# Social Statistics

*OLS With Multiple Variables*

November 28, 2023

# Warm Up 1

## *Regress age at first birth (`agekdbrn`) on years of education (`educ`)*

```
warmup_1 <- lm(agekdbrn ~ educ, data = gss_week11)

summary(warmup_1)
```

# Warm Up 1

```
Call:
lm(formula = agekdbrn ~ educ, data = gss_week11)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.5223  -3.5605  -0.9757   2.8548  27.0243 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.46735    0.80342   16.76   <2e-16 ***
educ         0.79237    0.05768   13.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.157 on 1055 degrees of freedom
  (868 observations deleted due to missingness)
```

# Warm Up 1

*Predict age at first birth for respondents with 16 years of education*

```
13.46735 + .79237*16
```

```
[1] 26.14527
```

# Warm Up 2

*Regress age at first birth (`agekdbrn`) on highest degree (`degree`). Use "College Degree" as the reference group.*

```r
gss_week11 <- mutate(gss_week11, degree = factor(degree,
       levels = c("Less Than HS", "HS Diploma",
       "Some College", "College Degree",
       "Grad/Prof Degree")))

gss_week11$degree <- relevel(factor(gss_week11$degree),
  ref = "College Degree")

warmup2 <- lm(agekdbrn ~ degree, data = gss_week11)

summary(warmup2)
```

# Warm Up 2

```
Call:
lm(formula = agekdbrn ~ degree, data = gss_week11)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.7426  -3.3598  -0.9965   2.6402  27.0035 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             27.3598     0.4000  68.400  < 2e-16 ***
degreeLess Than HS      -6.5101     0.5758 -11.307  < 2e-16 ***
degreeHS Diploma        -4.3633     0.5027  -8.680  < 2e-16 ***
degreeSome College      -3.3286     0.4917  -6.770 2.14e-11 ***
degreeGrad/Prof Degree   0.3829     0.5941   0.645    0.519    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Warm Up 2

*Predict age at first birth for respondents with a graduate or professional degree*

```r
27.3598 + .3829
```

```
[1] 27.7427
```

# Warm Up 3

*Regress having a first child at age 30 or later (`agekdbrn_30plus`) on religion (`religion`). Use "Protestant" as the reference group.*

```
gss_week11$religion <- relevel(factor(gss_week11$religion),
                               ref="Protestant")

warmup3 <- lm(agekdbrn_30plus ~ religion, data = gss_week11)

summary(warmup3)
```

# Warm Up 3

```
Call:
lm(formula = agekdbrn_30plus ~ religion, data = gss_week11)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4815 -0.1754 -0.1499 -0.1499  0.8501 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      0.149915   0.015756   9.515  < 2e-16 ***
religionCatholic 0.068709   0.028952   2.373  0.01782 *  
religionEastern  0.304631   0.082899   3.675  0.00025 ***
religionJewish   0.331567   0.075137   4.413 1.13e-05 ***
religionNone     0.009377   0.039216   0.239  0.81106    
religionOther    0.025524   0.052961   0.482  0.62995    
---
```

# Warm Up 3

## *Predict probability of having a first child at age 30 or later for Jewish respondents:*

```
.149915 + .331567
```

```
[1] 0.481482
```

# Introducing Multiple Regression

- So far, our models have had one X (even if it has more than one category)
- We want to adjust for possible confounding or spuriousness like we did with descriptive tables
  - → How do we *control for* other variables?
  - → Can we *explain away* the association between X and Y by controlling for other variables?

# Introducing Multiple Regression

- Find another variable, *hold it constant*, and see if the association between X and Y changes
- Can be categorical (Highest Degree, Year, Religion) or continuous (Years Since Marriage, Months Since Sister's First Birth)

# Introducing Multiple Regression

- We already saw that each additional year of education is associated with a delay of .79 years in the age at first birth.
- Perhaps religion explains some of the variation in both education and age at first birth. So let's *hold religion constant*.
- In R, include more variables by linking them to the independent variable with a plus sign

```
agekd_educ_religion <- lm(agekdbrn ~ educ + religion,
  data = gss_week11)

summary(agekd_educ_religion)
```

# Introducing Multiple Regression

```
Call:
lm(formula = agekdbrn ~ educ + religion, data = gss_week11)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.3572  -3.4731  -0.7609   2.5773  25.9809 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      13.32194    0.81313  16.384  < 2e-16 ***
educ              0.76259    0.05792  13.167  < 2e-16 ***
religionCatholic  1.54599    0.38639   4.001 6.75e-05 ***
religionEastern   3.44548    1.10633   3.114  0.00189 ** 
religionJewish    2.85010    1.01669   2.803  0.00515 ** 
religionNone     -0.14292    0.52325  -0.273  0.78480    
religionOther     1.24962    0.70710   1.767  0.07748 .  
```

# Introducing Multiple Regression

- Holding religion constant (or net of religion), each additional year of education is associated with a delay of .76 years in the age at first birth, on average
- To find the predicted values, think of the full equation:

$$\hat{y}_{agekdbrn} = \alpha + \beta_1(\text{educ}) + \beta_2(\text{Catholic}) + \beta_3(\text{Eastern}) + \beta_4(\text{Jewish}) + \beta_5(\text{None}) + \beta_6(\text{Other})$$

- Every prediction will have a value for education. Every prediction will also have a value for each binary religious category (even though they are mutually exclusive).

# Predictions From Multiple Regression

- For 16 years of education and Protestant (the reference category)

```r
13.32 + .76*16 + 1.55*0 + 3.45*0 + 2.85*0 - .14*0 + 1.25*0
```

```
[1] 25.48
```

- Try finding the predicted age at first birth for Catholic respondents with 16 years of education

```r
13.32 + .76*16 + 1.55*1 + 3.45*0 + 2.85*0 - .14*0 + 1.25*0
```

```
[1] 27.03
```

- Is the difference of 1.55 in the predictions between Protestants and Catholics *with the same years of education* statistically significant?

# Predictions From Multiple Regression

```
Call:
lm(formula = agekdbrn ~ educ + religion, data = gss_week11)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.3572  -3.4731  -0.7609   2.5773  25.9809 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      13.32194    0.81313  16.384  < 2e-16 ***
educ              0.76259    0.05792  13.167  < 2e-16 ***
religionCatholic  1.54599    0.38639   4.001 6.75e-05 ***
religionEastern   3.44548    1.10633   3.114  0.00189 ** 
religionJewish    2.85010    1.01669   2.803  0.00515 ** 
religionNone     -0.14292    0.52325  -0.273  0.78480    
religionOther     1.24962    0.70710   1.767  0.07748 .  
```

# Predictions From Multiple Regression

- What is the prediction for a respondent in an Eastern religion with 13 years of education?

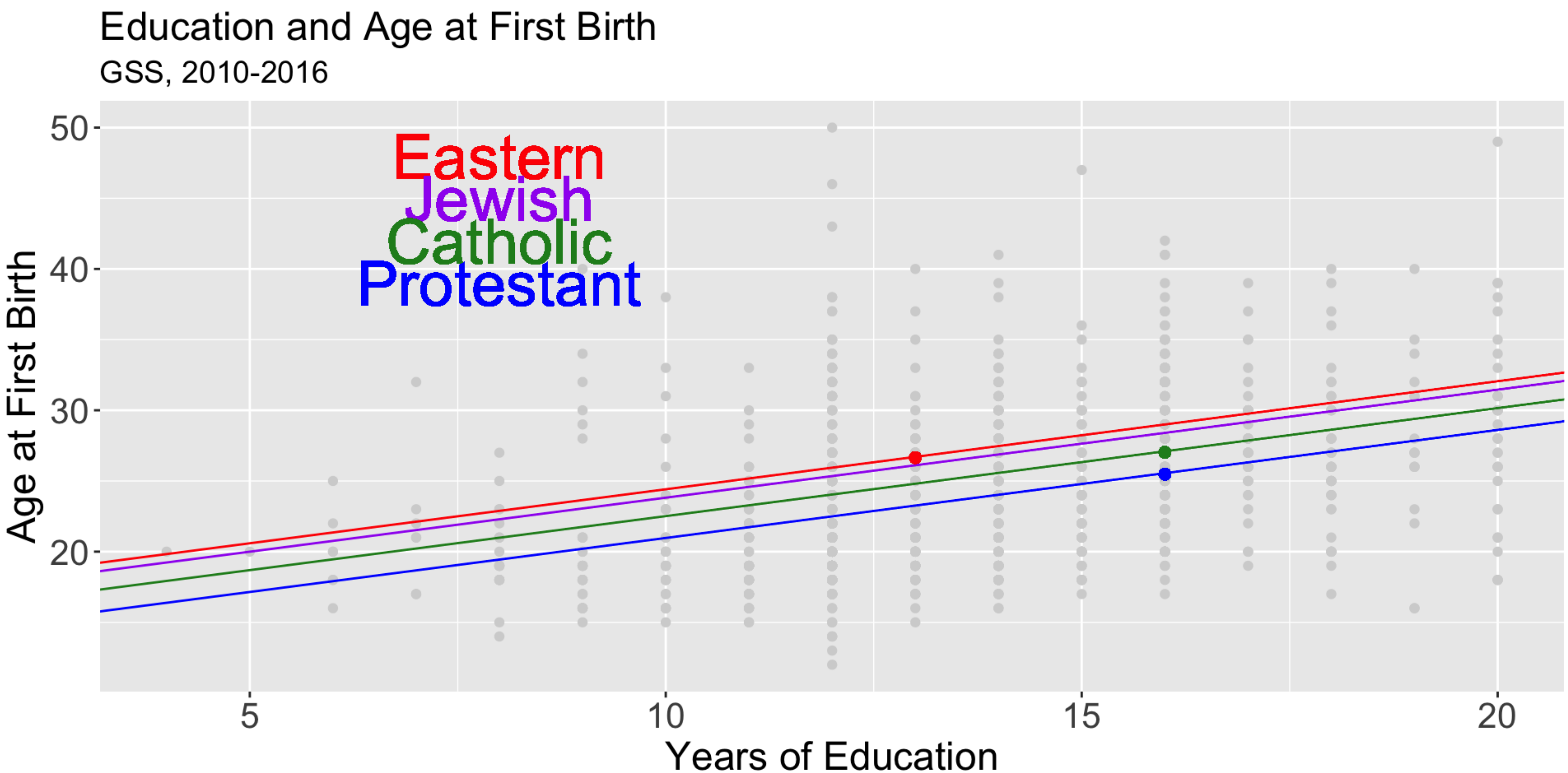
```
13.32 + .76*13 + 1.55*0 + 3.45*1 +2.85*0 - .14*0 + 1.25*0
```

```
[1] 26.65
```

# Plotting Multiple Regression

- How do we make sense of this in a plot?
- The beta for all groups is the coefficient for `educ`. So in this model the slopes are the same for each group.
- But the intercepts are different: use the intercept coefficient for the reference group, use the intercept and the respective coefficient for each other group

# Plotting Multiple Regression



# More And More Variables

- Models can continue adding control variables
- Let’s try regressing age at first birth on education, religion, and race

```r
agekd_educ_religion_race_model <-
lm(agekdbrn ~ educ + religion + racehisp,
data = gss_week11)

summary(agekd_educ_religion_race_model)
```

# More And More Variables

```
Call:
lm(formula = agekdbrn ~ educ + religion + racehisp, data = gss_week11)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.437  -3.485  -0.735   2.540  26.774 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       12.8134     0.8728  14.682  < 2e-16 ***
educ               0.7420     0.0589  12.599  < 2e-16 ***
religionCatholic   1.5079     0.4110   3.669 0.000256 ***
religionEastern    3.4348     1.2394   2.771 0.005681 ** 
religionJewish     2.6802     1.0187   2.631 0.008641 ** 
religionNone      -0.1464     0.5249  -0.279 0.780304    
religionOther      1.2509     0.7072   1.769 0.077226 .  
```

# More And More Variables

- Holding religion and race constant, each additional year of education is associated with a delay of .74 years in age at first birth, on average
- Controlling for education and race, Catholic women are 1.5 years older than Protestant women at their first birth, on average. This difference is significant.
- Net of education and religion, there is no significant difference in the age at first birth between Black women and women in the other race category, on average.
- Holding constant, controlling for, and net of can all be used interchangeably in these examples.

# More And More Variables

- Predictions still require the full equation
- What is the predicted age at first birth for a Black Protestant with 17 years of education?
- Black is the reference group for `racehisp` and Protestant is the reference group for `religion` so:

```
12.8134 + .7420*17
```

```
[1] 25.4274
```

# More And More Variables

- What is the predicted age at first birth for a Hispanic with no religious affiliation with 14 years of education?

```
12.8134 + .7420*14 - .1464 + .2763
```

```
[1] 23.3313
```

# Comparing Models

- How do we know if our model gets better when we add more control variables?
- In other words: how well does our X predict our Y?
- Without an X, only comparison is the difference between the observed Y and the mean of Y
- With an X, the measure of fit is the residual (the difference between the observed Y and the predicted Y)
- $\mathbf{r}^2$ is a function of both of these in the form of a ratio
  - → The proportional reduction in error from using the model

# R-Squared

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.41495    0.20125  -7.031 2.84e-12 ***
educ         0.22722    0.01442  15.762  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.78 on 1923 degrees of freedom
Multiple R-squared:  0.1144,    Adjusted R-squared:  0.1139
F-statistic: 248.4 on 1 and 1923 DF,  p-value: < 2.2e-16
```

# R-Squared

- Let's calculate $\mathbf{r}^2$ for the model regressing number of memberships on years of education
- $\mathbf{r}^2 = \dfrac{\sum (y-\bar{y})^2 - \sum (y-\hat{y})^2}{\sum (y-\bar{y})^2}$

```r
memnum_educ_model <-
lm(memnum ~ educ, data = gss_week11)

summary(memnum_educ_model)
```

# R-Squared

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.41495    0.20125  -7.031 2.84e-12 ***
educ         0.22722    0.01442  15.762  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.78 on 1923 degrees of freedom
Multiple R-squared:  0.1144,	Adjusted R-squared:  0.1139 
F-statistic: 248.4 on 1 and 1923 DF,  p-value: < 2.2e-16
```

# R-Squared

```
memnum_educ_model <- lm(memnum ~ educ, data = gss_week11)
```

```
gss_week11$pred_memnum <- memnum_educ_model$fitted.values

gss_week11$res_memnum <-
(gss_week11$memnum - gss_week11$pred_memnum)^2

gss_week11$dev_memnum <-
(gss_week11$memnum - mean(gss_week11$memnum))^2

rsquared <- ((sum(gss_week11$dev_memnum)) -
                  (sum(gss_week11$res_memnum))) /
                  sum(gss_week11$dev_memnum)

rsquared
```

```
[1] 0.1144075
```

# Properties of R-Squared

- Like correlation, always between 0 and 1
- Unlike correlation, always positive (since it is squared and a proportion)
- Closer to 1 means observations fall more tightly around the line (in a linear association)
- Will usually increase when you add variables to the model. But that does not necessarily mean the model is getting better.
- Remember, parsimony is still our goal

# Comparing Models

- If we regress number of memberships on education and age, it looks like the model is better since r-squared increases.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.958854   0.243859  -8.033 1.65e-15 ***
educ         0.234599   0.014486  16.195  < 2e-16 ***
age          0.009605   0.002451   3.919 9.22e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.773 on 1922 degrees of freedom
Multiple R-squared:  0.1214,	Adjusted R-squared:  0.1205 
F-statistic: 132.8 on 2 and 1922 DF,  p-value: < 2.2e-16
```

# Comparing Models

- But be careful: R-squared will almost always go up as you add variables, even if the variables are not significant.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -1.9264512  0.2546124  -7.566 5.93e-14 ***
educ            0.2334475  0.0146100  15.979  < 2e-16 ***
age             0.0095195  0.0024569   3.875  0.00011 ***
placeNortheast  0.0009124  0.1284526   0.007  0.99433    
placeSoutheast -0.0480122  0.1049289  -0.458  0.64731    
placeWest       0.0250015  0.1204691   0.208  0.83561    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.774 on 1919 degrees of freedom
Multiple R-squared:  0.1217,	Adjusted R-squared:  0.1194 
F-statistic: 53.16 on 5 and 1919 DF,  p-value: < 2.2e-16
```

# Adjusted R Squared

- Adjusted r-squared adjusts for the number of parameters in your model (but not for how good they are)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.41495    0.20125  -7.031 2.84e-12 ***
educ         0.22722    0.01442  15.762  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.78 on 1923 degrees of freedom
Multiple R-squared:  0.1144,    Adjusted R-squared:  0.1139
F-statistic: 248.4 on 1 and 1923 DF,  p-value: < 2.2e-16
```

# Adjusted R-Squared

```
# adjusted_rsquared =
# 1 - (((1 - rsquared)*(n-1)) / (n-k-1))

# n = sample size; k = number of variables

adjusted_rsquared <-
1 - (((1 - rsquared)*(1924-1)) / (1924-1-1))

adjusted_rsquared
```

```
[1] 0.1139467
```

# Adjusted R-Squared

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.41495    0.20125  -7.031 2.84e-12 ***
educ         0.22722    0.01442  15.762  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.78 on 1923 degrees of freedom
Multiple R-squared:  0.1144,    Adjusted R-squared:  0.1139 
F-statistic: 248.4 on 1 and 1923 DF,  p-value: < 2.2e-16
```