1 Regression Basics

Suppose we are interested in estimating the effect of a parent's incarceration on the child's income as an adult. We run the following linear regression:

$$Income_i = \alpha_0 + \alpha_1 ParentIncarcerated_i + u_i$$

1. Understanding OLS

a. What is the minimization problem solved by OLS?

b. What are the solutions to this minimization problem?

2. Regression components

a. The independent variable is:

b. The dependent variable is:

c. The error term is (how do we interpret the error term?):

d. The coefficient of interest is:

Now suppose that we run this regression in Stata and obtain the estimators $\hat{\alpha}_0$ and $\hat{\alpha}_1$:

```
. regress wages parent_incarcerated, r

Linear regression                               Number of obs     =      4,006
                                                F(1, 4004)        =      26.74
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0023
                                                Root MSE          =      52581

------------------------------------------------------------------------------
             |               Robust
       wages | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
parent_incarcerated |  -18257.69   3530.935    -5.17   0.000    -25180.29   -11335.09
       _cons |   57676.92   844.3675    68.31   0.000     56021.49    59332.35
------------------------------------------------------------------------------
```

3. Reading Stata output

a. What are the values of our estimates of α_0 and α_1?

b. Can we reject the null hypothesis $\alpha_1 = 0$? Why?

c. What is the confidence interval on α_1? How is it calculated?

d. What is the R^2, and what does it tell us?

e. How is the t-statistic estimated?

4. Regression interpretation

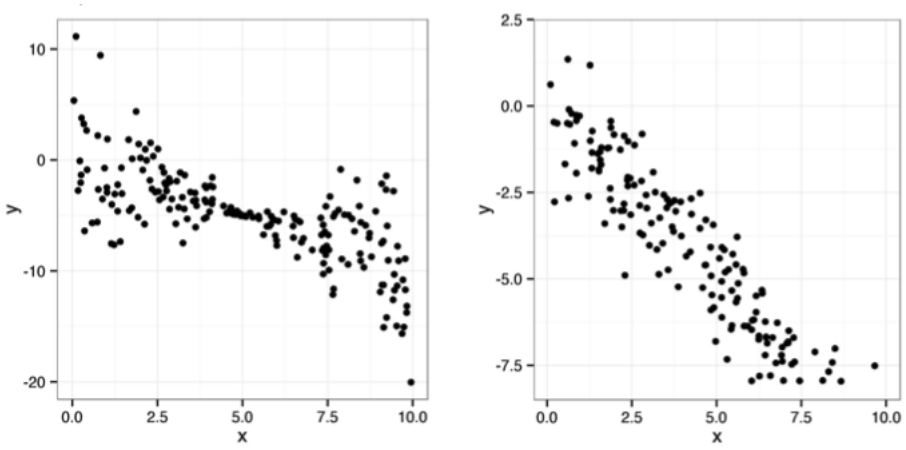
a. How do we interpret the coefficient on parent incarceration?

b. Does anything in the regression output tell us whether this is a causal estimate?

5. Homoskedasticity and heteroskedasticity

Last week, we discussed using robust standard errors when comparing means between groups with unequal variances. When our predictor is *continuous*, we say that the error term is **homoskedastic** if its variance is the same at all values of x: $\text{Var}(u_i|x_i) = c \; \forall \; i$. The error term is **heteroskedastic** if its variance differs depending on the value of x. Just as in the binary predictor case, using the `robust` or `hc2` option will provide unbiased standard errors whether the error term is heteroskedastic or homoskedastic, so it is always better to use one of the two.

Which of the graphs below shows data with heteroskedasticity? Which one shows homoskedasticity?



2 Omitted variable bias (OVB)

6. For there to be OVB, the omitted variable has to be:

1.

2.

7. What assumption of our regression is violated by omitted variable bias?

Going back to our example, suppose we think that parents' income would impact both their likelihood of incarceration and the child's future income. In this case, the "true" regression (or model) we are interested in is

$$Income_i = \beta_0 + \beta_1 ParentIncarcerated_i + \beta_2 ParentIncome_i + v_i$$

We can rewrite the $\hat{\alpha}_1$ we previously estimated as:

$$\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\gamma}_1$$

where $\hat{\gamma}_1$ is the slope coefficient on $ParentIncarcerated_i$ from an auxiliary regression of the omitted variable (parent income) on all of the included regressors (in this case, just $ParentIncarcerated_i$):

$$ParentIncome_i = \gamma_0 + \gamma_1 ParentIncarcerated_i + \varepsilon_i$$

We now have three equations:

1. Short form regression: $Income_i = \alpha_0 + \alpha_1 ParentIncarcerated_i + u_i$
2. Long form regression: $Income_i = \beta_0 + \beta_1 ParentIncarcerated_i + \beta_2 ParentIncome_i + v_i$
3. Auxiliary regression: $ParentIncome_i = \gamma_0 + \gamma_1 ParentIncarcerated_i + \varepsilon_i$

8. Given the three equations, derive the omitted variable bias formula ($\hat{\alpha}_i = \hat{\beta}_1 + \hat{\beta}_2 \hat{\gamma}_1$).

9. In this case, in which direction (positive or negative) is the bias likely to go? What part of the OVB formula reflects this?

Let's see what happens when we include a control for parent income in the regression:

```
. regress wages parent_incarcerated parent_income, r

Linear regression                               Number of obs     =      4,006
                                                F(2, 4003)        =      72.13
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0593
                                                Root MSE          =      51063

------------------------------------------------------------------------------
                     |               Robust
               wages | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
---------------------+--------------------------------------------------------------
parent_incarcerated |  -14484.39   3662.767    -3.95   0.000    -21665.45   -7303.323
      parent_income |   .2963409   .0265822    11.15   0.000     .2442249    .3484568
              _cons |   43176.54   1287.436    33.54   0.000     40652.45    45700.63
------------------------------------------------------------------------------
```

10. What is the interpretation of the coefficient on parent_incarcerated in the "long" regression? How does it differ from the "short" regression (the one not controlling for parent income)?

11. Are any of the following variables potential omitted variables (that could cause bias)? If so, which way do you think the bias would go? If not, why not?

a. Gender

b. Year of birth

c. Unemployment rate

3 Perfect multicollinearity and residual regression

Suppose we are interested in estimating the effect of education on income. We run the following linear regression:

$$Income_i = \alpha_0 + \alpha_1 Education_i + u_i$$

Now suppose that in the regression of income on education, we would like to control for person i's age and their years of work experience. In the data, we only observe their age, so we decide to define a new variable $Experience_i = Age_i - Education_i - 6$.

Recall from class that in a regression of the form $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$, we can obtain the same β_1 by running

$$\tilde{Y}_i = b_0 + \beta_1 \tilde{X}_i + u_i$$

where $\tilde{Y}_i$ and $\tilde{X}_i$ are the residuals from the regressions of Y and X on W:

$$X_i = \gamma_0 + \gamma_1 W_i + \tilde{X}_i$$
$$Y_i = \alpha_0 + \alpha_1 W_i + \tilde{Y}_i$$

12. Say we run a regression of our new experience variable on the age and education variables.

a. The constant would be:

b. The coefficient on Age_i would be:

c. The coefficient on $Education_i$ would be:

d. The R^2 would be:

e. The residuals would be:

13. What would happen if we tried to run our regression of income on education, age, and our new work experience variable?

Now suppose that we want to control for region in our returns to schooling regression. Region can be North, South, East, or West, and we have a dummy variable for each ($North_i = 1$ if person i lives in the North, 0 otherwise).

14. What would be a regression specification to control for region?

15. How do we interpret the coefficients on the dummy variables?