Regression Session Materials

2025-10-28

Background

In this problem set, you will analyze data from Cattaneo (2010), which examines the effect of maternal smoking during pregnancy on infant birthweight (measured in grams). This is the same dataset we used in our takehome exam. The data contains information on mothers, fathers, and newborns.

Your goal is to explore how help students understand how regression can be used to estimate and interpret the relationship between smoking and birth outcomes. We are trying to connect the principles of OLS regressions to the earlier material in the class (i.e., SDO, matching), the idea of rebalancing versus prediction off fitted lines, correct interpretation of regression coefficients and the R-squared using both OLS output and visualization of mean predicted outcomes, $\frac{\sum_i^N \widehat{y_i}}{N}$.

First, we will begin by importing the dataset from Cunningham's github repo with the `read_dta` command from the `haven` package.

```
data <- read_dta("https://raw.github.com/scunning1975/mixtape/master/cattaneo2.dta")
```

In the `cattaneo2` dataset are some of the following variables that we will use in the exercise.

Variables of Interest

Variable	Description
`bweight`	**Outcome variable**: Infant birthweight (grams)
`mbsmoke`	**Treatment variable**: Mother smoked during pregnancy (1 = yes)
`mage`	Mother's age
`medu`	Mother's education (years)
`mrace`	Mother's race (1 = White)
`fedu`	Father's education (years)
`fage`	Father's age
`fhisp`	Father's Hispanic indicator variable (1 = yes)
`frace`	Father's race (1 = White)
`deadkids`	Number of previous newborns who died
`nprenatal`	Number of prenatal visits
`order`	Birth order of the infant
`monthslb`	Months since last birth
`prenatal1`	1 if first prenatal visit was in 1st trimester
`alcohol`	1 if alcohol consumed during pregnancy
`fbaby`	1 if first baby

Bivariate regression

First we will run regressions and focus on calculation and interpretation, as well as visualization. Remember that the outcome is `bweight` and our treatment variable is `mbsmoke`.

Question 1 (1 points)

Compute the mean birthweight for babies of smoking and nonsmoking mothers. Calculate the simple difference in outcomes (SDO) between the two groups.

```r
mean_treat  <- mean(data$bweight[data$mbsmoke == 1], na.rm = TRUE)
mean_control <- mean(data$bweight[data$mbsmoke == 0], na.rm = TRUE)
sdo <- mean_treat - mean_control
mean_treat; mean_control; sdo
```

```
## [1] 3137.66
```

```
## [1] 3412.912
```

```
## [1] -275.2519
```

Prompt:

1. In your own words, interpret the SDO.

The SDO is a difference in the mean birth weight for the women who smoked and the women who did not smoke during pregnancy.

2. What treatment assignment mechanism is needed in order for the SDO to have a causal interpretation?

Recall that the SDO is equal to the ATE plus the selection bias term plus the heterogenous treatment effect bias term. Under randomization of the treatment (i.e., treatment is assigned independent of potential outcomes), selection bias is zero and the ATT equals the ATU. This causes both bias terms to cancel out, and thus the SDO has a clear causal interpretation as an ATE. This is based on the work done in the first half of the semester as well as the focus of our midterm.

Question 2 (4 points)

Run a simple linear regression of birthweight on smoking (`bweight ~ mbsmoke`). Save the output as `fit1` and display the estimated coefficients.

```r
fit1 <- lm(bweight ~ mbsmoke, data = data)
tidy(fit1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    3413.      9.26     369.  0
## 2 mbsmoke        -275.     21.5      -12.8 4.68e-37
```

Prompt:

3. Interpret and compare the intercept term from this model to the mean birth weight of the control group from question 1.

The intercept is the average birth weight for the control group. In the earlier question 1, you can see that that number was 3412.912. Here it is rounded, but it is the same number – 3413.

Thus the intercept in a regression is the average of the outcome when the dummy variable is set equal to zero – which would be when $D = 0$, or when we are evaluating the control group.

3. Interpret and compare the regression coefficient on `mbsmoke` to the SDO from question 1.

The regression coefficient labeled in our output as "`2 mbsmoke`" is -275. Recall from question 1 that the SDO was -275.2519. This is the same number, just with the numbers to the right of the decimal rounded to the 0th decimal point.

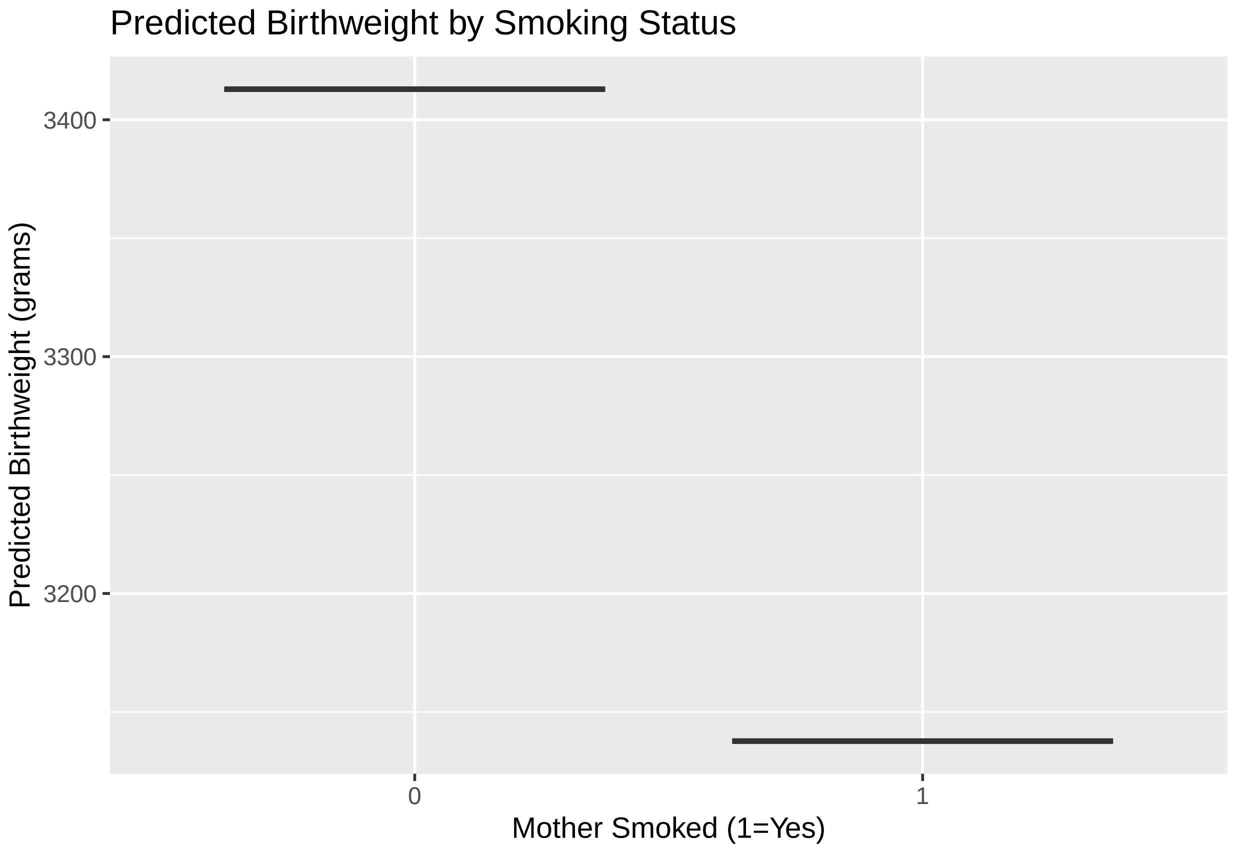
4. What treatment assignment mechanism would give the coefficient on `mbsmoke` a causal interpretation?

The answer is the same as in question 2 – randomization of the treatment. As these are the same calculations, then the answer would be the same. Thus we want to help students see that the OLS specification `bweight ~ mbsmoke` generates means and differences, and that the coefficient on the treatment variable, `mbsmoke`, is the same calculation as the SDO since OLS is taking two means and differencing them.

Question 3 (2 points)

Create a visualization of average birthweight for mathers who smoked versus those who didn't smoke using the `fit1` model.

```r
data$predicted <- predict(fit1)
ggplot(data, aes(x = factor(mbsmoke), y = predicted)) +
  geom_boxplot(fill = "skyblue", alpha = 0.6) +
  labs(x = "Mother Smoked (1=Yes)", y = "Predicted Birthweight (grams)",
       title = "Predicted Birthweight by Smoking Status")
```



Prompt:

5. Interpret the gap between these two horizontal lines and compare it with the SDO as well as the coefficient on `mbsmoke` from Quesiton 2.

Visually, the gap in the two means is the SDO. For students who have not connected this, the hope is that they can. But the hope is also that they see the role that `predict(fit1)` is playing. Students should understand that given the intercept term and the coefficient on the treatment variable, that OLS is generated predicted values and we are plotting that. But, given there are only two values to the treatment, predicted values are the same for $D = 1$ and $D = 0$, respectively. That's why the visualization is a flat line (unlike what we will see below when we include more controls).

Multivariate regression

Now we will move into regressions that include both the treatment variable, `mbsmoke`, as well as additional variables.

Question 4 (6 points)

Now rerun the regression of `bweight` onto `mbsmoke` but add maternal age (`mage`) as a control variable, save the output as `fit2`, display and interpret the results.

```
fit2 <- lm(bweight ~ mbsmoke + mage, data = data)
tidy(fit2)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  3177.       41.0      77.5  0
## 2 mbsmoke      -261.       21.5     -12.1  2.58e-33
## 3 mage            8.79      1.49      5.90 3.88e- 9
```

Prompt:

6. Compared to the previous regression, how did including `mage` change the coefficient on `mbsmoke`?

When we included `mage`, the coefficient on `mbsmoke` went from -275 to -261. This suggests that there had been some selection bias assuming that the `mage` was a confounder causing both the treatment and the outcome. The fact that the coefficient went down also indicates that some of the difference in means can be attributed to differences in maternal age in the two groups of women. Some of this was simply because the treatment group, in other words, had younger women. You might show them this, though I don't here. The treatment group of smokers had an average `mage` of 25.2 whereas the non-smokers had an average `mage` of 26.8. Once we include `mage` as a covariate in the regression, the conditional on maternal age, difference in mean birthweight was slightly smaller than we found in the SDO.

Question 5 (6 points)

Extend the model to include parental education levels (`medu` and `fedu`). Save this as `fit3`.

```
fit3 <- lm(bweight ~ mbsmoke + mage + medu + fedu, data = data)
tidy(fit3)
```

```
## # A tibble: 5 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  3074.       51.0      60.3  0
## 2 mbsmoke      -242.       21.9     -11.0  6.50e-28
## 3 mage            6.12      1.64      3.74 1.89e- 4
## 4 medu            5.29      4.15      1.27 2.03e- 1
## 5 fedu            8.41      2.76      3.05 2.33e- 3
```

Prompt:

7. Compare the smoking coefficient across the three regressions (`fit1`, `fit2`, `fit3`).

When we include `mage`, `medu`, and `fedu`, the coefficient falls even further. In `fit2`, the coefficient on `mbsmoke` had been -261, but with these three covariate controls, the coefficient is now -242. This again because of a correlation between the treatment variable and each of these, which can also be shown if you wanted by simply documented the imbalance. Treatment group had lower `mage` (maternal age), lower `medu` (maternal education) by around 1.3 years of schooling, and lower `fedu` (father education) by almost 2 years.

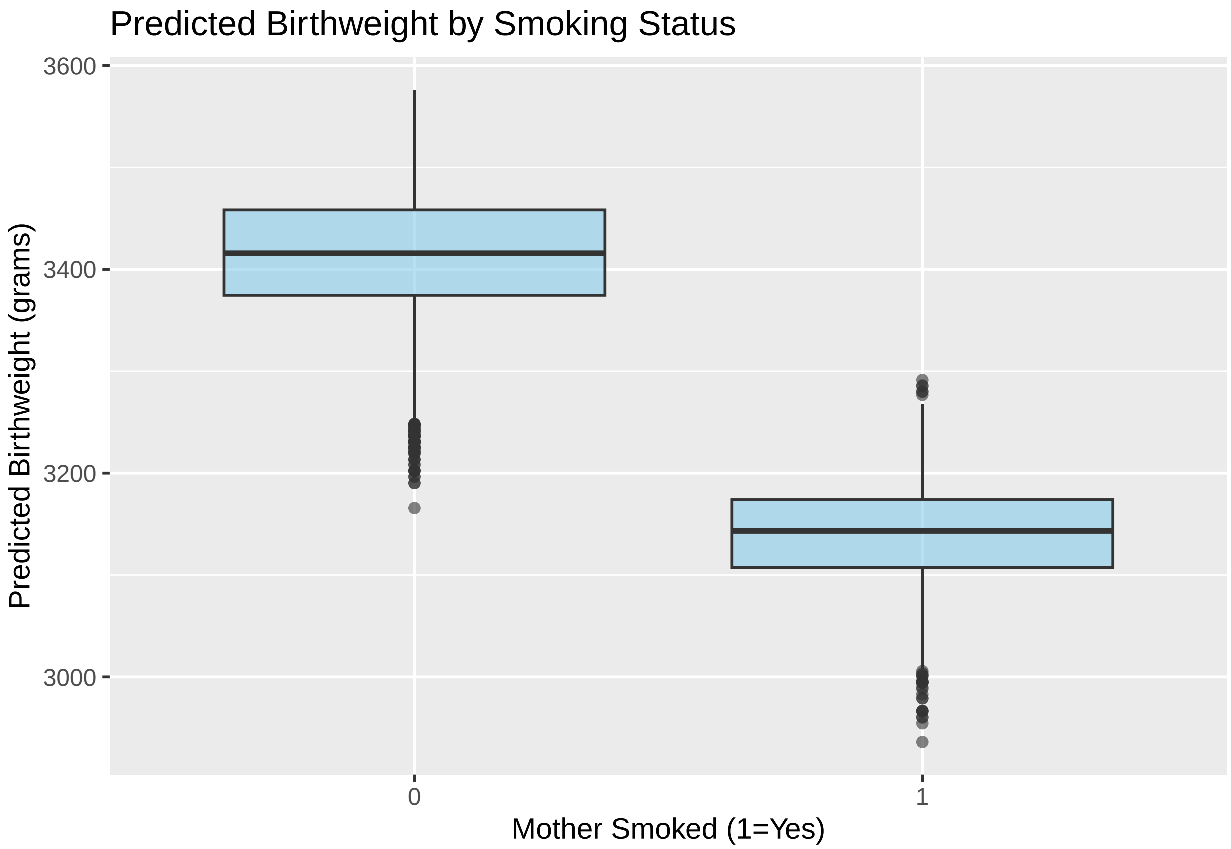
8. What does the pattern suggest about the importance of omitted variables in estimating the causal effect of maternal smoking on infant birth weight?

The omitted variables are not merely variables we did not control for which caused the outcome, but rather, variables we did not control for which were imbalanced between the two groups. Recall that under independence, balance is guaranteed in a large sample, but we do not have it here. OLS is operating similarly to matching in that it's finding differences in means across values of the covariates and then predicting those outcomes given each person's unique values of the covariates. We can see this in the next question visually.

Question 6 (5 points)

Similar to question 3, create a visualization of predicted birthweight for smoking vs. nonsmoking mothers using the model that controlled for `mage`, `medu` and `fedu` (`fit3`).

```r
data$predicted <- predict(fit3)
ggplot(data, aes(x = factor(mbsmoke), y = predicted)) +
  geom_boxplot(fill = "skyblue", alpha = 0.6) +
  labs(x = "Mother Smoked (1=Yes)", y = "Predicted Birthweight (grams)",
       title = "Predicted Birthweight by Smoking Status")
```



Prompt:

9. Compare this visualization to the visualization we did in Question 3 – how are they similar and how are they different?

Students should be encouraged to compare the visualization from question 3 to the visualization in question 6. Notice that the visualization in question 3 is two means a flat line. Students should be asked why that would be the case, but the visualization in question 6 shows thicker boxes with a horizontal black line. What might be going on here?

What is going on is the fitted values differ within the treatment versus the control group because of how many different values of `medu`, `fedu` and `mage` there are in the sample. The number of combinations yield different values of $\widehat{y_i}$ – the predicted value of birth weight given treatment state and the three covariates – and thus we end up with a thick box representing variation in predicted outcomes. But the underlying mechanics of prediction, as can be seen in the `predict(fit3)` command, is the same.

Question 7 (2 points)

Compute and interpret the R-squared for each regression model.

```
rsq_values <- data.frame(
  Model = c("fit1", "fit2", "fit3"),
  R2 = c(summary(fit1)$r.squared,
         summary(fit2)$r.squared,
         summary(fit3)$r.squared)
)
rsq_values
```

```
##   Model         R2
## 1  fit1 0.03426361
## 2  fit2 0.04145808
## 3  fit3 0.04531109
```

Prompt:

10. How did the explanatory power of the model as measured using R-squared change as you included more and more covariates as controls?

As we include more covariates, the R-squared grows. This is mechanical – insofar as the covariate in the model has *any* correlation with the outcome, then we will explain more of it, and the R-squared will increase. You might say to the student that the R-squared “weakly increases” as we include more covariates. It can go up, but it cannot go down, as we add a control in (on top of what we already had) and the stronger the correlation between that variable and the outcome (conditional on controls), the larger the increase in R-squared it is.

You might then say that ultimately the value of the R-squared differs depending on whether we are trying to explain variation in the outcome versus identification of causal effects. The tasks are distinct. In causal inference, we are looking for unbiased estimates of the treatment effect, but in pure prediction, it’s more like we are trying to accurately predict the outcome. Students should be encouraged to compare those two tasks and articulate their own underestanding of how they differ, but also how the OLS regression does them both (but stores that information in different places).

Question 8 (6 points)

Run a regression including the complete set of conditioning variables below. Compare the coefficient on `mbsmoke` from this regression to earlier models.

```
covars <- c("mage", "medu", "mmarried", "mhisp", "foreign", "fage", "fedu", "fhisp", "frace", "deadkids

formula_full <- as.formula(
  paste("bweight ~ mbsmoke +", paste(covars, collapse = " + "))
)

fit_full <- lm(formula_full, data = data)
broom::tidy(fit_full)
```

```
## # A tibble: 18 x 5
##    term        estimate std.error statistic  p.value
##    <chr>          <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept) 2893.       63.2      45.7   0
##  2 mbsmoke     -226.       22.0     -10.3   1.52e-24
##  3 mage           0.975     2.14      0.455 6.49e- 1
##  4 medu           4.79      4.24      1.13  2.58e- 1
##  5 mmarried      49.1      23.8       2.07  3.89e- 2
##  6 mhisp         65.5      66.3       0.989 3.23e- 1
##  7 foreign      -19.3      39.4      -0.491 6.24e- 1
##  8 fage          -1.65      1.19     -1.39  1.65e- 1
##  9 fedu          -0.779     3.07     -0.253 8.00e- 1
## 10 fhisp        -92.1      62.7      -1.47  1.42e- 1
## 11 frace        213.       24.8       8.58  1.27e-17
## 12 deadkids     -25.0      18.8      -1.33  1.85e- 1
## 13 nprenatal     32.5       2.53     12.8   5.73e-37
## 14 monthslb      -0.221     0.332    -0.664 5.07e- 1
## 15 order         20.9      10.8       1.93  5.38e- 2
## 16 prenatal1    -91.4      24.0      -3.81  1.41e- 4
## 17 alcohol      -29.6      46.4      -0.637 5.24e- 1
## 18 fbaby        -66.7      26.9      -2.48  1.30e- 2
```

```r
full_rsq_values <- data.frame(
  Model = c("fit_full"),
  R2 = c(summary(fit_full)$r.squared)
)

full_rsq_values
```

```
##      Model        R2
## 1 fit_full 0.1102817
```

Prompt:

11. How does the coefficient on `mbsmoke` change as we add all covariates?

Now we include even more covariates and when we do the the coefficient on `mbsmoke` falls to -226. Thus the imbalance in these covariates combined with their partial association with the outcome suggests that simply trying to account for that imbalance can remove some of the bias.

Conceptually, the challenge for the TF is the degree to which they want to connect matching's way of re-balancing the covariates (a balancing property is more or less the result of matching) and how OLS regressions do so. OLS regression fit coefficients which are then used to predict outcomes before then taking a difference in means. We call this a functional form based prediction. It does not directly rebalance; rather we fit lines (even in higher dimensional spaces). But the goal is similar in both cases – address the piece of the SDO that can be attributed to difference in the distribution of covariates through either explicit rebalancing (matching) or fitted predictions (OLS).

12. How did the R-squared change as we included all covariates?

Again, the inclusion of more covariates pushed up the R-squared from around 0.045 to 0.11. Our model now can account for around 11% of all variation in birth weight. There is still a lot of the variation that is *not* explained by these covariates, though.

13. Why might including many controls strengthen or weaken the estimated effect?

Insofar as the two groups have differences in the average values of these covariates, then the SDO will be partly based on the unique role each has in predicting the outcome. Thus if `alcohol` is much more common among the treatment group (which it is) than the control group, then simply accounting for that can push the coefficient on the treatment variable away from the earlier regression coefficients.

Summary (Discussion)

Summarize your findings about the relationship between maternal smoking and infant birthweight. Do your regression results support the hypothesis that smoking causes lower birthweight? What assumptions are required for this conclusion to be causal?