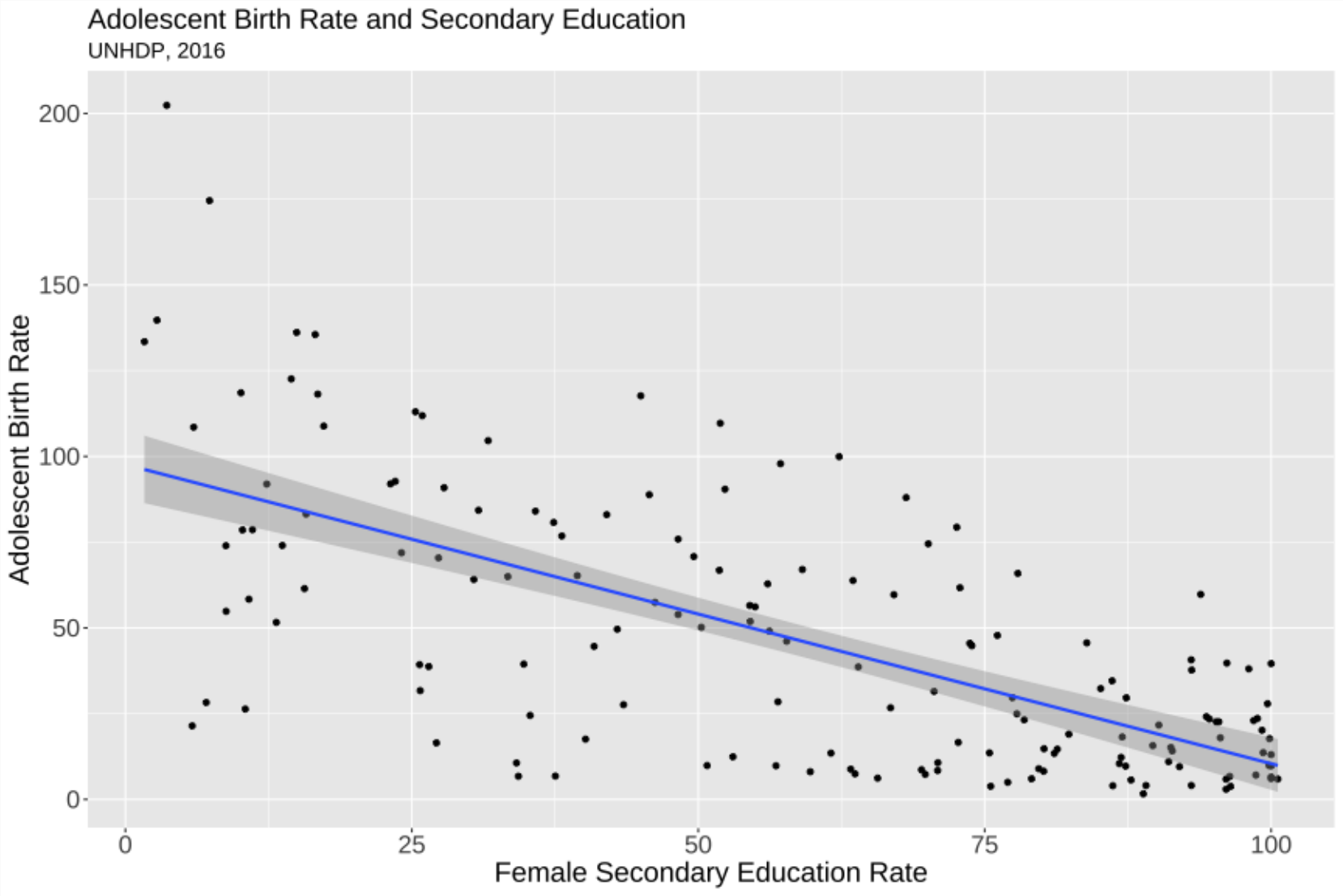
```
Call:
lm(formula = adolescent_birth_rate ~ female_secondary_educ, data = hdi)

Residuals:
    Min      1Q  Median      3Q     Max
-71.129 -17.494  -2.911  16.475 107.874

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           97.65697    5.08673   19.20   <2e-16 ***
female_secondary_educ -0.87232    0.07635  -11.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.11 on 157 degrees of freedom
Multiple R-squared:  0.454,    Adjusted R-squared:  0.4505
F-statistic: 130.5 on 1 and 157 DF,  p-value: < 2.2e-16
```

Warm Up - Model



Warm Up - Prediction

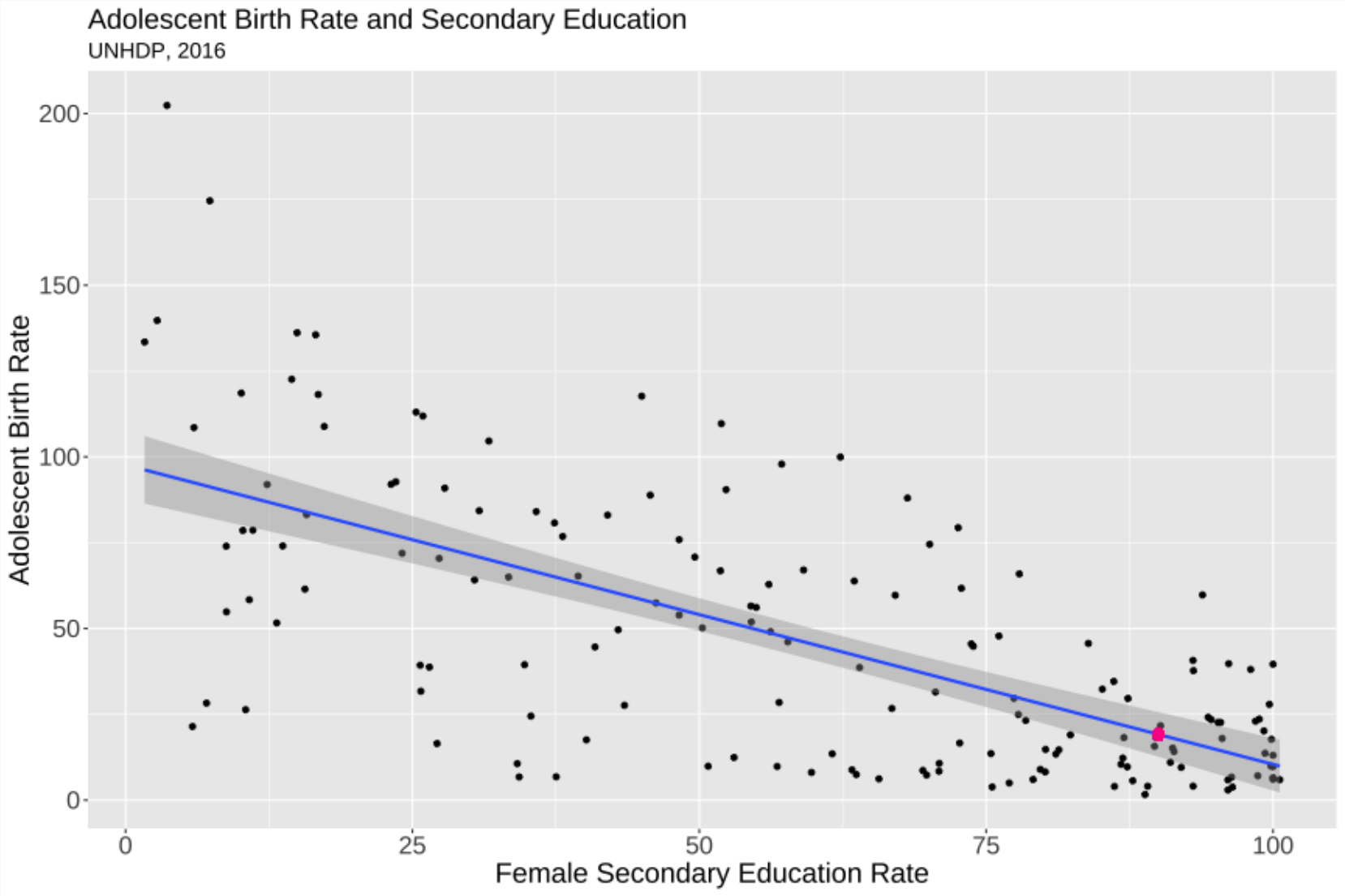
Find the predicted adolescent birth rate when 90% of female residents of a country complete some secondary education.

```
# Model: y = 97.65697 - 0.87232X

97.65697 - 0.87232*90
```

```
## [1] 19.14817
```

Warm Up - Plotting Prediction



Moving Forward With Regression

So far we have seen a basic model: a continuous dependent variable and one continuous independent variable

Today we'll extend the basic model to show how regression works with categorical variables, starting with binary (0/1) *independent* variables

Before regression, how would you find the mean adolescent birth rate for countries without high scores on the human development index (`hdi_rank_hi==0`) and for countries with high scores on the human development index (`hdi_rank_hi==1`)?

Moving Forward With Regression

One option would be `mean()` with indexing:

```
mean(hdi$adolescent_birth_rate[hdi$hdi_rank_hi==0])
```

```
## [1] 59.70372
```

```
mean(hdi$adolescent_birth_rate[hdi$hdi_rank_hi==1])
```

```
## [1] 12.90428
```

Moving Forward With Regression

May be more efficient to use `group_by()` and `summarize()`:

```
birthrates <- hdi  |>
  group_by(hdi_rank_hi) |>
  summarize(mean_birthrate = mean(adolescent_birth_rate))

birthrates
```

```
## # A tibble: 2 × 2
##   hdi_rank_hi mean_birthrate
##         <int>          <dbl>
## 1           0           59.7
## 2           1           12.9
```

What is the difference between the two means?

```
mean(hdi$adolescent_birth_rate[hdi$hdi_rank_hi==0]) -
     mean(hdi$adolescent_birth_rate[hdi$hdi_rank_hi==1])
```

```
## [1] 46.79945
```

Binary Independent Variables

Regressing `adolescent_birth_rate` on `hdi_rank_hi` will give us the exact same information

```
birthrate_rankhi_model <-
      lm(adolescent_birth_rate ~ hdi_rank_hi,
            data = hdi)

summary(birthrate_rankhi_model)
```

Binary Independent Variables

```
Call:
lm(formula = adolescent_birth_rate ~ hdi_rank_hi, data = hdi)

Residuals:
    Min      1Q  Median      3Q     Max
-53.518 -20.340  -3.219  14.579 142.680

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   59.704      3.121  19.132  < 2e-16 ***
hdi_rank_hi  -46.799      5.740  -8.154 1.05e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.03 on 157 degrees of freedom
Multiple R-squared:  0.2975,    Adjusted R-squared:  0.293
F-statistic: 66.48 on 1 and 157 DF,  p-value: 1.054e-13
```

Binary Independent Variables

As with continuous variables, the intercept - α - is the mean value of our dependent variable, Y, when our independent variable, X, is 0

- The mean for `adolescent_birth_rate` is 59.704 for the countries that are not in the high hdi group

Binary Independent Variables

The coefficient for our independent variable - β - is the difference in the mean of our dependent variable between cases with a 0 and 1 for our independent variable

- The mean for `adolescent_birth_rate` for the countries that are in the high hdi group is 46.799 points lower than the mean for `adolescent_birth_rate` for the countries that are not in the high hdi group

Same intuition as before: a one-unit increase in X is associated with a change in Y of β. But now that one-unit increase is moving from a value of 0 to 1 for the binary variable.

Binary Independent Variables

Before regression, how would you have tested if the difference in means is significant?

The t value and p value are for a t-test of the difference with one small change. OLS requires the assumption that the sample variances are equal:

```
t.test(hdi$adolescent_birth_rate[hdi$hdi_rank_hi==1],
       hdi$adolescent_birth_rate[hdi$hdi_rank_hi==0],
       var.equal = TRUE)
```

Binary Independent Variables

```
    Two Sample t-test

data:  hdi$adolescent_birth_rate[hdi$hdi_rank_hi == 1] and hdi$adolescent_birth_r
t = -8.1536, df = 157, p-value = 1.054e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -58.13648 -35.46241
sample estimates:
mean of x mean of y
 12.90428  59.70372
```

Same as t value and p value in regression output!

Binary Independent Variables

```
Call:
lm(formula = adolescent_birth_rate ~ hdi_rank_hi, data = hdi)

Residuals:
    Min      1Q  Median      3Q     Max
-53.518 -20.340  -3.219  14.579 142.680

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   59.704      3.121  19.132  < 2e-16 ***
hdi_rank_hi  -46.799      5.740  -8.154 1.05e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.03 on 157 degrees of freedom
Multiple R-squared:  0.2975,    Adjusted R-squared:  0.293
F-statistic: 66.48 on 1 and 157 DF,  p-value: 1.054e-13
```

Binary Independent Variables - Exercise

Use regression to test the significance of the difference in the maternal mortality ratio between countries where higher percentages of females than males completed some secondary education.

The maternal mortality ratio is the number of women who die per 100,000 live births; use the `maternal_mortality_ratio` variable. Create the education variable; call it `female_more_schl`.

```
hdi <- hdi |>
  mutate(female_more_schl =
           ifelse(female_secondary_educ > male_secondary_educ,1,0))
```

Binary Independent Variables - Exercise

```
mortality_morefemeduc_model <-
lm(maternal_mortality_ratio ~
      female_more_schl,
      data = hdi)

summary(mortality_morefemeduc_model)
```

Binary Independent Variables - Exercise

```
Call:
lm(formula = maternal_mortality_ratio ~ female_more_schl, data = hdi)

Residuals:
    Min      1Q  Median      3Q     Max
-178.30 -160.30  -60.11   64.39 1178.70

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        181.30      20.30   8.931 1.07e-15 ***
female_more_schl  -106.19      38.59  -2.752  0.00662 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 217.7 on 157 degrees of freedom
Multiple R-squared:  0.04601,    Adjusted R-squared:  0.03994
F-statistic: 7.573 on 1 and 157 DF,  p-value: 0.006624
```

Binary Independent Variables - Exercise

In words: Countries where there is not a higher percentage of females than males completing some secondary education (`female_more_schl==0`) have, on average, maternal mortality ratios of 181.30.

Countries where there is a higher percentage of females than males completing some secondary education (`female_more_schl==1`) have, on average, maternal mortality ratios 106.19 points lower than countries where higher percentages of males than females complete some secondary education.

This difference is significant (p-value = 0.007)

Categorical Independent Variables

What if the independent variable has more than one category?

One category becomes the *reference group*

The α is the average for that reference group

The coefficients are the differences in means for each category compared to the reference group

The t-test compares the differences in means to the null hypothesis that the real difference between the reference group and the given category is actually zero

Categorical Independent Variables

Let's create a new variable called `hdi_rank_cat` which has four categories of hdi_rank: 1-51, 52-106, 107-147, 148-188.

```
hdi <- mutate(hdi, hdi_rank_cat =
      ifelse(hdi_rank %in% 1:51, 1,
      ifelse(hdi_rank %in% 52:106, 2,
      ifelse(hdi_rank %in% 107:147, 3, 4))))
```

Note: Usually easier to create categories numerically to keep them in the right order. Add their labels/levels later.

Categorical Independent Variables

Try regressing the adolescent birth rate variable on this new hdi rank category variable

```
birthrate_hdicat_model1 <-
lm(adolescent_birth_rate ~
      hdi_rank_cat, data = hdi)

summary(birthrate_hdicat_model1)
```

Categorical Independent Variables

```
Call:
lm(formula = adolescent_birth_rate ~ hdi_rank_cat, data = hdi)

Residuals:
    Min      1Q  Median      3Q     Max
-65.812 -14.548  -0.705  14.528 110.236

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -16.863      4.740  -3.557 0.000495 ***
hdi_rank_cat   27.253      1.861  14.648  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.61 on 157 degrees of freedom
Multiple R-squared:  0.5774,    Adjusted R-squared:  0.5748
F-statistic: 214.5 on 1 and 157 DF,  p-value: < 2.2e-16
```

Categorical Independent Variables

That approach treats the hdi rank categories as a continuous variable. That could work okay when the mean difference between categories is equal (but that's a big assumption).

Preferable to have a separate coefficient comparing the mean for each category to the mean for the reference category.

When a variable has multiple categories, make sure R knows the variable is a factor variable.

There are two options for how to do this...

Categorical Independent Variables

Option 1: Use the variable as a factor just for this model:

```
birthrate_hdicat_model2 <-
lm(adolescent_birth_rate ~ factor(hdi_rank_cat),
      data = hdi)

summary(birthrate_hdicat_model2)
```

Categorical Independent Variables

```
Call:
lm(formula = adolescent_birth_rate ~ factor(hdi_rank_cat), data = hdi)

Residuals:
    Min      1Q  Median      3Q     Max
-69.833 -13.319  -3.356  13.624 106.215

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             12.904      3.731   3.459 0.000701 ***
factor(hdi_rank_cat)2   22.411      5.249   4.270 3.40e-05 ***
factor(hdi_rank_cat)3   48.017      5.809   8.266 5.84e-14 ***
factor(hdi_rank_cat)4   83.265      5.918  14.069  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.58 on 155 degrees of freedom
Multiple R-squared:  0.5839,    Adjusted R-squared:  0.5759
F-statistic: 72.51 on 3 and 155 DF,  p-value: < 2.2e-16
```

In Words

- The mean adolescent birth rate for countries in the reference group (category 1) is 12.9. This mean is significantly different from zero (p-value = 0.001).

- The mean adolescent birth rate for countries in category 2 is 22.41 points higher than the mean for countries in category 1. This difference is significant (p-value < 0.001).

In Words

- The mean adolescent birth rate for countries in category 3 is 48.02 points higher than the mean for countries in category 1. This difference is significant (p-value < 0.001).
- The mean adolescent birth rate for countries in category 4 is 83.27 points higher than the mean for countries in category 1. This difference is significant (p-value < 0.001).

Categorical Independent Variables

Option 2: Make the variable a factor before setting up your model:

```
# This is all you need...

hdi <- mutate(hdi, hdi_rank_cat = factor(hdi_rank_cat))
```

```
# But you might want to also add labels...

hdi <- mutate(hdi, hdi_rank_cat = factor(hdi_rank_cat,
                labels = c("Very High", "High", "Medium" ,"Low")))
```

Categorical Independent Variables

Then use your factor variable in the model (without the need to restate that it is a factor variable)

```
birthrate_hdicat_model3 <-
lm(adolescent_birth_rate ~ hdi_rank_cat,
      data = hdi)

summary(birthrate_hdicat_model3)
```

Categorical Independent Variables

```
Call:
lm(formula = adolescent_birth_rate ~ hdi_rank_cat, data = hdi)

Residuals:
    Min      1Q  Median      3Q     Max
-69.833 -13.319  -3.356  13.624 106.215

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          12.904      3.731   3.459 0.000701 ***
hdi_rank_catHigh     22.411      5.249   4.270 3.40e-05 ***
hdi_rank_catMedium   48.017      5.809   8.266 5.84e-14 ***
hdi_rank_catLow      83.265      5.918  14.069  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.58 on 155 degrees of freedom
Multiple R-squared:  0.5839,    Adjusted R-squared:  0.5759
F-statistic: 72.51 on 3 and 155 DF,  p-value: < 2.2e-16
```

Categorical Independent Variables

Same estimates, but the labels may help with the interpretation

Remember, the intercept is the mean value of the dependent variable for the reference group. In this example, that is the `very high` hdi group.

The coefficients compare the mean values of the dependent variable between the reference category and each other level of the independent variable

Use this approach with any categorical variable: race, class, religion, region, degree, marital status, etc.

Categorical Independent Variables

By default, the first category becomes the reference group. Good practice is to use the group with the most cases as your reference group.

If you want to change the reference group, use `relevel()`. In this example, we will make the "Low" category the reference group:

```
hdi$hdi_rank_cat <-
      relevel(hdi$hdi_rank_cat, ref = "Low")

birthrate_hdi_model4 <-
lm(adolescent_birth_rate ~
      hdi_rank_cat,
      data = hdi)

summary(birthrate_hdi_model4)
```

Categorical Independent Variable

```
Call:
lm(formula = adolescent_birth_rate ~ hdi_rank_cat, data = hdi)

Residuals:
    Min      1Q  Median      3Q     Max
-69.833 -13.319  -3.356  13.624 106.215

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             96.169      4.594  20.933  < 2e-16 ***
hdi_rank_catVery High  -83.265      5.918 -14.069  < 2e-16 ***
hdi_rank_catHigh       -60.854      5.894 -10.325  < 2e-16 ***
hdi_rank_catMedium     -35.248      6.398  -5.509 1.47e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.58 on 155 degrees of freedom
Multiple R-squared:  0.5839,    Adjusted R-squared:  0.5759
F-statistic: 72.51 on 3 and 155 DF,  p-value: < 2.2e-16
```

Compare To...

```
Call:
lm(formula = adolescent_birth_rate ~ hdi_rank_cat, data = hdi)

Residuals:
    Min      1Q  Median      3Q     Max
-69.833 -13.319  -3.356  13.624 106.215

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         12.904      3.731   3.459 0.000701 ***
hdi_rank_catHigh    22.411      5.249   4.270 3.40e-05 ***
hdi_rank_catMedium  48.017      5.809   8.266 5.84e-14 ***
hdi_rank_catLow     83.265      5.918  14.069  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.58 on 155 degrees of freedom
Multiple R-squared:  0.5839,    Adjusted R-squared:  0.5759
F-statistic: 72.51 on 3 and 155 DF,  p-value: < 2.2e-16
```

Binary Dependent Variables

Let's switch to a categorical *dependent* variable

Key point: OLS can only handle a categorical dependent variable that is binary (two categories, 0 and 1)

Like a model with a continuous dependent variable, an OLS model with a binary dependent variable estimates the mean of that dependent variable

- The mean of a binary variable is the probability of having a 1 for that variable, so this is called a *linear probability model*

A good reminder to think about what variables would make sense as binary dependent variables.

Binary Dependent Variables

Let's go back to our variable identifying countries where higher percentages of females than males completed some secondary education (`female_more_schl`). Regress this measure on `schooling_mean`.

```
femalemore_schooling_model1 <-
lm(female_more_schl ~
      schooling_mean,
      data = hdi)

summary(femalemore_schooling_model1)
```

Binary Dependent Variables

```
Call:
lm(formula = female_more_schl ~ schooling_mean, data = hdi)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4086 -0.3344 -0.2093  0.6181  0.8285

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.04113    0.10279   0.400   0.6896
schooling_mean  0.02749    0.01127   2.438   0.0159 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4419 on 157 degrees of freedom
Multiple R-squared:  0.03648,    Adjusted R-squared:  0.03035
F-statistic: 5.945 on 1 and 157 DF,  p-value: 0.01588
```

In Words

Each additional year in a country's mean years of schooling is associated with an increase of 2.75 percentage points in the probability that females attend secondary school at higher rates than males, on average.

Categorical Dep *And* Ind Variables

Regress `female_more_schl` on `hdi_rank_cat`

```
femalemore_hdi_model2 <-
lm(female_more_schl ~
      hdi_rank_cat,
      data = hdi)

summary(femalemore_hdi_model2)
```

Categorical Dep *And* Ind Variables

```
##
## Call:
## lm(formula = female_more_schl ~ hdi_rank_cat, data = hdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4167 -0.3030 -0.2766  0.5833  0.9677
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.03226    0.07770   0.415 0.678593
## hdi_rank_catVery High  0.24434    0.10010   2.441 0.015771 *
## hdi_rank_catHigh       0.38441    0.09968   3.856 0.000168 ***
## hdi_rank_catMedium     0.27077    0.10821   2.502 0.013373 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4326 on 155 degrees of freedom
## Multiple R-squared:  0.08847,    Adjusted R-squared:  0.07083
## F-statistic: 5.015 on 3 and 155 DF,  p-value: 0.002404
```

In Words...

In three percent of the countries in the "Low" hdi category, higher percentages of females than males complete some secondary education.

In Words...

In twenty-seven percent (.03 + .24) of the countries in the "Very High" hdi category, higher percentages of females than males complete some secondary education. This percentage is significantly higher than the percentage of countries in the "Low" hdi category with this outcome (p-value = 0.016).

In Words...

In forty-one percent (.03 + .38) of the countries in the "High" hdi category, higher percentages of females than males complete some secondary education. This percentage is significantly higher than the percentage of countries in the "Low" hdi category with this outcome (p-value < 0.001).

In Words...

In thirty percent (.03 + .27) of the countries in the "Medium" hdi category, higher percentages of females than males complete some secondary education. This percentage is significantly higher than the percentage of countries in the "Low" hdi category with this outcome (p-value = 0.013).

Categorical Dep *And* Ind Variables

Scatterplots work better when both the dependent and independent variables are continuous

Stick with reporting a table or a simple barplot for a binary dependent variable

Next week we'll see how categorical independent variables affect plots

Final Project

Time to choose a research question!

If you have a dataset in mind, let's chat

If you do not have data you want to use, start with the GSS

Choose one independent variable: age or educ (in years)

Choose one control variable: sex, race, marital status, class, or religion

Final Project

Find a continuous dependent variable using the `gssr` package or the GSS website: https://gssdataexplorer.norc.org

- Could be a scale
- The variable should be included in at least one of the 2010-2018 surveys
- Your next assignment is to find your variable and come up with a research question.