BS2280 - Econometrics 1

Lecture 2 - Part 2: Introduction to Simple Regression Analysis

by Dr Yichen Zhu

Outline

1. Simple Regression Analysis

2. Regression using R

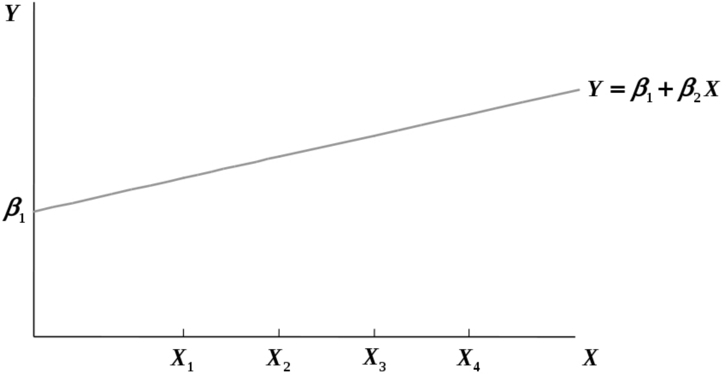
Student task

1. Assume that $\beta_1 = 1$ and $\beta_2 = 2$. Draw the regression line.
2. Show how the regression line will change when $\beta_1 = 0$
3. Show how the regression line will change when $\beta_2 = -1$

Exact relationship

$$Y = \beta_1 + \beta_2 X$$

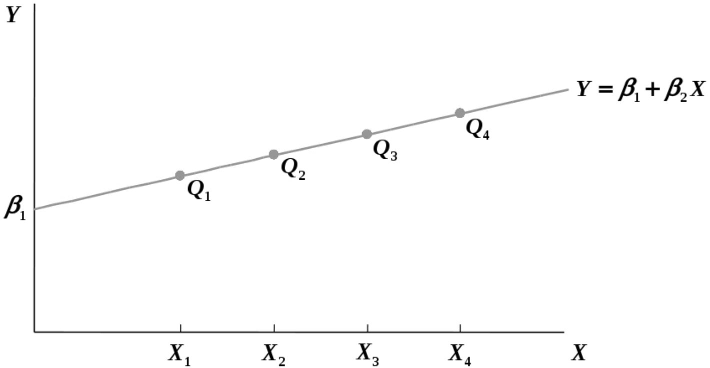
Equation above states an exact relationship between X and Y. If we know the value of X, we know exactly what value of Y will come out.



Exact relationship

$$Y = \beta_1 + \beta_2 X$$

Equation above states an exact relationship between X and Y. If we know the value of X, we know exactly what value of Y will come out.



Exact relationship

Question: Do you think most real-world relationships are exact or perfectly precise?
Take, for instance, the relationship between education and wages.
If two students possess the same number of years of education, can we expect them to earn identical wages?

$$Y = \beta_1 + \beta_2 X$$

Simple Linear Regression Model

- Most relationships are not exact - a value of X can produce different values of Y
- Therefore, we need to add a disturbance term u to the regression equation indicates that the relationship is not exact
- E.g.: People with same education will earn not exactly the same

$$Y = \beta_1 + \beta_2 X + u$$

- This equation is the “simple regression model”
- The term simple comes from the fact that there is only one variable X on the right-hand side

Simple Linear Regression Model

- Most relationships are not exact - a value of X can produce different values of Y
- Therefore, we need to add a disturbance term u to the regression equation indicates that the relationship is not exact
- E.g.: People with same education will earn not exactly the same

$$Y = \beta_1 + \beta_2 X + u$$

- This equation is the “simple regression model”
- The term simple comes from the fact that there is only one variable X on the right-hand side

Simple Linear Regression Model

- Most relationships are not exact - a value of X can produce different values of Y
- Therefore, we need to add a disturbance term u to the regression equation indicates that the relationship is not exact
- E.g.: People with same education will earn not exactly the same

$$Y = \beta_1 + \beta_2 X + u$$

- This equation is the “simple regression model”
- The term simple comes from the fact that there is only one variable X on the right-hand side

Simple Linear Regression Model

- Most relationships are not exact - a value of X can produce different values of Y
- Therefore, we need to add a disturbance term u to the regression equation indicates that the relationship is not exact
- E.g.: People with same education will earn not exactly the same

$$Y = \beta_1 + \beta_2 X + u$$

- This equation is the “simple regression model”
- The term simple comes from the fact that there is only one variable X on the right-hand side

Simple Linear Regression Model

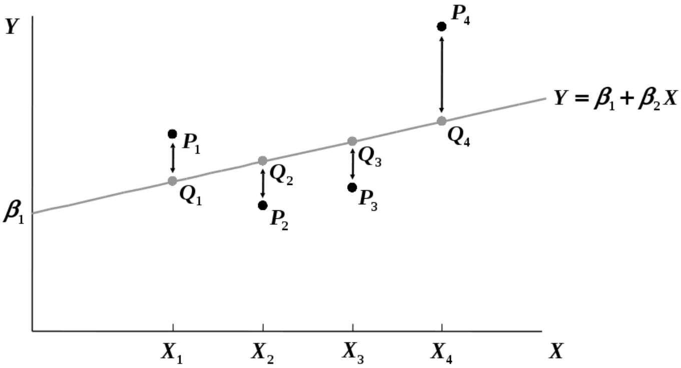
- Most relationships are not exact - a value of X can produce different values of Y
- Therefore, we need to add a disturbance term u to the regression equation indicates that the relationship is not exact
- E.g.: People with same education will earn not exactly the same

$$Y = \beta_1 + \beta_2 X + u$$

- This equation is the “simple regression model”
- The term simple comes from the fact that there is only one variable X on the right-hand side

Simple Linear Regression Model

In the real-world, we should have P points rather than Q points.
E.g. If we assume X is education and Y is wage
P_1: less years of education but earn more
P_3: more years of education but earn less



Simple Linear Regression Model - Cross-sectional data

- We focus on cross-sectional data (many individuals, one time point)
- We therefore add the i subscript

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

- Non-random component: $Y_i = \beta_1 + \beta_2 X_i$
- Random component: u_i

$$\underset{\substack{\text{dependent variable}\\ \text{explained variable}\\ \text{regressand}\\ \text{left-hand side variable}}}{Y_i} = \beta_1 + \beta_2 \underset{\substack{\text{independent variable}\\ \text{explanatory variable}\\ \text{regressor}\\ \text{right-hand side variable}}}{X_i} + \underset{\substack{\text{disturbance term}\\ \text{error term}\\ \text{noise term}}}{u_i} \tag{2}$$

- β_1 and β_2: parameters of the regression model

Simple Linear Regression Model - Cross-sectional data

- We focus on cross-sectional data (many individuals, one time point)
- We therefore add the i subscript

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

- Non-random component: $Y_i = \beta_1 + \beta_2 X_i$
- Random component: u_i

$$\begin{array}{ccccc} Y_i & = \beta_1 + \beta_2 & X_i & + & u_i \\ \text{dependent variable} & & \text{independent variable} & & \text{disturbance term} \\ \text{explained variable} & & \text{explanatory variable} & & \text{error term} \\ \text{regressand} & & \text{regressor} & & \text{noise term} \\ \text{left-hand side variable} & & \text{right-hand side variable} & & \end{array} \tag{2}$$

- β_1 and β_2: parameters of the regression model

Simple Linear Regression Model - Cross-sectional data

- We focus on cross-sectional data (many individuals, one time point)
- We therefore add the i subscript

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

- Non-random component: $Y_i = \beta_1 + \beta_2 X_i$
- Random component: u_i

$$\begin{array}{cccc} Y_i & = \beta_1 + \beta_2 & X_i & + & u_i \\ \text{dependent variable} & & \text{independent variable} & & \text{disturbance term} \\ \text{explained variable} & & \text{explanatory variable} & & \text{error term} \\ \text{regressand} & & \text{regressor} & & \text{noise term} \\ \text{left-hand side variable} & & \text{right-hand side variable} & & \end{array} \tag{2}$$

- β_1 and β_2: parameters of the regression model

Simple Linear Regression Model - Cross-sectional data

- We focus on cross-sectional data (many individuals, one time point)
- We therefore add the i subscript

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

- Non-random component: $Y_i = \beta_1 + \beta_2 X_i$
- Random component: u_i

$$\begin{array}{ccccc}
Y_i & = \beta_1 + \beta_2 & X_i & + & u_i \\
\text{dependent variable} & & \text{independent variable} & & \text{disturbance term} \\
\text{explained variable} & & \text{explanatory variable} & & \text{error term} \\
\text{regressand} & & \text{regressor} & & \text{noise term} \\
\text{left-hand side variable} & & \text{right-hand side variable} & &
\end{array} \tag{2}$$

- β_1 and β_2: parameters of the regression model

Simple Linear Regression Model - Cross-sectional data

- We focus on cross-sectional data (many individuals, one time point)
- We therefore add the i subscript

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

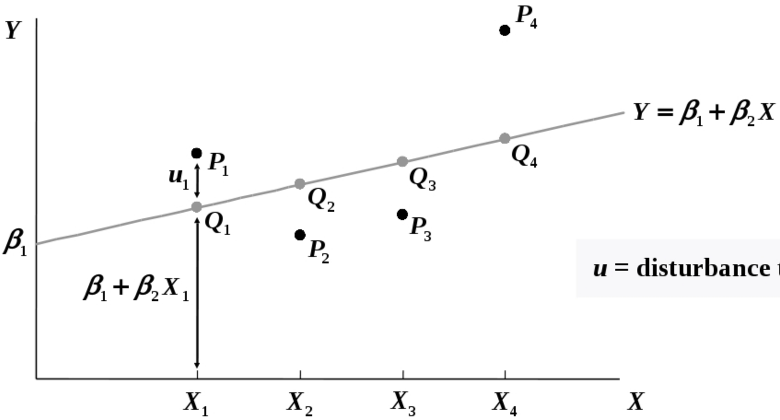
- Non-random component: $Y_i = \beta_1 + \beta_2 X_i$
- Random component: u_i

$$\begin{array}{ccccc} Y_i & = \beta_1 + \beta_2 & X_i & + & u_i \\ \text{dependent variable} & & \text{independent variable} & & \text{disturbance term} \\ \text{explained variable} & & \text{explanatory variable} & & \text{error term} \\ \text{regressand} & & \text{regressor} & & \text{noise term} \\ \text{left-hand side variable} & & \text{right-hand side variable} & & \end{array} \tag{2}$$

- β_1 and β_2: parameters of the regression model

Simple Linear Regression Model - Cross-sectional data

For each individual, we should have one disturbance term.
E.g. P_1 has the disturbance term u_1



u = disturbance term

Why does the disturbance term exist? There are several reasons.

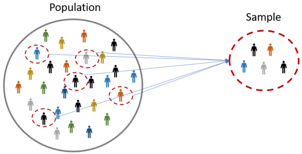
$$Y_i = \beta_1 + \beta_2 X_i + \underset{\substack{\text{Reasons: 1) Omitted variables} \\ \text{2) Aggregation of variables} \\ \text{3) Mis-specification of functional form} \\ \text{4) Measurement error}}}{u_i} \tag{3}$$

- These can arise due to:
  1) Omitted variables - variables we know affect Y but we cannot observe/measure
  2) Aggregation of variables - aggregate consumption
  3) Mis-specification of functional form - use linear instead of non-linear relationship
  4) Measurement error - variables measured with error which translate into error terms

Student task

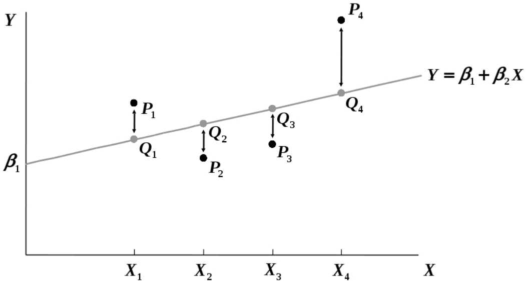
- Assume that Y_i is the mark achieved in the Principles of Microeconomics exam and X_i is the time spent revising for the exam of student i. While there is positive relationship between those two variables ($\beta_2 > 0$) the relationship won't be exact. Explain what factors could be captured by the disturbance term.

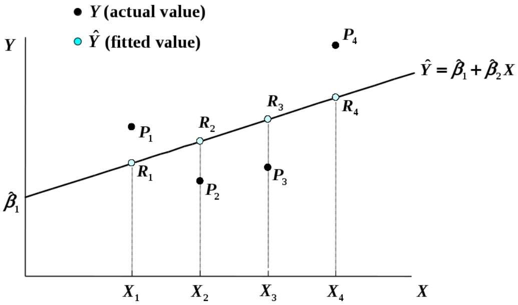
Simple Linear Regression Model - Population vs. Sample



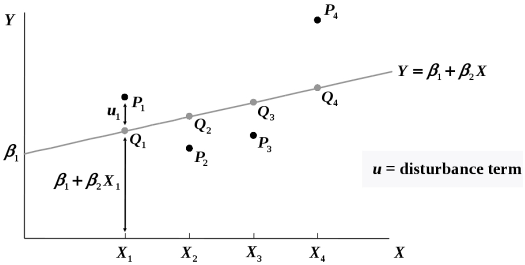
- When we hypothesize a relationship between X and Y, we do so thinking in terms of the population
- But population not observable - time consuming, costly, sometimes downright impossible!
- Use sample data from the population
- In the population, a unique value exists for each of β_1 and β_2
- We do not know what these values are, as we cannot observe the population
- The basis of regression analysis and much of econometrics is to use sample data to estimate these unknown population values

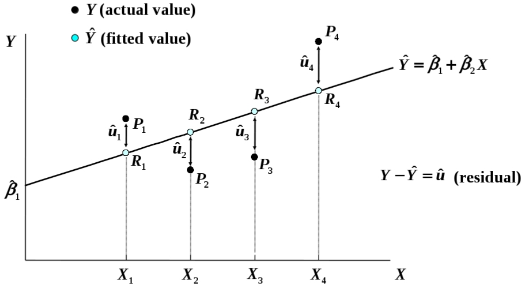
Simple Linear Regression Model - Population vs. Sample




● Y (actual value)
○ $\hat{Y}$ (fitted value)
Y
P₄
$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$
R₃
R₄
R₂
P₁
R₁
P₃
P₂
$\hat{\beta}_1$
X₁
X₂
X₃
X₄
X

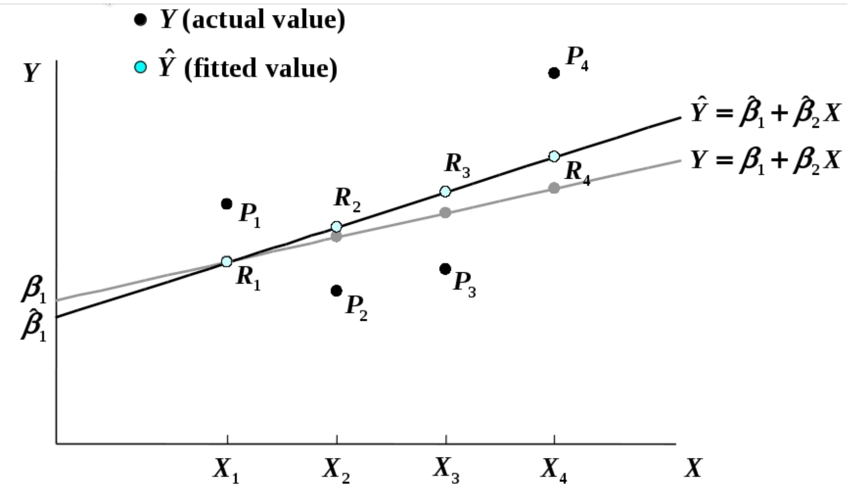
Simple Linear Regression Model - Population vs. Sample





Population	Sample
$Y_i = \beta_1 + \beta_2 X_i + u_i$ parameters β_1 and β_2 u disturbance term	$Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$ coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ $\hat{u}$ residual

Simple Linear Regression Model - Population vs. Sample



Simple Linear Regression Model

- Two aspects to the estimation part:
  - quantification
  - statistical inference
- Quantification is a technique to calculate estimated values on β_1 and β_2
- Estimated values (coefficients): $\hat{\beta}_1$ and $\hat{\beta}_2$ (note the ˆ symbol)
- Statistical inference: are $\hat{\beta}_1$ and $\hat{\beta}_2$ statistically significant?
- Today, we focus on quantification.

Simple Linear Regression Model

- Two aspects to the estimation part:
  - quantification
  - statistical inference
- Quantification is a technique to calculate estimated values on β_1 and β_2
- Estimated values (coefficients): $\hat{\beta}_1$ and $\hat{\beta}_2$ (note the ^symbol)
- Statistical inference: are $\hat{\beta}_1$ and $\hat{\beta}_2$ statistically significant?
- Today, we focus on quantification.

Simple Linear Regression Model

- Two aspects to the estimation part:
  - quantification
  - statistical inference
- Quantification is a technique to calculate estimated values on β_1 and β_2
- Estimated values (coefficients): $\hat{\beta}_1$ and $\hat{\beta}_2$ (note the ^ symbol)
- Statistical inference: are $\hat{\beta}_1$ and $\hat{\beta}_2$ statistically significant?
- Today, we focus on quantification.

Simple Linear Regression Model

- Two aspects to the estimation part:
  - quantification
  - statistical inference
- Quantification is a technique to calculate estimated values on β_1 and β_2
- Estimated values (coefficients): $\hat{\beta}_1$ and $\hat{\beta}_2$ (note the ^ symbol)
- Statistical inference: are $\hat{\beta}_1$ and $\hat{\beta}_2$ statistically significant?
- Today, we focus on quantification.

Simple Linear Regression Model

- Two aspects to the estimation part:
  - quantification
  - statistical inference
- Quantification is a technique to calculate estimated values on β_1 and β_2
- Estimated values (coefficients): $\hat{\beta}_1$ and $\hat{\beta}_2$ (note the ˆ symbol)
- Statistical inference: are $\hat{\beta}_1$ and $\hat{\beta}_2$ statistically significant?
- Today, we focus on quantification.

Simple Linear Regression Model

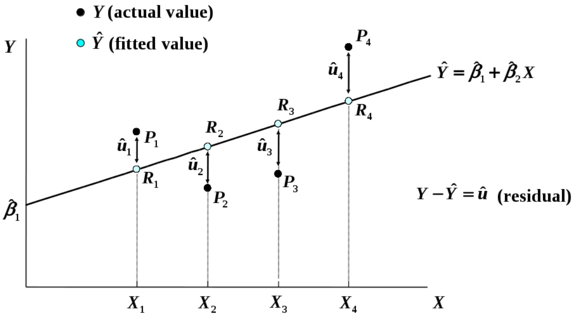
- Two aspects to the estimation part:
  - quantification
  - statistical inference
- Quantification is a technique to calculate estimated values on β_1 and β_2
- Estimated values (coefficients): $\hat{\beta}_1$ and $\hat{\beta}_2$ (note the ˆ symbol)
- Statistical inference: are $\hat{\beta}_1$ and $\hat{\beta}_2$ statistically significant?
- Today, we focus on quantification.

Simple Linear Regression Model

- Two aspects to the estimation part:
  - quantification
  - statistical inference
- Quantification is a technique to calculate estimated values on β_1 and β_2
- Estimated values (coefficients): $\hat{\beta}_1$ and $\hat{\beta}_2$ (note the ˆ symbol)
- Statistical inference: are $\hat{\beta}_1$ and $\hat{\beta}_2$ statistically significant?
- Today, we focus on quantification.

Ordinary Least Squares (OLS)

- Intuitively, OLS is fitting a line through the sample points such that the sum of squared residuals (RSS) is as small as possible, hence the term least squares
- The residual, $\hat{u}$, is an estimate of the error term, u, and is the difference between the fitted line (sample regression function) and the sample point
- OLS aims to find the values of the linear regression model's coefficients that **minimise the sum of the squared residuals (RSS).**



Ordinary Least Squares (OLS)

Given the intuitive idea of fitting a line, we can set up a formal minimisation problem

- **Step 1.** Write out the sum of squared residuals (RSS) formula

$$Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

so,

$$\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

$$RSS = \sum_{i=1}^{n} \hat{u}_i{}^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

Expand (Derivations will be thoroughly explained in the post-lecture videos):

$$RSS = \sum_{i=1}^{n} Y_i^2 + n\hat{\beta}_1^2 + \hat{\beta}_2^2 \sum X_i^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} Y_i - 2\hat{\beta}_2 \sum_{i=1}^{n} X_i Y_i + 2\hat{\beta}_1\hat{\beta}_2 \sum_{i=1}^{n} X_i$$

- **Step 2.** We want to choose our coefficients such that we minimise the expression of RSS taking First Order Conditions:

$$RSS = \sum_{i=1}^{n} Y_i^2 + n\hat{\beta}_1^2 + \hat{\beta}_2^2 \sum_{i=1}^{n} X_i^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} Y_i - 2\hat{\beta}_2 \sum_{i=1}^{n} X_i Y_i + 2\hat{\beta}_1\hat{\beta}_2 \sum_{i=1}^{n} X_i$$

First Order Conditions (Derivations will be thoroughly explained in the post-lecture videos):

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = 2n\hat{\beta}_1 - 2\sum_{i=1}^{n} Y_i + 2\hat{\beta}_2 \sum_{i=1}^{n} X_i = 0$$

$$\frac{\partial RSS}{\partial \hat{\beta}_2} = 2\hat{\beta}_2 \sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i Y_i + 2\hat{\beta}_1 \sum_{i=1}^{n} X_i = 0$$

The necessary condition for a relative extremum (maximum or minimum) is that the first-order derivative be zero.

Ordinary Least Squares (OLS)

- **Step 3.** Solving for FOC's (the normal equations):

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$$

$$\begin{aligned}
\hat{\beta}_2 &= \frac{(\sum_{i=1}^{n} XY) - n\bar{X}\bar{Y}}{(\sum_{i=1}^{n} X^2) - n\bar{X}^2} \\
&= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \frac{Cov(X, Y)}{Var(X)}
\end{aligned}$$

Summary of OLS slope estimate

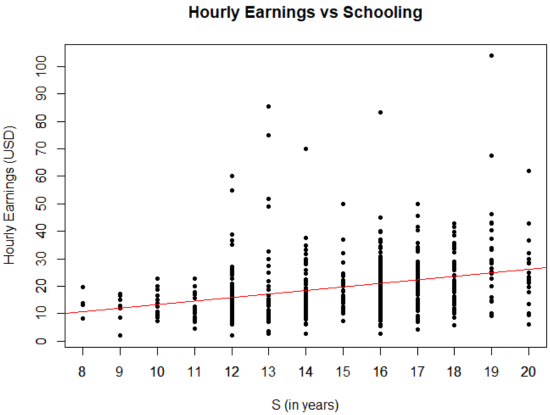
- The slope estimate is the sample covariance between X and Y divided by the sample variance of X
- If X and Y are positively correlated, the slope will be positive
- If X and Y are negatively correlated, the slope will be negative
- Only need X to vary in our sample

Example: Earnings and education

- Let's consider a concrete example of a simple regression model on a real life data set using R.
- We use the National Longitudinal Survey of Youth 1997 (NLSY97) to model the relationship between hourly earnings and years of schooling
- We use two variables from this data set:
  - EARNINGS - hourly earnings ($)
  - S - years of schooling
  - We aim to estimate the following regression model:

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

Example: Earnings and education



- Red line is the linear regression line derived through OLS

Example: Earnings and education

```
> # R command to run regression EARNINGS on S
> earnfit <- lm(EARNINGS~S, data=EAWE21.simple)

> # print regression output
> earnfit

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

Coefficients:
(Intercept)            S
     0.7647       1.2657
```

- The output above shows the estimated $\hat{\beta}_1$ and $\hat{\beta}_2$

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

What to do next:

- Attempt homework 2
- Read chapter 1.1 - 1.5 of Dougherty

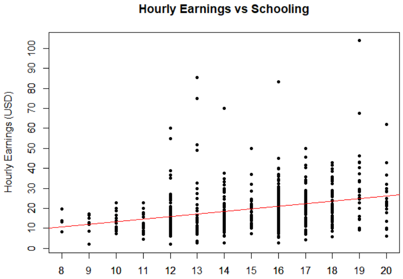
APPENDIX

- R code to generate scatter plot of EARNINGS and S used in this lecture:

```
plot(EAWE21.simple$EARNINGS~EAWE21.simple$S,
        main="Hourly Earnings vs Schooling",
    xlab ="S (in years)",
    ylab = "Hourly Earnings (USD)", pch=20)
axis(side=1, at=c(8:20))
axis(side=2, at=seq(0, 100, by= 10))
```

- Add regression line:

```
abline(lm(EARNINGS~S,data=EAWE21.simple), col="red" )
```


Hourly Earnings vs Schooling
Hourly Earnings (USD)