Recap of Multiple Linear Regression

Dr. Soliman

2024-09-19

1. In your own words, explain what each line of code below does. I have marked each relevant line of code with a comment, which correspond to the description you will need to make below the code chunk. However, for the first action, where there is a question, comment directly next to it. This is to get you used to commenting your code in class.

```r
# what does the line below do? WRITE HERE: reads in the csv dataset from dropbox
# and assigns it to the star_df object
star_df <- read.csv("https://www.dropbox.com/s/bf1fog8yasw3wjj/star_data.csv?dl=1")

star_df_homework <- star_df %>% # part a
  filter(complete.cases(.)) %>% # part b
  filter(grade == 2) %>% # part c
    mutate(small = (star == "small"), # part d
           regular = (star == "regular"), # part e
           regular_plus = (star == "regular+aide"), # part f
           sum = small + regular + regular_plus) # part g
```

Part a: Upon completing the below actions, it stores the updated data set with new columns to a new object called star_df_homework. In this specific line, it assigns what we read in, star_df, to star_df_homework. The subsequent pipe is signaling that we will be doing something else after. Remember that these could all be on the same line if desired, but its easier to read and comment the current way.

Part b: This line uses the filter function to select only data entries that have complete information, i.e., ones with no missing information/ NA's

Part c: This line also uses the filter function to select the data for only entries where the student is in second grade

Part d: This line creates a dummy variable called small using thr mutate function and filling the new dummy variable with either TRUE or FALSE for that observaation, corresponding to whether the class size is small or not (TRUE meaning class is small, FALSE meaning class is not small (i.e., regular or regular+aide))

Part e: another dummy variable but this one is called regular and it gets filled with either TRUE or FALSE, corresponding to whether or not class size is regular

Part f: another dummy variable called regular_plus and this one is also filled with either TRUE or FALSE, corresponding to whether or not class size is regular+aide.

Part g: creating another dummy variable called sum that holds 0's or 1's. It adds up the values of three previous dummy variables, but it can only ever be one of the classes at a time, so the this varaible will always have the value 1.

2. When you run the code below to generate simple tables, why are there so fewer observations in the second line of code? What does each value in the variable `star` mean?

The first line is for the full raw dataset, while the second uses our filtered data set (of complete cases and grade 2). Therefore, it mechanically must have fewer observations, as we removed data. The variable star is the treatment variable and captures what class the student was

assigned to: small, regular, or regular+aide, and the corresponding number of students in each treatment arm.

```
table(star_df$star)
```

```
##
##      regular regular+aide        small
##         9192         9589         8015
```

```
table(star_df_homework$star)
```

```
##
##      regular regular+aide        small
##         1945         2033         1694
```

Make sure to choose the correct dataset for this and future sections.

3. Tabulate the variable `school`, as we did in the previous question but just for the *analysis dataset*, to make sure you know the values it takes on. Then regress `math` on `school`. Interpret the coefficients. What's the omitted category? Do you find the results surprising? If so, why? What might be an omitted variable?

```
table(star_df_homework$school)
```

```
##
## inner-city      rural   suburban      urban
##       1199       2665       1476        332
```

```
lm(math ~ school, star_df_homework)
```

```
##
## Call:
## lm(formula = math ~ school, data = star_df_homework)
##
## Coefficients:
##    (Intercept)     schoolrural  schoolsuburban     schoolurban
##         561.56           30.18           16.93           20.33
```

The `school` variable is a categorical variable indicating the school location type. There are four categories: inner-city, suburban, rural and urban.

The ommitted category in the regression is inner-city, meaning that the intercept represents the average math score of students in inner-city schools and that the other coefficients should be interpreted relative to the expected math score of students in inner-city schools. In particular, on average, students in rural schools score 30 points higher than those in inner-city schools, those in suburban schools score almost 17 points higher, and those in urban schools score about 20 points higher. Inner-cities tend to be poorer areas which might explain that they score so much lower than students in other locations. An omitted variable might be some student- or school-evel variable of disadvantage.

4. Compute the share of students qualifying for free lunch (i.e. `lunch` equals "free") by school location category. Hint: one option is to use the group_by() function then %>% to summarise(mean(variable == "value")). What do you observe? Add `lunch` to the previous question's regression. How do the coefficients change?

```
star_df_homework %>%
    group_by(school) %>%
    summarise(mean(lunch == "free"))
```

```
## # A tibble: 4 x 2
```

```
##   school     `mean(lunch == "free")`
##   <chr>                        <dbl>
## 1 inner-city                   0.897
## 2 rural                        0.389
## 3 suburban                     0.335
## 4 urban                        0.443
```

As expected, the share of students qualifying for free lunches, a common indicator of disadvantage, is significantly higher in inner-city schools (almost 90% of such students) while it is between 34% and 44% of students in other locations.

```r
lm(math ~ school, star_df_homework)
```

```
##
## Call:
## lm(formula = math ~ school, data = star_df_homework)
##
## Coefficients:
##    (Intercept)     schoolrural  schoolsuburban     schoolurban
##         561.56           30.18           16.93           20.33
```

```r
lm(math ~ school + lunch, star_df_homework)
```

```
##
## Call:
## lm(formula = math ~ school + lunch, data = star_df_homework)
##
## Coefficients:
##    (Intercept)     schoolrural  schoolsuburban     schoolurban   lunchnon-free
##        559.228          18.750           4.289          10.114          22.517
```

We observe that the coefficients on the rural, suburban and urban decrease dramatically once the free lunch status of students are taken into account. Note that students not qualifying for free lunch score significantly higher, on average, than those who do, controlling for the school's location category.

5. Regress `math` on `star` and interpret the coefficients. Then, regress `math` on `star`, `gender`, `ethnicity`, `lunch`, `degree`, `experience` and `school`. Recalling that this is a randomized experiment, does it look like the randomization was well done?

```r
lm(math ~ star, star_df_homework)
```

```
##
## Call:
## lm(formula = math ~ star, data = star_df_homework)
##
## Coefficients:
##      (Intercept)  starregular+aide         starsmall
##          577.696             2.696             8.942
```

The interpretation of the coefficients is as usual a comparison of conditional means. That is, students in small classes score, on average 8.94 points higher than those in regular classes (the omitted category) while students in regular + aide classes score only 2.70 points higher.

```r
reg_all <- lm(math ~ star + gender + ethnicity + lunch + degree +
                experience + school, star_df_homework)
reg_all
```

```
##
```

```
## Call:
## lm(formula = math ~ star + gender + ethnicity + lunch + degree +
##     experience + school, data = star_df_homework)
##
## Coefficients:
##       (Intercept)   starregular+aide          starsmall         gendermale
##         557.64898            2.01899            7.89579           -1.07263
## ethnicityamindian     ethnicityasian      ethnicitycauc  ethnicityhispanic
##         -21.91483           42.16429           18.61906           15.04510
##    ethnicityother       lunchnon-free       degreemaster          degreephd
##          54.67051           18.40555           -1.29047            0.97823
##  degreespecialist          experience        schoolrural      schoolsuburban
##          12.11634           -0.05146            3.99491           -4.65503
##       schoolurban
##          -3.97039
```

The coefficients on the treatment variable `star` decrease very slightly compared to the previous simple regression. This is expected considering this was a randomized experiment and therefore we can suppose that the randomization seems to have worked. If the randomization had been very poor, then accounting for all these factors should alter the coefficients on `star` more substantially.

6. (Optional) What's the adjusted R^2 from the previous multiple regression? How do you interpret it? What might you deduce about the importance of observable individual, teacher and school characteristics in explaining educational outcomes?

```r
summary(reg_all)
```

```
##
## Call:
## lm(formula = math ~ star + gender + ethnicity + lunch + degree +
##     experience + school, data = star_df_homework)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -137.882  -27.033   -1.892   25.174  144.834
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       557.64898    1.68995 329.979  < 2e-16 ***
## starregular+aide    2.01899    1.30760   1.544 0.122635
## starsmall           7.89579    1.36425   5.788 7.52e-09 ***
## gendermale         -1.07263    1.08674  -0.987 0.323675
## ethnicityamindian -21.91483   28.97355  -0.756 0.449457
## ethnicityasian     42.16429   11.90591   3.541 0.000401 ***
## ethnicitycauc      18.61906    1.76609  10.543  < 2e-16 ***
## ethnicityhispanic  15.04510   14.53924   1.035 0.300810
## ethnicityother     54.67051   13.71246   3.987 6.78e-05 ***
## lunchnon-free      18.40555    1.25646  14.649  < 2e-16 ***
## degreemaster       -1.29047    1.17370  -1.099 0.271603
## degreephd           0.97823    7.08286   0.138 0.890157
## degreespecialist   12.11634    5.50130   2.202 0.027674 *
## experience         -0.05146    0.06398  -0.804 0.421259
## schoolrural         3.99491    2.09764   1.904 0.056898 .
## schoolsuburban     -4.65503    1.90889  -2.439 0.014775 *
## schoolurban        -3.97039    2.91817  -1.361 0.173703
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.86 on 5655 degrees of freedom
## Multiple R-squared:  0.1489, Adjusted R-squared:  0.1465
## F-statistic: 61.85 on 16 and 5655 DF,  p-value: < 2.2e-16
```

The adjusted R^2 of the previous regression is about 0.15, which means that that model explains about 15% of the variance in students' math scores. This implies that 85% of the variance in math scores remains unaccounted for. In simple terms, class size, gender, ethnicity, free lunch status, teacher experience and degree, and school location, explain very little of the differences in math scores between the students in this dataset. Extrapolating a bit, one may argue that most of the differences in educational outcomes do not appear to be explained by these factors. This DOES NOT mean that they may not have important causal effects on educational outcomes.

7. Regress `math` on `gender` and `experience` (the teacher's experience). Interpret the coefficients. How would these regression results look like visually?

```r
lm(math ~ gender + experience, star_df_homework)
```

```
##
## Call:
## lm(formula = math ~ gender + experience, data = star_df_homework)
##
## Coefficients:
## (Intercept)   gendermale   experience
##   581.35255     -0.91383      0.03453
```

The intercept coefficient corresponds to the average math score of female students, keeping teacher experience constant. Recall that because we have both a numeric variable and a dummy variable as regressors, the intercept can be interpreted as the intercept for the line of the omitted category, in this case female. The coefficient on gender represents the expected difference in math scores between male and female students, holding teacher experience constant. The coefficient is negative but small, implying that on average male students score slightly lower than their female peers, holding teacher experience constant. Graphically, this implies that the line for male students lies below that for females. The coefficient on experience corresponds to the expected change in math score associated, on average, to an increase in a teacher's experience by 1 year, accounting for student gender. As such, an additional year of teacher experience is associated, on average, with a 0.03 increase in math scores, keeping gender constant. This is a very small effect, since you would need 100 years of increased experience to expect math scores to increase by 3 points.