Homework 1

Jonathan Pedroza

1/7/2020

In this assignment you will be working with a subset of the publicly available version of the data from the National Longitudinal Study of Adolescent to Adult Health (Add Health). This data can be found on the course website in Canvas under the “Files” tab. Look for data file “AddHealth.dta.”

Begin by downloading the data file to your computer and opening it in STATA. One way to open a data set is to use the “Open” tab in the File menu and proceed to where you have saved the data file. Double click the file name and it should load in STATA.

For this assignment, check out these two resources:

* A STATA commands quick-reference handout (located on Canvas in the “STATA code”sub folder of “Files”).
* The “Hints for HW1” document available in the Homework 1 folder on Canvas.

On the due date please hand in a hardcopy of your STATA output and answers to all questions in this assignment. You can either hand in a document with your answers (e.g., Word doc) with STATA output attached or you can incorporate answers to the questions directly into the STATA output in the form of comments. I prefer typed assignments in Times New Roman size 12 or Arial size 11.

health <- data %>%   
 dplyr::select(aid,  
 schoolid,  
 bmi\_w1,  
 sex,  
 parent\_highestedu,  
 hrs\_sleep,  
 age\_w1,  
 multpov\_w1)

1. Begin by obtaining basic statistics (number of observations, mean, standard deviation, min/max values) on the following variables in the data set. When necessary you can round values to the nearest two decimal places.

**Answer:**

Age, *n* = 6501, *M* = 15.53, *sd* = 1.78, minimum value = 11, maximum value = 21

BMI, *n* = 6291, *M* = 22.49, *sd* = 4.42, minimum value = 11.22, maximum value = 56.43

Hours of sleep, *n* = 6477, *M* = 7.81, *sd* = 1.43, minimum value = 1, maximum value = 20

three\_var <- health %>%   
 dplyr::select(age\_w1, bmi\_w1, hrs\_sleep)  
  
psych::describe(three\_var)

## vars n mean sd median trimmed mad min max range skew  
## age\_w1 1 6501 15.53 1.78 16.0 15.55 1.48 11.00 21.00 10.00 -0.04  
## bmi\_w1 2 6291 22.49 4.42 21.5 21.97 3.38 11.22 56.43 45.21 1.50  
## hrs\_sleep 3 6477 7.81 1.43 8.0 7.77 1.48 1.00 20.00 19.00 0.37  
## kurtosis se  
## age\_w1 -0.83 0.02  
## bmi\_w1 3.91 0.06  
## hrs\_sleep 2.36 0.02

1. Generate the following two variables:
2. female – code as 0 = male, 1 = female
3. poor – code so that 0 = no, 1 = yes. In this case, being poor is defined as having a multpov\_w1 value of < =1 (i.e., family is at or below the federal poverty line), or alternatively being non-poor (poor = 0) means having multpov\_w1 > 1.

health %>%   
 count(sex)

## # A tibble: 3 x 2  
## sex n  
## <dbl> <int>  
## 1 1 3147  
## 2 2 3356  
## 3 NA 1

describe(health$multpov\_w1)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 4928 3.11 3.94 2.5 2.61 1.85 0 83.25 83.25 10.21 162.01  
## se  
## X1 0.06

# 1 = male, 2 = female

**Answer:**  
**health <- health** %>%mutate**(female =** recode**(sex, "1" = 'male',  
 "2" = 'female'),  
 poor =** case\_when**(multpov\_w1** <= **1** ~ **1,  
 multpov\_w1** > **1** ~ **0),  
 poor =** recode**(poor, '1' = 'poor',  
 '0' = 'not\_poor'),  
 female =** as.factor**(female),  
 poor =** as.factor**(poor))**

1. In STATA, output basic statistics (n mean sd min max) of bmi\_w1 and hrs\_sleep for each category of the following variables: female and poor.
2. Is mean BMI higher among males or females (compare raw scores, don’t worry about statistical significance of the difference)?

**Answer:**

BMI is higher in males (*M* = 22.66) than females (*M* = 22.33).

1. What is the mean hours slept per night among adolescents from poor families? from non-poor families?

**Answer:**

The average number of hours of sleep for adolescents in poor families is 7.91 hours. In non-poor families, adolescents sleep 7.82 hours on average.

describeBy(health$bmi\_w1, group = health$female)

##   
## Descriptive statistics by group   
## group: **female**## vars n mean sd median trimmed mad min max range skew  
## X1 1 3204 **22.33** 4.45 21.28 21.76 3.32 12.03 46.36 34.34 1.51  
## kurtosis se  
## X1 3.31 0.08  
## --------------------------------------------------------   
## group: **male**## vars n mean sd median trimmed mad min max range skew  
## X1 1 3087 **22.66** 4.38 21.72 22.19 3.46 11.22 56.43 45.21 1.5  
## kurtosis se  
## X1 4.6 0.08

describeBy(health$hrs\_sleep, group = health$female)

##   
## Descriptive statistics by group   
## group: **female**  
## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 3343 **7.77** 1.45 8 7.73 1.48 1 20 19 0.41 2.97  
## se  
## X1 0.03  
## --------------------------------------------------------   
## group: **male**  
## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 3134 **7.85** 1.41 8 7.82 1.48 1 15 14 0.34 1.66  
## se  
## X1 0.03

describeBy(health$bmi\_w1, group = health$poor)

##   
## Descriptive statistics by group   
## group: **not\_poor**  
## vars n mean sd median trimmed mad min max range skew  
## X1 1 3881 **22.35** 4.42 21.39 21.83 3.33 12.03 56.43 44.41 1.62  
## kurtosis se  
## X1 4.9 0.07  
## --------------------------------------------------------   
## group: **poor**  
## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 887 **22.85** 4.44 21.85 22.33 3.49 11.22 42 30.78 1.27 2.19  
## se  
## X1 0.15

describeBy(health$hrs\_sleep, group = health$poor)

##   
## Descriptive statistics by group   
## group: **not\_poor**  
## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 3977 **7.82** 1.38 8 7.78 1.48 3 15 12 0.33 1.47  
## se  
## X1 0.02  
## --------------------------------------------------------   
## group: **poor**  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 931 **7.91** 1.61 8 7.86 1.48 1 20 19 0.71 4.75 0.05

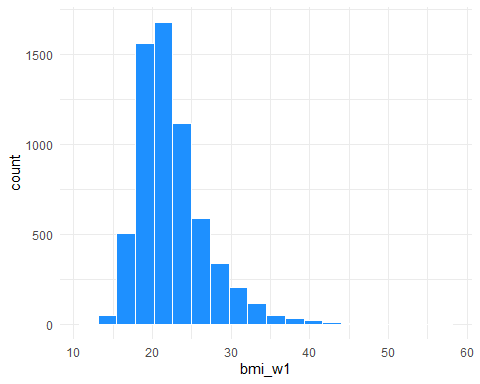
1. Generate histograms to visualize the distribution of bmi\_w1 and hrs\_sleep in the sample. For each histogram, give a brief description of the distribution (i.e., normally distributed, skewed, etc?)

**Answer:**

For the full sample, BMI is positively skewed slightly, however it is still representative of a normal distribution. Hours of sleep represents a normal distribution.

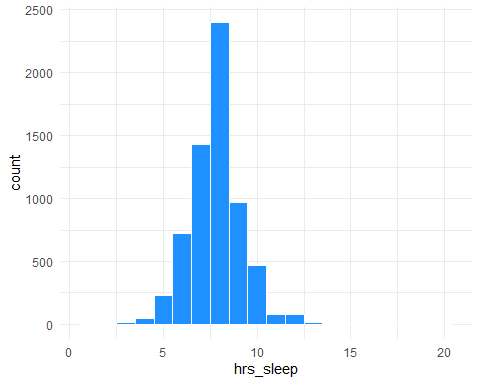
health %>%   
 ggplot(aes(bmi\_w1)) +  
 geom\_histogram(color = 'white', fill = 'dodgerblue', bins = 20) +  
 theme\_minimal()

## Warning: Removed 213 rows containing non-finite values (stat\_bin).



health %>%   
 ggplot(aes(hrs\_sleep)) +  
 geom\_histogram(color = 'white', fill = 'dodgerblue', bins = 20) +  
 theme\_minimal()

## Warning: Removed 27 rows containing non-finite values (stat\_bin).



1. Using notation from class, write each of the following linear regression models. Be sure to include residual terms and index the appropriate terms by *i*.

**Answer:**

1. Fit single-level linear regression models in STATA for each of the models outlined in Question 5.
2. Compare the Mean Square Error (MSE) in Model 1 to MSE in Model 2. Compare the MSE in Model 3 to MSE in Model 4. What might account for the differences?

**Answer:**

The differences in MSE between models 1 and 2 is based on the inclusion of predictors (i.e., female, age, and parent education). The inclusion of predictors explained more of the variation than the intercept alone. The same can be said for the differences in MSE between model 3 and model 4.

1. Use your results from Model 2 to answer the following questions. In this case consider a “beta” parameter to be statistically significant if it has p-value < 0.05.
2. Do mean BMI values differ significantly between males and females in the sample? Describe any differences (i.e., do males or females have a higher BMI score on average, and by how much).
3. Interpret the age\_w1 beta parameter in terms of statistical significance and what the change in predicted BMI is for every unit change in age (e.g., “for every 1 year increase in age, the predicted BMI…”).
4. How do mean BMI values differ between adolescents whose parents have attained different educational levels? Describe in terms of statistical significance, direction, and magnitude.

**Answers:**

A. While adjusting for age and parent education, male adolescents had a significantly higher BMI than female adolescents. Male adolescents’ BMI was .26 (*β* = .03) higher on average than female adolescents.

B. While adjusting for gender/sex and parent education, age was a significant predictor of BMI, where every 1 year increase in age, the predicted BMI had a .42 (*β* = .17) increase.

C. In comparison to adolescents with parents who have a college degree, adolescents with parents that had less than a high school degree had a significantly higher BMI of 1.26 (*β* = .09) on average. Similarly, adolescents with parents that had a high school diploma had a higher BMI score of .61 (*β* = .06) on average over adolescents with college graduate parents. Parents with some college also reported having a significantly higher BMI score of .54 (*β* = .05) on average when compared to parents with college degrees.

model1 <- lm(bmi\_w1 ~ 1, data = health)  
sum1 <- summary(model1)  
sum1

##   
## Call:  
## lm(formula = bmi\_w1 ~ 1, data = health)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.272 -2.949 -0.990 1.903 33.942   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.49217 0.05568 404 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.416 on 6290 degrees of freedom  
## (213 observations deleted due to missingness)

mse1 <- mean(sum1$residuals^2)  
  
health <- health %>%   
 mutate(parent\_ed = as.factor(parent\_highestedu),  
 parent\_ed = recode(parent\_ed, '1' = 'Less than high school',  
 '2' = 'completed high school',  
 '3' = 'some college',  
 '4' = 'college degree'),  
 parent\_ed = relevel(parent\_ed, ref = 'college degree'))  
  
model2 <- lm(bmi\_w1 ~ female + age\_w1 + parent\_ed, data = health)  
sum2 <- summary(model2)  
sum2

##   
## Call:  
## lm(formula = bmi\_w1 ~ female + age\_w1 + parent\_ed, data = health)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.785 -2.852 -0.959 1.757 33.868   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.41131 0.49019 31.439 < 2e-16 \*\*\*  
## femalemale 0.26222 0.10990 2.386 0.0171 \*   
## age\_w1 0.41908 0.03115 13.455 < 2e-16 \*\*\*  
## parent\_edLess than high school 1.25930 0.19798 6.361 2.15e-10 \*\*\*  
## parent\_edcompleted high school 0.60629 0.14107 4.298 1.75e-05 \*\*\*  
## parent\_edsome college 0.53859 0.13833 3.894 9.98e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.324 on 6204 degrees of freedom  
## (294 observations deleted due to missingness)  
## Multiple R-squared: 0.03884, Adjusted R-squared: 0.03807   
## F-statistic: 50.14 on 5 and 6204 DF, p-value: < 2.2e-16

betas.lm(model2)

## beta se.beta  
## femalemale 0.02973583 0.01246211  
## age\_w1 0.16801965 0.01248753  
## parent\_edLess than high school 0.08507614 0.01337513  
## parent\_edcompleted high school 0.06054383 0.01408680  
## parent\_edsome college 0.05494145 0.01411080

mse2 <- mean(sum2$residuals^2)  
  
cbind**(mse1, mse2)**

**## mse1 mse2  
## [1,] 19.50002 18.68132**

1. Use your results from Model 4 to answer the following questions. In this case consider a “beta” parameter to be statistically significant if it has p-value < 0.05.
2. Do hours slept differ significantly between males and females in the sample, and if so by how much?
3. Do hours slept differ significantly between adolescents from poor and non-poor families, and if so by how much?
4. Interpret the age\_w1 beta parameter in terms of statistical significance and what the change in predicted hours slept is for every unit change in age (e.g., “for every 1 year increase in age, the expected number of hours slept…”).

**Answers:**

A. Hours of sleep was significantly different in males and females, with males engaging in .14 (*β* = .05) more hours of sleep on average than females while adjusting for whether adolescents were below the poverty line and their age.

B. Hours of sleep did not differ between adolescents who were in poor and non-poor families.

C. The association between age and hours slept was significant. For every 1 year increase in age, the expected number of hours slept reduced by -.20 (*β* = -.25).

model3 <- lm(hrs\_sleep ~ 1, data = health)  
sum3 <- summary(model3)  
sum3

##   
## Call:  
## lm(formula = hrs\_sleep ~ 1, data = health)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.8082 -0.8082 0.1918 1.1918 12.1918   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.80824 0.01778 439.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.431 on 6476 degrees of freedom  
## (27 observations deleted due to missingness)

mse3 <- mean(sum3$residuals^2)  
  
model4 <- lm(hrs\_sleep ~ female + poor + age\_w1, data = health)  
betas.lm(model4)

## beta se.beta  
## femalemale 0.04784185 0.01383852  
## poorpoor 0.02544890 0.01382805  
## age\_w1 -0.24584834 0.01383863

sum4 <- summary(model4)  
sum4

##   
## Call:  
## lm(formula = hrs\_sleep ~ female + poor + age\_w1, data = health)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.2648 -0.7710 -0.0355 0.6741 12.2729   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.84337 0.17540 61.820 < 2e-16 \*\*\*  
## femalemale 0.13666 0.03953 3.457 0.000551 \*\*\*  
## poorpoor 0.09269 0.05036 1.840 0.065773 .   
## age\_w1 -0.20056 0.01129 -17.765 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.383 on 4904 degrees of freedom  
## (1596 observations deleted due to missingness)  
## Multiple R-squared: 0.06237, Adjusted R-squared: 0.06179   
## F-statistic: 108.7 on 3 and 4904 DF, p-value: < 2.2e-16

mse4 <- mean(sum4$residuals^2)  
  
cbind**(mse3, mse4)**

**## mse3 mse4  
## [1,] 2.047219 1.911707**

1. In future assignments we will explore clustering of adolescents by school attended.
2. For now, use the tabulate command to generate output specifying how many students are in each school.
3. Which school has the most students (provide schoolid number), and how many students are in this school?
4. Which school has the fewest students (provide schoolid number), and how many students are in this school?

**Answers:**

B. The school with the most students was school 166 with 122 students.

C. The school with the least amount of students was school 216 with 12 students.

health %>%   
 group\_by(schoolid) %>%   
 summarize(n = n()) %>%  
 arrange(desc(n))

## # A tibble: 132 x 2  
## schoolid n  
## <dbl> <int>  
## 1 **166 122**  
## 2 191 100  
## 3 136 98  
## 4 113 97  
## 5 114 92  
## 6 121 87  
## 7 126 87  
## 8 141 85  
## 9 108 84  
## 10 304 83  
## # ... with 122 more rows

health %>%   
 group\_by(schoolid) %>%   
 summarize(n = n()) %>%  
 arrange(n)

## # A tibble: 132 x 2  
## schoolid n  
## <dbl> <int>  
## 1 **216 12**  
## 2 224 17  
## 3 271 19  
## 4 254 20  
## 5 292 21  
## 6 103 22  
## 7 273 22  
## 8 276 22  
## 9 277 22  
## 10 102 23  
## # ... with 122 more rows