BS2280 - Econometrics 1

Lecture 5 - Part 1: Multiple Regression Analysis I

Dr. Yichen Zhu

Structure of today's lecture

Intended Learning Outcomes

- Understanding the differences between a simple and a multiple regression model
- Interpret the coefficients of the multiple regression model

Background

- Simple regression model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{1}$$

- Assumes that Variable Y is affected by only **one** variable X on the right-hand side
- Variations in Y could be sufficiently explained by variations in X only
- That is often too **simplistic**!!!
- More likely the case that several (observed) variables X will affect Y
- Example
- What factors other than years of schooling can affect wages of graduates?

$$EARNINGS_i = \beta_1 + \beta_2 S_i + u_i$$

Multiple Regression Model: Notations

- The multiple regression model allows two or more X variables in the model
- Hence, Y will depend on several X variables
- How do we symbolise these variables in our multiple regression model?

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

- Different textbooks use different notations!
- Example
- We now extend the simple regression model by adding another variable to it, i.e. out-of-school years of experience (EXP)

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

Multiple Regression Model: Notations

- The multiple regression model allows two or more X variables in the model
- Hence, Y will depend on several X variables
- How do we symbolise these variables in our multiple regression model?

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

- Different textbooks use different notations!
- Example
- We now extend the simple regression model by adding another variable to it, i.e. out-of-school years of experience (EXP)

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

Multiple Regression Model: Notations

- The multiple regression model allows two or more X variables in the model
- Hence, Y will depend on several X variables
- How do we symbolise these variables in our multiple regression model?

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

- Different textbooks use different notations!
- Example
- We now extend the simple regression model by adding another variable to it, i.e. out-of-school years of experience (EXP)

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

Multiple Regression Model: Notations

- The multiple regression model allows two or more X variables in the model
- Hence, Y will depend on several X variables
- How do we symbolise these variables in our multiple regression model?

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

- Different textbooks use different notations!
- Example
- We now extend the simple regression model by adding another variable to it, i.e. out-of-school years of experience (EXP)

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

Multiple Regression Model: Notations

- The multiple regression model allows two or more X variables in the model
- Hence, Y will depend on several X variables
- How do we symbolise these variables in our multiple regression model?

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$$

- Different textbooks use different notations!
- Example
- We now extend the simple regression model by adding another variable to it, i.e. out-of-school years of experience (EXP)

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

Multiple Regression Model: Notations

- The multiple regression model allows two or more X variables in the model
- Hence, Y will depend on several X variables
- How do we symbolise these variables in our multiple regression model?

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$$

- Different textbooks use different notations!
- Example
- We now extend the simple regression model by adding another variable to it, i.e. out-of-school years of experience (EXP)

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

Multiple Regression Model: Considerations

1. Which of the many potentially important X_{ki} variables are relevant to the model?
2. How can we introduce these X variables in our regression model? Linear or Non-linear?
3. How can we distinguish between the effect of each one of the X_{ki} variables on Y_i ?

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- Know the specific effect of each independent variable!
- How can we make sure that we get the effects of one more year of schooling (S_i) on hourly earnings ($EARNINGS_i$)?
- How can we get the effects of one more out-of-school year of experience (EXP_i) on hourly earnings ($EARNINGS_i$)?

Multiple Regression Model: Considerations

1. Which of the many potentially important X_{ki} variables are relevant to the model?
2. How can we introduce these X variables in our regression model? Linear or Non-linear?
3. How can we distinguish between the effect of each one of the X_{ki} variables on Y_i ?

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- Know the specific effect of each independent variable!
- How can we make sure that we get the effects of one more year of schooling (S_i) on hourly earnings ($EARNINGS_i$)?
- How can we get the effects of one more out-of-school year of experience (EXP_i) on hourly earnings ($EARNINGS_i$)?

Multiple Regression Model: Considerations

1. Which of the many potentially important X_{ki} variables are relevant to the model?
2. How can we introduce these X variables in our regression model? Linear or Non-linear?
3. How can we distinguish between the effect of each one of the X_{ki} variables on Y_i ?

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- Know the specific effect of each independent variable!
- How can we make sure that we get the effects of one more year of schooling (S_i) on hourly earnings ($EARNINGS_i$)?
- How can we get the effects of one more out-of-school year of experience (EXP_i) on hourly earnings ($EARNINGS_i$)?

Multiple Regression Model: Considerations

1. Which of the many potentially important X_{ki} variables are relevant to the model?
2. How can we introduce these X variables in our regression model? Linear or Non-linear?
3. How can we distinguish between the effect of each one of the X_{ki} variables on Y_i ?

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- Know the specific effect of each independent variable!
- How can we make sure that we get the effects of one more year of schooling (S_i) on hourly earnings ($EARNINGS_i$)?
- How can we get the effects of one more out-of-school year of experience (EXP_i) on hourly earnings ($EARNINGS_i$)?

Multiple Regression Model: Considerations

1. Which of the many potentially important X_{ki} variables are relevant to the model?
2. How can we introduce these X variables in our regression model? Linear or Non-linear?
3. How can we distinguish between the effect of each one of the X_{ki} variables on Y_i ?

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- Know the specific effect of each independent variable!
- How can we make sure that we get the effects of one more year of schooling (S_i) on hourly earnings ($EARNINGS_i$)?
- How can we get the effects of one more out-of-school year of experience (EXP_i) on hourly earnings ($EARNINGS_i$)?

Multiple Regression Model: Considerations

1. Which of the many potentially important X_{ki} variables are relevant to the model?
2. How can we introduce these X variables in our regression model? Linear or Non-linear?
3. How can we distinguish between the effect of each one of the X_{ki} variables on Y_i ?

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- Know the specific effect of each independent variable!
- How can we make sure that we get the effects of one more year of schooling (S_i) on hourly earnings ($EARNINGS_i$)?
- How can we get the effects of one more out-of-school year of experience (EXP_i) on hourly earnings ($EARNINGS_i$)?

Multiple Regression Model: Estimations

- Once we have decided how many X we want in the model, we start by estimating the coefficients.

Simple Regression Model	**Multiple Regression Model**
$Y_i = \beta_1 + \beta_2 X_i + u_i$ $\hat{\beta}_1$ and $\hat{\beta}_2$	$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$ $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, ..., $\hat{\beta}_k$
$Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$	$Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + ... + \hat{\beta}_k X_{ki} + \hat{u}_i$
$\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$	$\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - ... - \hat{\beta}_k X_{ki}$
min $RSS = \sum_{i=1}^{n} \hat{u}_i{}^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$	min $RSS = \sum_{i=1}^{n} \hat{u}_i{}^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - ... - \hat{\beta}_k X_{ki})^2$
FOC's. $\frac{\partial RSS}{\partial \hat{\beta}_1} = 0$ and $\frac{\partial RSS}{\partial \hat{\beta}_2} = 0$	FOC's. $\frac{\partial RSS}{\partial \hat{\beta}_1} = 0$, $\frac{\partial RSS}{\partial \hat{\beta}_2} = 0$, $\frac{\partial RSS}{\partial \hat{\beta}_3} = 0$, ..., $\frac{\partial RSS}{\partial \hat{\beta}_k} = 0$
$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$ $\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$	labour intensive and complex Such calculations are best done using matrix algebra

Example: Determinants of Earnings

- We used a simple regression model to analyse the impact of years of schooling on hourly wags.

$$EARNINGS_i = \beta_1 + \beta_2 S_i + u_i$$

```
> lm(EARNINGS~S, data=EAWE21)

Call:
lm(formula = EARNINGS ~ S, data = EAWE21)

Coefficients:
(Intercept)            S
     0.7647       1.2657
```

$$\widehat{EARNINGS}_i = 0.765 + 1.266 S_i$$

Example: Determinants of Earnings

- We now extend the simple regression model by adding another variable to it, i.e. out-of-school years of experience (*EXP*)

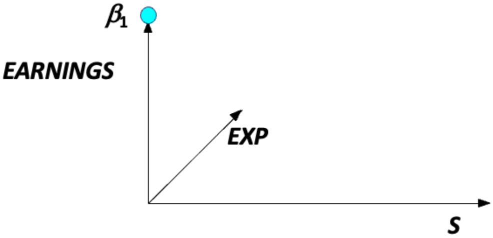
$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

```
> lm(EARNINGS~S+EXP, data=EAWE21)

Call:
lm(formula = EARNINGS ~ S + EXP, data = EAWE21)

Coefficients:
(Intercept)            S          EXP
   -14.6683       1.8776       0.9833
```

$$\widehat{EARNINGS}_i = -14.668 + 1.877 S_i + 0.983 EXP_i$$

Multiple Regression Model: Geometrical Interpretation

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

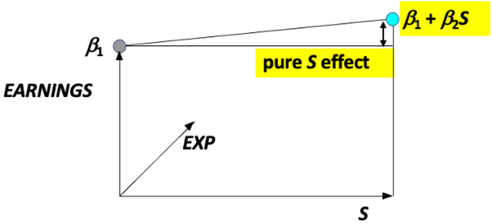


- The model has three dimensions, one each for *EARNINGS*, *S*, and *EXP*.
- The starting point for investigating the determination of *EARNINGS* is the intercept, β_1.
- Literally the intercept gives *EARNINGS* for those respondents who have no schooling and no work experience. However, there were no respondents with less than 6 years of schooling. Hence a literal interpretation of β_1 would be unwise.

Multiple Regression Model: Geometrical Interpretation

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- The next term on the right side of the equation gives the effect of variations in S.
- Pure S effect

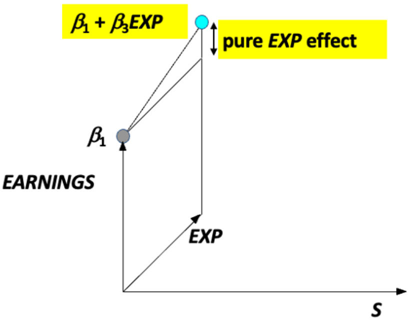


- A one year increase in S causes $EARNINGS$ to increase by β_2 dollars, **holding** EXP **constant**.

Multiple Regression Model: Geometrical Interpretation

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- The third term gives the effect of variations in *EXP*
- Pure *EXP* effect

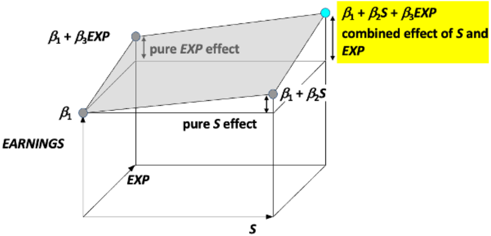


- A one year increase in *EXP* causes earnings to increase by β_3 dollars, **holding *S* constant**.

Multiple Regression Model: Geometrical Interpretation

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- Combine effects of S and EXP

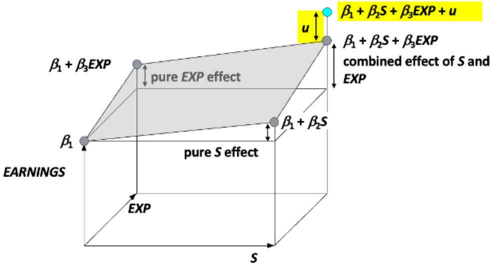


- Different combinations of S and EXP give rise to values of $EARNINGS$ which lie on the plane shown in the diagram, defined by the equation $EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i$.
- This is the nonstochastic (nonrandom) component of the model.

Multiple Regression Model: Geometrical Interpretation

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- The final element of the model is the disturbance term, u_i.

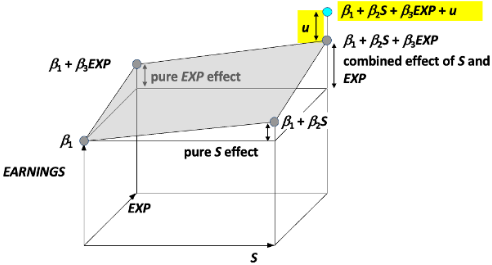


- This causes the actual values of *EARNINGS* to deviate from the plane.
- In this observation, u_i happens to have a positive value.
- This is the stochastic (random) component of the model.
- A sample consists of a number of observations generated in this way. Note that the interpretation of the model does not depend on whether S and EXP are correlated or not.

Multiple Regression Model: Geometrical Interpretation

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- The final element of the model is the disturbance term, u_i.

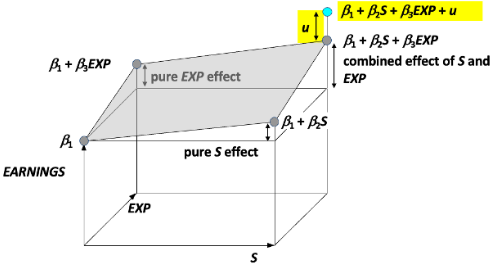


- This causes the actual values of *EARNINGS* to deviate from the plane.
- In this observation, u_i happens to have a positive value.
- This is the stochastic (random) component of the model.
- A sample consists of a number of observations generated in this way. Note that the interpretation of the model does not depend on whether S and EXP are correlated or not.

Multiple Regression Model: Geometrical Interpretation

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- The final element of the model is the disturbance term, u_i.

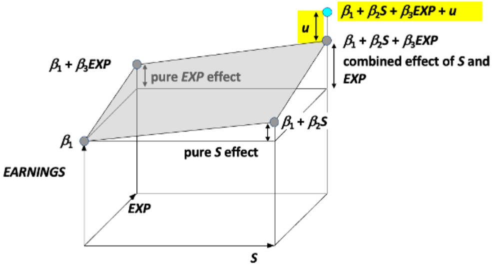


- This causes the actual values of *EARNINGS* to deviate from the plane.
- In this observation, u_i happens to have a positive value.
- This is the stochastic (random) component of the model.
- A sample consists of a number of observations generated in this way. Note that the interpretation of the model does not depend on whether *S* and *EXP* are correlated or not.

Multiple Regression Model: Geometrical Interpretation

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- The final element of the model is the disturbance term, u_i.

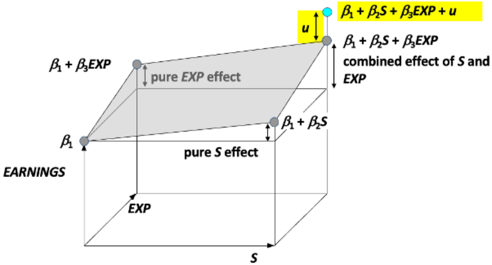


- This causes the actual values of *EARNINGS* to deviate from the plane.
- In this observation, u_i happens to have a positive value.
- This is the stochastic (random) component of the model.
- A sample consists of a number of observations generated in this way. Note that the interpretation of the model does not depend on whether *S* and *EXP* are correlated or not.

Multiple Regression Model: Geometrical Interpretation

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- The final element of the model is the disturbance term, u_i.



- This causes the actual values of *EARNINGS* to deviate from the plane.
- In this observation, u_i happens to have a positive value.
- This is the stochastic (random) component of the model.
- A sample consists of a number of observations generated in this way. Note that the interpretation of the model does not depend on whether S and EXP are correlated or not.

Interpretation of Coefficients

Simple Regression Model	Multiple Regression Model
$Y_i = \beta_1 + \beta_2 X_i + u_i$ $\hat{\beta}_1$ and $\hat{\beta}_2$	$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$ $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, ..., \hat{\beta}_k$
$\hat{\beta}_1$: Intercept	$\hat{\beta}_1$: Intercept
$\hat{\beta}_2$: A one unit change in X leads to a $\hat{\beta}_2$ unit change in Y	$\hat{\beta}_2$: On average, a one unit change in X_2 leads to a $\hat{\beta}_2$ unit change in Y, **controlling for the effects of other** X **variables**
	$\hat{\beta}_3$: On average, a one unit change in X_3 leads to a $\hat{\beta}_3$ unit change in Y, **controlling for the effects of other** X **variables** ...
	$\hat{\beta}_k$: On average, a one unit change in X_k leads to a $\hat{\beta}_k$ unit change in Y, **controlling for the effects of other** X **variables**

Interpretation of Coefficients

What does controlling for the effects of other variables mean?

- Holding all other variables constant: Other X variables do not change when specific X variable of interest is changing
- If they all change at the same time, it would be difficult to assess the effect of a change in the specific X variable on Y.
- For example, a change in X_2 variable could increase Y, but a change in X_3 variable could decrease Y and so on. This would not be informative.

Interpretation of Coefficients

What does controlling for the effects of other variables mean?

- Holding all other variables constant: Other X variables do not change when specific X variable of interest is changing
- If they all change at the same time, it would be difficult to assess the effect of a change in the specific X variable on Y.
- For example, a change in X_2 variable could increase Y, but a change in X_3 variable could decrease Y and so on. This would not be informative.

Interpretation of Coefficients

What does controlling for the effects of other variables mean?

- Holding all other variables constant: Other X variables do not change when specific X variable of interest is changing
- If they all change at the same time, it would be difficult to assess the effect of a change in the specific X variable on Y.
- For example, a change in X_2 variable could increase Y, but a change in X_3 variable could decrease Y and so on. This would not be informative.

Interpretation of Coefficients

```
> summary(earnfit2)

Call:
lm(formula = EARNINGS ~ S + EXP, data = EAWE21)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.098  -6.440  -2.113   3.782  76.907 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -14.6683     4.2884  -3.420 0.000677 ***
S             1.8776     0.2237   8.392 5.01e-16 ***
EXP           0.9833     0.2098   4.686 3.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.13 on 497 degrees of freedom
Multiple R-squared:  0.1242,	Adjusted R-squared:  0.1207 
F-statistic: 35.24 on 2 and 497 DF,  p-value: 4.86e-15
```

Interpretation of Coefficients

$$\widehat{EARNINGS}_i = -14.668 + 1.877 S_i + 0.983 EXP_i$$

- We need to attach units of measurement to X and Y as per the data set being used!!!!
- Determining whether each coefficient is statistically significant uses the same concept as with the simple regression model
- In our example:
  On average, every additional schooling year increases hourly earnings by $1.88, controlling for the effects of other X variables
  On average, every additional year of out-of-school experience completed raises hourly earnings by $0.98, ceteris paribus
- Controlling for the effects of other X variables or Ceteris paribus means: if two individuals, e.g. Yichen and Chiara, have the same years of out of school experience (EXP), then if Chiara completes an additional grade of schooling (S) compared to Yichen, we predict that Chiara will earn a $1.88 higher hourly rate.

Interpretation of Coefficients

$$\widehat{EARNINGS}_i = -14.668 + 1.877S_i + 0.983EXP_i$$

- We need to attach units of measurement to X and Y as per the data set being used!!!!
- Determining whether each coefficient is statistically significant uses the same concept as with the simple regression model
- In our example:
  On average, every additional schooling year increases hourly earnings by \$1.88, controlling for the effects of other X variables
  On average, every additional year of out-of-school experience completed raises hourly earnings by \$0.98, ceteris paribus
- Controlling for the effects of other X variables or Ceteris paribus means: if two individuals, e.g. Yichen and Chiara, have the same years of out of school experience (EXP), then if Chiara completes an additional grade of schooling (S) compared to Yichen, we predict that Chiara will earn a \$1.88 higher hourly rate.

Interpretation of Coefficients

$$\widehat{EARNINGS}_i = -14.668 + 1.877 S_i + 0.983 EXP_i$$

- We need to attach units of measurement to X and Y as per the data set being used!!!!
- Determining whether each coefficient is statistically significant uses the same concept as with the simple regression model
- In our example:
  On average, every additional schooling year increases hourly earnings by \$1.88, controlling for the effects of other X variables
  On average, every additional year of out-of-school experience completed raises hourly earnings by \$0.98, ceteris paribus
- Controlling for the effects of other X variables or Ceteris paribus means: if two individuals, e.g. Yichen and Chiara, have the same years of out of school experience (EXP), then if Chiara completes an additional grade of schooling (S) compared to Yichen, we predict that Chiara will earn a \$1.88 higher hourly rate.

Interpretation of Coefficients

$$\widehat{EARNINGS}_i = -14.668 + 1.877 S_i + 0.983 EXP_i$$

- We need to attach units of measurement to X and Y as per the data set being used!!!!
- Determining whether each coefficient is statistically significant uses the same concept as with the simple regression model
- In our example:
  On average, every additional schooling year increases hourly earnings by \$1.88, controlling for the effects of other X variables
  On average, every additional year of out-of-school experience completed raises hourly earnings by \$0.98, ceteris paribus
- Controlling for the effects of other X variables or Ceteris paribus means: if two individuals, e.g. Yichen and Chiara, have the same years of out of school experience (EXP), then if Chiara completes an additional grade of schooling (S) compared to Yichen, we predict that Chiara will earn a \$1.88 higher hourly rate.

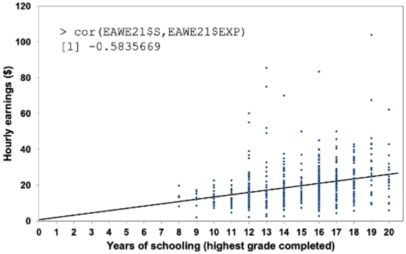
Impact of Omitted Variables

Simple regression model:

$$\widehat{EARNINGS}_i = 0.765 + 1.266 S_i$$

Multiple regression model:

$$\widehat{EARNINGS}_i = -14.668 + 1.877 S_i + 0.983 EXP_i$$



- Schooling is negatively correlated with work experience!
- Regression line underestimates the impact of schooling on earnings.

- Years of schooling is negatively correlated with work experience!
- Simple regression line underestimates the impact of Years of schooling on earnings.