

# **THE MULTIPLE REGRESSION MODEL I**

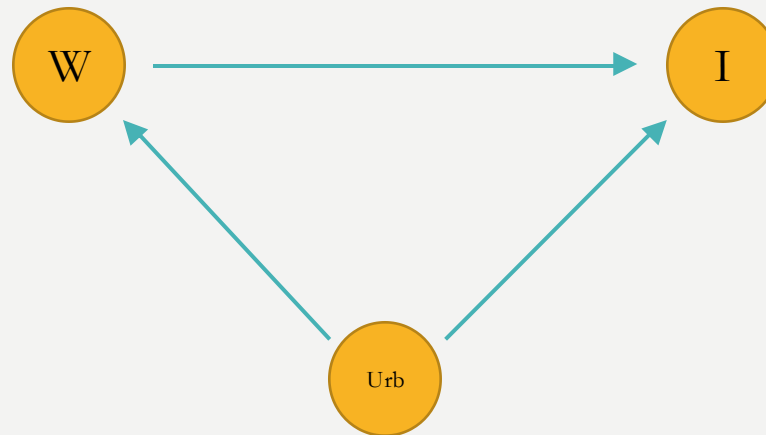
**PROF. SEBASTIÁN VALLEJO VERA**

- The primary drawback in using simple regression analysis for empirical work is that it is very difficult to draw *ceteris paribus* conclusions about how  $x$  affects  $y$ : the key assumption SLR.4—that all other factors affecting  $y$  are uncorrelated with  $x$ ,  $E(\mu|x) = 0$ —is often unrealistic.

# EXAMPLE

- Go to code for 7 Multiple Regression Model I.R / Vignette 6.1

- In this case, we have the following DAG:



Or, we have the following population regression functions:

$$\begin{aligned} \text{income} &= 1 + 2 * \text{urban} - 0.3 * \text{wages} + \mu \\ \text{wages} &= 3 * \text{urban} + v \end{aligned}$$

Yet we are estimating the following OLS regression:

$$\widehat{\text{income}}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \text{wages}_i + \widehat{\mu}_i$$

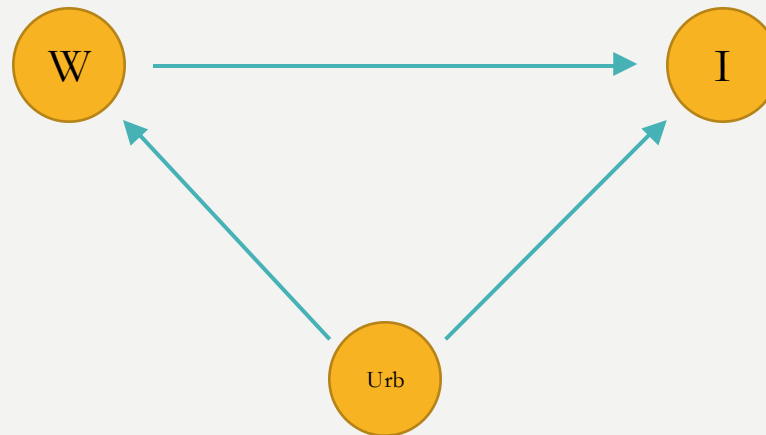
Where did *urban* go?

- That's right, to the error term,  $\hat{\mu}$ .
- BUT... if *urban* is going to the error term,  $\hat{\mu}$ , and we also know that

$$wages = 3 * urban + v$$

- Then  $E(\mu|wages) \neq 0$  !! We are violating the very important SLR.4, thus rendering our OLS estimates **biased**...

- What to do?



- We can remove everything from wages that is explained by urban, and remove everything income that is explained by urban, and we would have left...

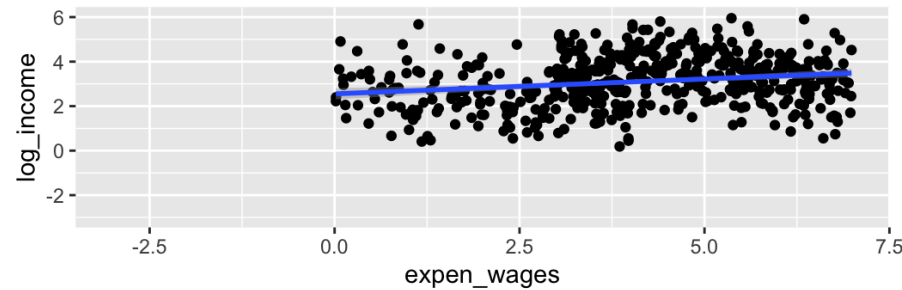
- Go to code for 7 Multiple Regression Model I.R / Vignette 6.2



- Here is a graphical representation of what just happened:

factor(urban) 0 1

0. Relation between wages and income. Beta = 0.13



1. Relation between wages and income divided by urban.



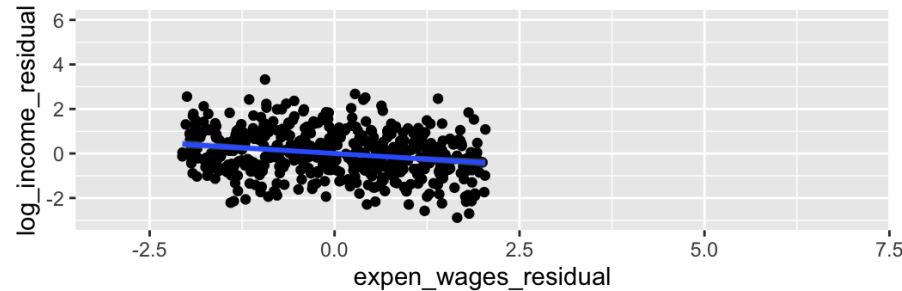
2. We remove the difference of wages explained by urban.

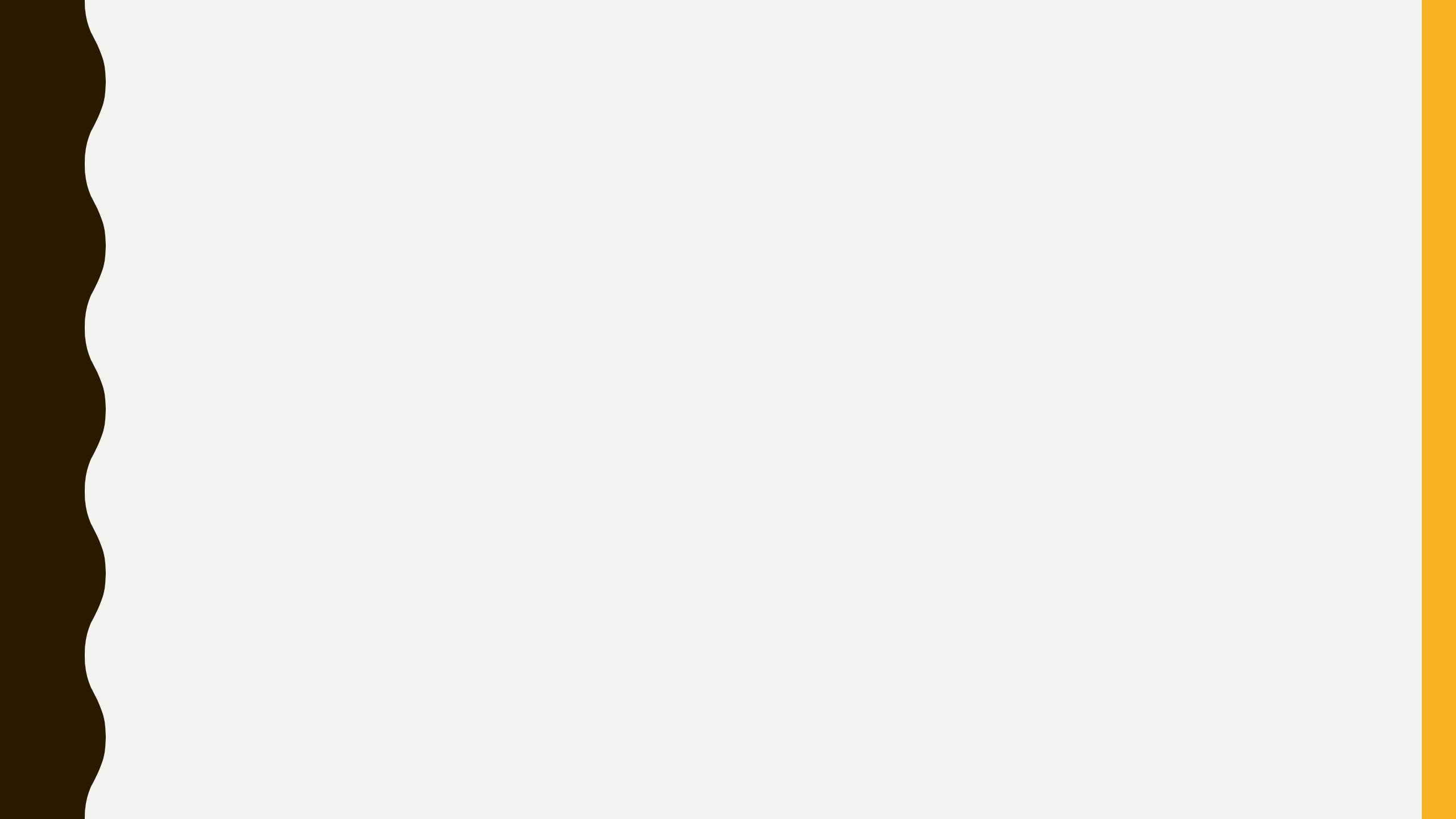


3. We remove the difference of income explained by urban.



4. We analyze what is left. Beta = -0.22





- We call this, to...

# CONTROL

(A VARIABLE)



- Multiple regression analysis is more amenable to *ceteris paribus* analysis because it allows us to explicitly **control** for many other factors that simultaneously affect the dependent variable (i.e., by removing factors in  $\mu$  and estimating their impact on  $y$ ).

The general **multiple linear regression model** (also called the *multiple regression model*) with  $k$  independent variables can be written in the population as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \mu$$

where:

- $\beta_0$  is the intercept
- $\beta_1$  is the slope parameter associated with  $x_1$
- $\beta_2$  is the slope parameter associated with  $x_2$ , and so on
- $\mu$  is the error term or disturbance and contains all the other factors that affect  $y$  other than  $x_1, x_2, x_3, \dots, x_k$ .

In the general case with  $k$  independent variables, we seek estimates  $\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_k$  in the equation:

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \dots + \widehat{\beta}_k x_k$$

The OLS estimates,  $k + 1$  of them, are chosen to minimize the sum of squared residuals in a sample of  $n$  observations:

$$\sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} + \dots + \widehat{\beta}_k x_{ik})^2$$



From the **OLS regression line** (also called the sample regression function, SRF):

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_k x_k$$

the estimates  $\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_k$  have partial effect, or *ceteris paribus*, interpretations.

For example, the coefficient on  $x_1$  measures the change in  $\hat{y}$  due to a one-unit increase in  $x_1$ , holding all other independent variables ( $x_2, x_3, \dots, x_k$ ) fixed.

That is,  $\Delta \hat{y} = \widehat{\beta}_1 \Delta x_1$ , holding  $x_2, x_3, \dots, x_k$  fixed. Thus, we have controlled for the variables  $x_2, x_3, \dots, x_k$  when estimating the effect of  $x_1$  on  $y$ .

The intercept  $\beta_0$  is the predicted value of  $y$  when  $x_1 = 0, x_2 = 0, \dots$ , and  $x_k = 0$ . (Generally nonsensical)

From the **OLS regression line**, we can obtain a fitted or predicted value for each observation  $i$ :

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik}$$

by plugging the values of the independent variables for observation  $i$  into the equation.


The residual for observation  $i$  is defined just as in the simple regression case:

$$\hat{\mu}_i = y_i - \hat{y}_i$$

There is a residual for each observation. If  $\hat{\mu} > 0$ , then  $\hat{y}_i$  is below  $y_i$ , which means that, for this observation,  $y_i$  is underpredicted. If  $\hat{\mu} < 0$ , then  $\hat{y}_i$  is greater than  $y_i$  and  $y_i$  is overpredicted.

The OLS fitted values and residuals have some important properties that are immediate extensions from the single variable case:

1. The sample average of the residuals is zero ( $\sum_{i=1}^n \hat{\mu}_i = 0$ ) and so  $\hat{\bar{y}} = \bar{y}$
2. The sample covariance between each independent variable and the OLS residuals is zero ( $\sum_{i=1}^n x_{ij} \hat{\mu}_i = 0$  for all  $j$ ). Consequently, the sample covariance between the OLS fitted values and the OLS residuals is zero ( $\sum_{i=1}^n \hat{y}_i \hat{\mu}_i = 0$ )
3. The point  $(\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_k, \bar{y})$  is always on the OLS regression line:  $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \dots + \hat{\beta}_k \bar{x}_k$



# **IN-CLASS EXERCISES**

# PART I

Using the 'fertil2' dataset from 'wooldridge' on women living in the Republic of Botswana in 1988,

- estimate the regression equation of the number of children on education, age of the mother and electricity;
- interpret  $\widehat{\beta}_0, \widehat{\beta}_{education}, \widehat{\beta}_{age}, \widehat{\beta}_{electricity}$  ;
- verify that  $\sum_{i=1}^n \widehat{y}_i \widehat{\mu}_i = 0$ ;
- verify that  $\sum_{i=1}^n education_i \widehat{\mu}_i = 0$ ;
- re-estimate the regression equation of the number of children on education, age of the mother and electricity but also include the square of age of the mother. How do you now interpret the effect of age on the number of children?

# PART II

Using the 'wage1' dataset from 'wooldridge' on US workers in 1976,

- estimate the regression equation of the hourly wage (*wage*) in dollars (\$) on education (*educ*) and experience (*exper*);
- how do  $\widehat{\beta}_0$ ,  $\widehat{\beta}_{educ}$ ,  $\widehat{\beta}_{exper}$  change when the units of wages are in ¢ instead of \$;
- compare the estimated coefficient for education to the one obtained by regressing only the hourly wage on education. Interpret.