Social Statistics

Regression With Categorical Variables

November 16, 2023

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 (`secondary_educ_female`) is the percentage 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 ~ secondary_educ_female,
            data = hdi)

summary(birthrate_seceduc_model)
```

Warm Up - Model

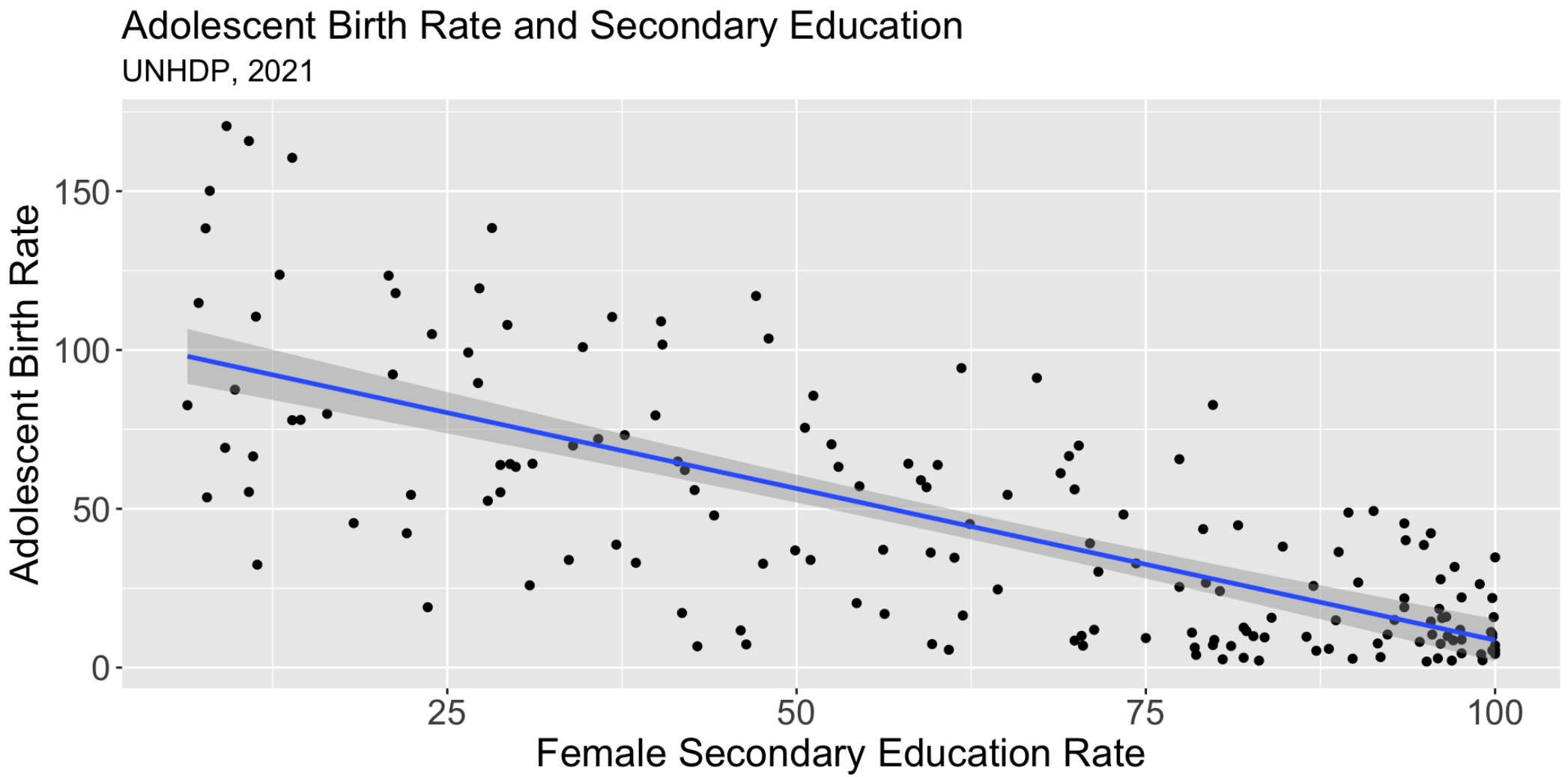
```
Call:
lm(formula = adolescent_birth_rate ~ secondary_educ_female, data = hdi)

Residuals:
    Min      1Q  Median      3Q     Max
-62.581 -15.571  -2.331  16.537  75.170

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           104.11409    4.76590   21.85   <2e-16 ***
secondary_educ_female  -0.95477    0.06922  -13.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.87 on 168 degrees of freedom
Multiple R-squared:  0.531, Adjusted R-squared:  0.5282
```

Warm Up - Model



Warm Up - Prediction

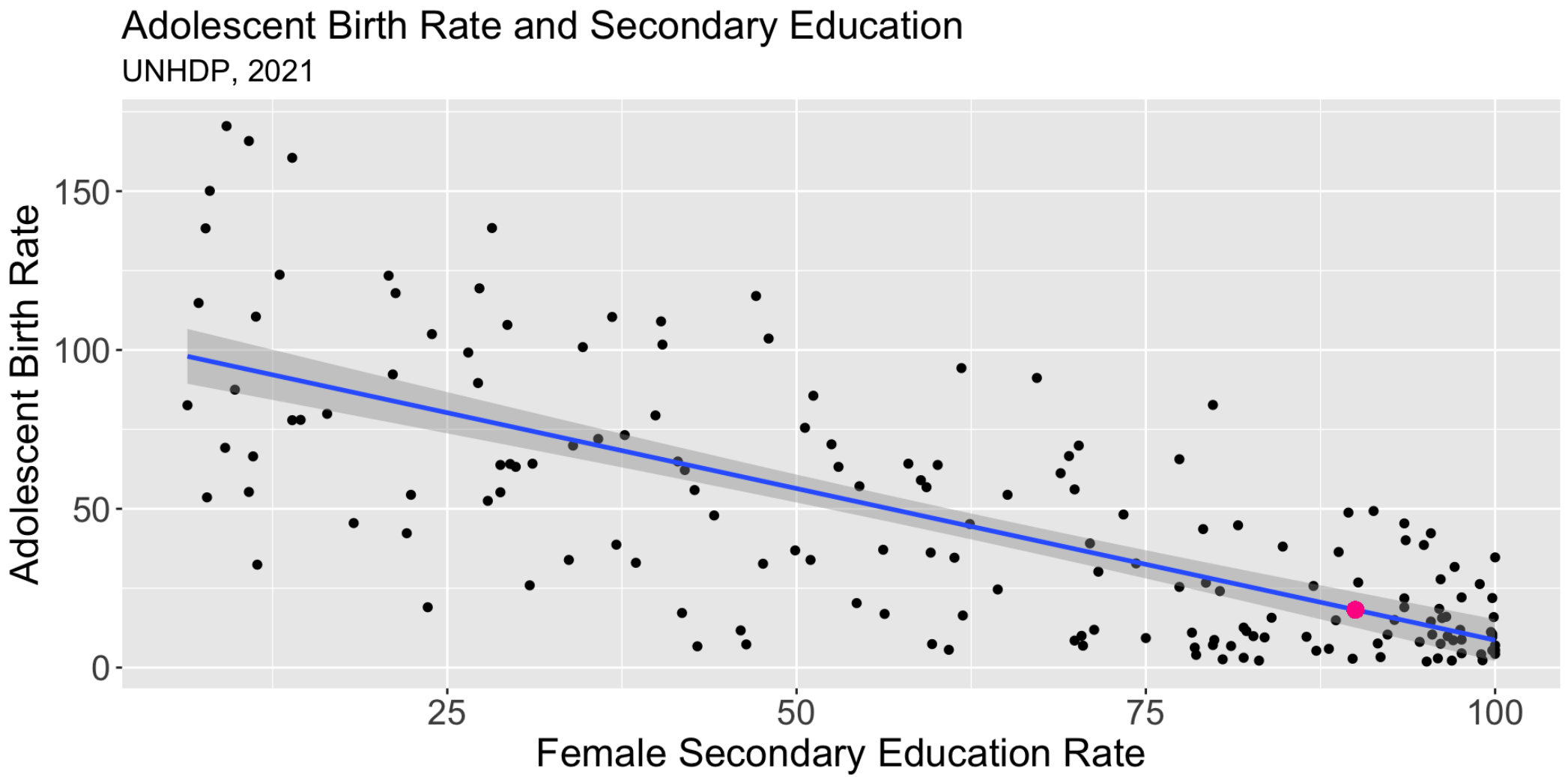
- Find the predicted adolescent birth rate when 90% of female residents of a country complete some secondary education.

```
# Model: y = 104.11409 - 0.95477 X

104.11409 - 0.95477 * 90
```

```
[1] 18.18479
```

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 compare the mean adolescent birth rate for countries in the highest category of human development (`hdi` >= 80) and the mean adolescent birth rate for all other countries?

Moving Forward With Regression

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

```
1 mean(hdi$adolescent_birth_rate[hdi$hdi>=80])
```

```
[1] 13.50968
```

```
1 mean(hdi$adolescent_birth_rate[hdi$hdi<80])
```

```
[1] 62.83333
```

Moving Forward With Regression

- May be more efficient to create a binary variable, and then use `group_by()` and `summarise()`:

```r
# To create binary variable...

hdi <- hdi  |>
  mutate(hdi_rank_hi = ifelse(hdi >= 80, 1, 0))
```

```r
# Get means for high and low hdi countries...

hdi |>
  group_by(hdi_rank_hi) |>
  summarise(mean_birthrate = mean(adolescent_birth_rate))
```

```
# A tibble: 2 × 2
  hdi_rank_hi mean_birthrate
        <dbl>          <dbl>
1           0           62.8
2           1           13.5
```

- What is the difference between the two means?

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

```
[1] 49.32366
```

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
-56.13 -14.48  -4.46  12.55 107.67

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.833      2.996  20.970   <2e-16 ***
hdi_rank_hi  -49.324      4.962  -9.941   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.14 on 168 degrees of freedom
Multiple R-squared:  0.3704,    Adjusted R-squared:  0.3666
```

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 62.833 for the countries that are not in the high hdi group

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

Binary Independent Variables

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?
- In the OLS output, 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:

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

```
    Two Sample t-test

data:  hdi$adolescent_birth_rate[hdi$hdi_rank_hi == 1] and
hdi$adolescent_birth_rate[hdi$hdi_rank_hi == 0]
t = -9.941, df = 168, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -59.11881 -39.52850
sample estimates:
```

Binary Independent Variables

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

Residuals:
   Min     1Q Median     3Q    Max
-56.13 -14.48  -4.46  12.55 107.67

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.833      2.996  20.970   <2e-16 ***
hdi_rank_hi  -49.324      4.962  -9.941   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.14 on 168 degrees of freedom
Multiple R-squared:  0.3704,    Adjusted R-squared:  0.3666
```

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(secondary_educ_female > secondary_educ_male,1,0))
```

Binary Independent Variables - Exercise

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

summary(mortality_morefemeduc_model)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-183.87 -168.87  -52.36   50.64  964.13

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        185.87      19.98   9.301  < 2e-16 ***
female_more_schl  -125.02      41.72  -2.997  0.00314 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 228.7 on 168 degrees of freedom
Multiple R-squared:  0.05074,   Adjusted R-squared:  0.04509
```

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 185.87.
- 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 125.02 points lower than countries where higher percentages of males than females complete some secondary education.
- This difference is significant (p-value = 0.003)

Categorical Independent Variables

- What if the independent variable has more than one category?
- One category becomes the *reference group* and 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_cat` which has four categories of `hdi`: 80-100, 70-79, 55-69, 55 or below.

```
hdi <- mutate(hdi, hdi_cat =
       ifelse(hdi >= 80, 1,
       ifelse(hdi >= 70 & hdi < 80, 2,
       ifelse(hdi >= 55 & hdi < 70, 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 category variable

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

summary(birthrate_hdicat_model1)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-62.722 -11.416  -2.774  12.633  75.378

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -16.605      4.013  -4.138 5.54e-05 ***
hdi_cat       27.932      1.627  17.165  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 23.65 on 168 degrees of freedom
Multiple R-squared:  0.6369,    Adjusted R-squared:  0.6347
```

Categorical Independent Variables

- That approach treats the hdi categories variable 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_cat), # Asserts hdi_cat is a factor
      data = hdi)
```

```
summary(birthrate_hdicat_model2)
```

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

Residuals:
   Min     1Q Median     3Q    Max
-66.18 -10.86  -3.31  11.71  73.22

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        13.510      2.998   4.507 1.24e-05 ***
factor(hdi_cat)2   21.776      4.717   4.616 7.79e-06 ***
factor(hdi_cat)3   51.671      4.946  10.447  < 2e-16 ***
factor(hdi_cat)4   85.074      5.250  16.206  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In Words

- The mean adolescent birth rate for countries in the reference group (category 1) is 13.510. This mean is significantly different from zero (p-value < .05).
- The mean adolescent birth rate for countries in category 2 is 21.776 points higher than the mean for countries in category 1. This difference is significant.
- The mean adolescent birth rate for countries in category 3 is 51.671 points higher than the mean for countries in category 1. This difference is significant.
- The mean adolescent birth rate for countries in category 4 is 85.074 points higher than the mean for countries in category 1. This difference is significant.

Categorical Independent Variables

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

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

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

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

hdi <- mutate(hdi, hdi_cat = factor(hdi_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)

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

summary(birthrate_hdicat_model3)
```

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

Residuals:
   Min     1Q Median     3Q    Max
-66.18 -10.86  -3.31  11.71  73.22

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     13.510      2.998   4.507 1.24e-05 ***
hdi_catHigh     21.776      4.717   4.616 7.79e-06 ***
hdi_catMedium   51.671      4.946  10.447  < 2e-16 ***
hdi_catLow      85.074      5.250  16.206  < 2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

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

Categorical Independent Variables

```
1 summary(birthrate_hdi_model4)
```

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

Residuals:
   Min     1Q Median     3Q    Max
-66.18 -10.86  -3.31  11.71  73.22

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        98.583      4.310  22.876  < 2e-16 ***
hdi_catVery High  -85.074      5.250 -16.206  < 2e-16 ***
hdi_catHigh       -63.298      5.642 -11.218  < 2e-16 ***
hdi_catMedium     -33.403      5.835  -5.724 4.75e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compare To...

```
1 summary(birthrate_hdicat_model3)
```

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

Residuals:
   Min     1Q Median     3Q    Max
-66.18 -10.86  -3.31  11.71  73.22

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     13.510      2.998   4.507 1.24e-05 ***
hdi_catHigh     21.776      4.717   4.616 7.79e-06 ***
hdi_catMedium   51.671      4.946  10.447  < 2e-16 ***
hdi_catLow      85.074      5.250  16.206  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

  → Since the mean of a binary variable is the probability of having a 1 for that variable, 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)
```

Binary Dependent Variables

```
1 summary(femalemore_schooling_model1)
```

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

Residuals:
     Min       1Q   Median       3Q      Max 
-0.36336 -0.28966 -0.16858 -0.05473  0.84985 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.007784   0.093984  -0.083  0.93409
schooling_mean  0.026322   0.009816   2.682  0.00806 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4142 on 168 degrees of freedom
Multiple R-squared:  0.04105,   Adjusted R-squared:  0.03534 
```

Categorical Dep *And* Ind Variables

- Regress `female_more_schl` on `hdi_cat`

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

summary(femalemore_hdi_model2)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.45238 -0.22581 -0.13889 -0.03333  0.96667

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.03333    0.07297   0.457   0.6484
hdi_catVery High  0.19247    0.08889   2.165   0.0318 *
hdi_catHigh       0.41905    0.09554   4.386 2.04e-05 ***
hdi_catMedium     0.10556    0.09880   1.068   0.2869
---
```

In Words...

- In three percent of the countries in the “Low” hdi category, higher percentages of females than males complete some secondary education.
- In twenty-two percent (.03 + .19) 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 = 0.0318).

In Words...

- In forty-five percent (.03 + .42) 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 fourteen percent (.03 + .11) of the countries in the “Medium” hdi category, higher percentages of females than males complete some secondary education. This percentage is not significantly higher than the percentage of countries in the “Low” hdi category with this outcome (p-value = 0.2869).

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. Recall that we have seen GSS, Pew, UNHDP, Opportunity Insights, eviction lab data across cities, hurricanes, race and names in resumes, and more
- 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, but preference is for a continuous variable
- The variable should be included in at least one of the 2010-2021 surveys
- Your next assignment is to find your variables and come up with a research question.