Introduction to Statistical Methods
SOC-GA 2332

Lecture 4: Regression I

Siwei Cheng

NEW YORK UNIVERSITY

Why Regression?

Three perspectives:

Why Regression?

Three perspectives:

- Perspective 1: Regression as describing a relationship
- Perspective 2: Regression as prediction (for conditional mean)
- Perspective 3: Regression as a causal relationship (or not!)

P1: Regression as describing a relationship

- We can write Y as a function of X.
- X can be multi-dimensional (i.e. including more than one variables). For example, we can write Y as a linear function of X_1, X_2, and X_3:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \tag{1}$$

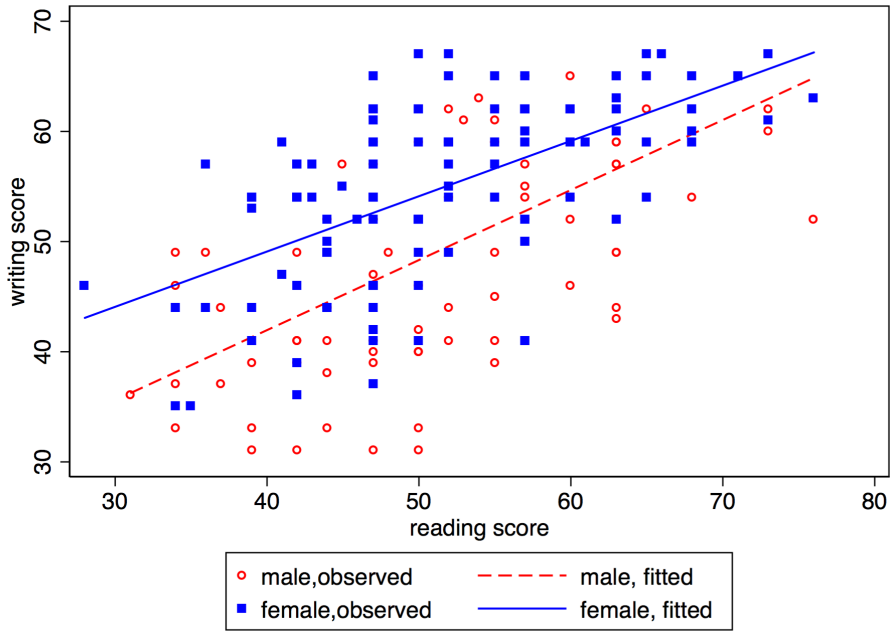
- For example, let Y=attitudes towards immigration policies. We may hypothesize that Y is a function of: X_1=perceived economic consequence of immigrant population, X_2=frequency of contact with immigrants, X_3=own educational attainment...

P1: Regression as describing a relationship

- In the case of binary regression, there is only one variable in X:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2}$$

- Note that even in the case of a binary regression, we are NOT necessarily making the assumption that Y is determined by only one variable X_1. Why?



P2: Regression as prediction (for conditional mean)

P2: Regression as prediction (for conditional mean)

- Regression can also be seen as giving us the conditional mean of Y at a given value of X. This conditional mean is denoted by $E(Y|X)$.
- In Equation (2), suppose $E(\epsilon|X) = 0$ for all values of X, then we have:

$$E(Y|X) = \beta_0 + \beta_1 X. \tag{3}$$

- For example, when $X = x_0$, $E(Y|x_0) = \beta_0 + \beta_1 \cdot x_0$.
- In other words, the linear regression equation (and estimated coefficients, such as β_0 and β_1), gives us the formula to predict the expectation of Y given a value of X.

P2: Regression as prediction (for conditional mean)

- Let's look at a visual illustration from Angrist & Pischke Page 31:

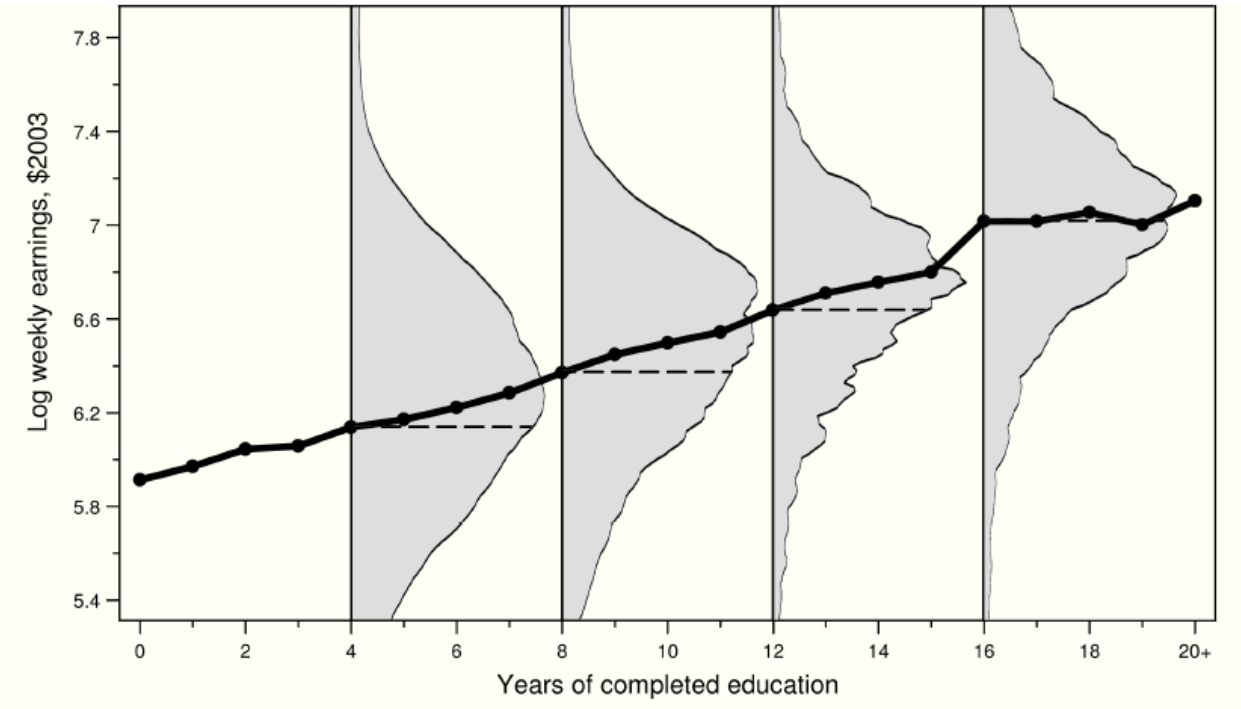


Figure 3.1.1: Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49 in the 1980 IPUMS 5 percent file.

P3: Regression as a causal relationship (or not!)

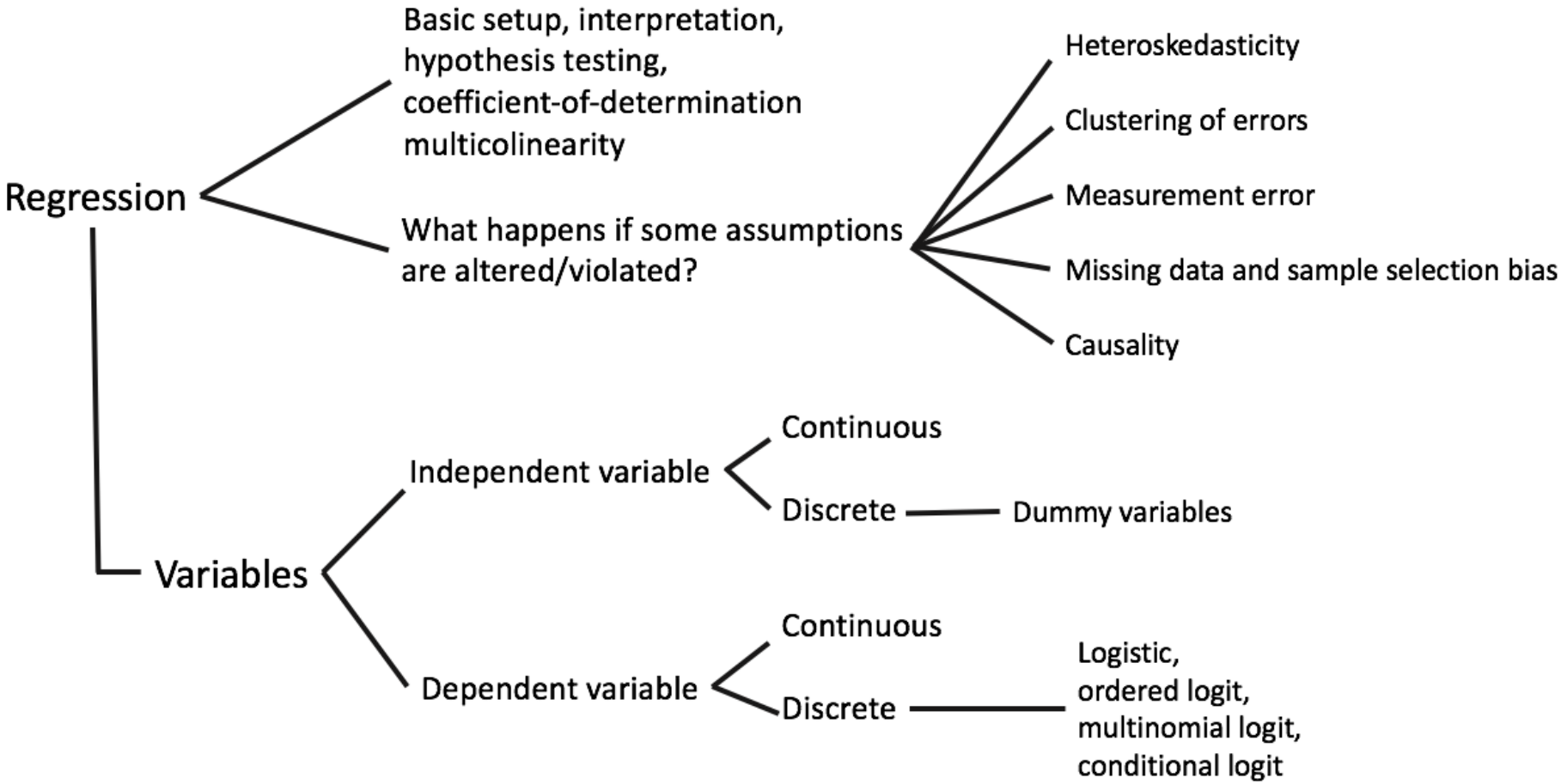
P3: Regression as a causal relationship (or not!)

- Does a coefficient estimated from a regression model imply a causal relationship? (e.g. if $\beta_1 \neq 0$, does this mean "X causes Y"?)

P3: Regression as a causal relationship (or not!)

- Does a coefficient estimated from a regression model imply a causal relationship? (e.g. if $\beta_1 \neq 0$, does this mean "X causes Y"?)
- NOT necessarily!
- **Under some assumptions**, regression coefficients can be interpreted as "causal" effects, but we should be very careful as to whether those assumptions are met.
- We will devote a separate lecture on **causality**.

Regression: A Roadmap



The Linear Regression: Setting up the Model

- Let's start with the case of bivariate linear regression (describing the relationship between two variables). The regression equation relates y (dependent variable) to x (independent variable):

$$Y = \underbrace{\beta_0 + \beta_1 X}_{deterministic} + \underbrace{\epsilon}_{stochastic} \tag{4}$$

- **Note: upper case letters** are often used to denote random variables; **lower case letters** are often used to denote values on individual observations. For example, we can also use lower case letters to denote one observation i to express the relationship as described in equation (4):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{5}$$

- β_0 is the intercept, β_1 is the slope, and ϵ is the error term.

The Linear Regression: Setting up the Model

- If we take the expectation on both sides of Equation (4).

$$E(Y|X) = \beta_0 + \beta_1 X + E(\epsilon|X) \tag{6}$$

$$\textit{If} \quad E(\epsilon|X) = 0 \quad \textit{then} \quad E(Y|X) = \beta_0 + \beta_1 X \tag{7}$$

- Then we have **predicated value** of Y:

$$\hat{Y} = E(Y|X) = \beta_0 + \beta_1 X \tag{8}$$

Interpreting Model Parameters

- Prediction equation:

$$\hat{Y} = E(Y|X) = \beta_0 + \beta_1 X$$

- The **intercept parameter** β_0 is the predicated value when X=0:

$$\hat{Y}_{X=0} = E(Y|X) = \beta_0 + \beta_1 \cdot 0 = \beta_0 \quad (9)$$

- The **slope parameter** β_1 tells us how much $\hat{Y}$ changes in response to changes in X, that is:

$$\Delta\hat{Y} = \beta_1 \Delta X \quad (10)$$

- The **error term** ϵ. captures the deviation of observed Y from predicted Y:

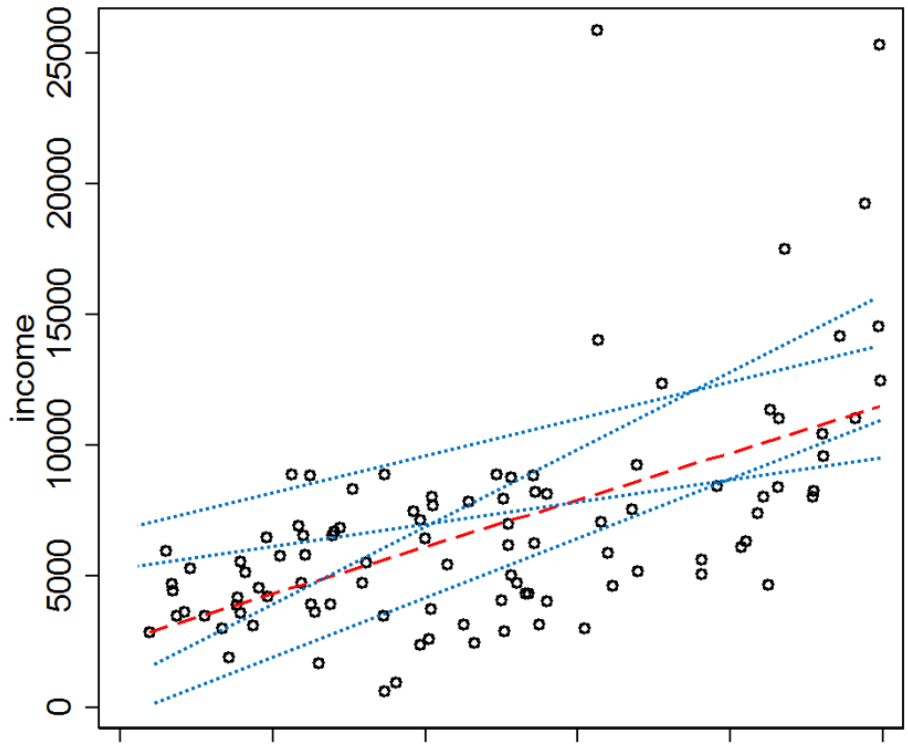
$$\epsilon = Y - \hat{Y} \quad (11)$$

Estimating Model Parameters

- The prediction equation is a **line** that describes the relationship between X and Y:

$$\hat{Y} = E(Y|X) = \beta_0 + \beta_1 X \tag{12}$$

- Once we have the estimated parameters, we will be able to draw the prediction line:



Estimating Model Parameters

- Now, going back to the linear regression prediction equation:

$$\hat{Y} = E(Y|X) = \beta_0 + \beta_1 X \tag{13}$$

- Note that there are two unknown parameters in this prediction equation: intercept β_0 and slope β_1.
- When we do regression analysis, we are looking for a "prediction equation" that helps us determine the relationship between X and Y in the **population** based on our **sample** data.

Estimating Model Parameters

- Now, our question is: how to find "good" estimators for β_0 and β_1?
- The statistical methods that we commonly use to estimate these two parameters are called **Ordinary Least Squares (OLS)** estimation. The two parameters estimated under this method are called **OLS estimators**.
- We use $\hat{\beta}_0$ and $\hat{\beta}_1$ to denote our estimator of population parameters β_0 and β_1. The predicted value of Y is thus given by: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.
- For each observation i, the predicted value for y_i is: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

The OLS Estimator

- The OLS estimator starts with model **residuals**.
- Residual for observation i is given by the difference between observed value y_i and estimated value $\hat{y}_i$. We know that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, so the residual for observation i is given by:

$$\epsilon_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \tag{14}$$

The OLS Estimator

- The OLS estimator starts with model **residuals**.
- Residual for observation i is given by the difference between observed value y_i and estimated value $\hat{y}_i$. We know that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, so the residual for observation i is given by:

$$\epsilon_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \tag{14}$$

- The OLS estimators minimize the **sum of squared errors/residuals (SSE)**:

$$\text{SSE} = \Sigma_{i=1}^{n} \epsilon_i^2 = \Sigma[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \tag{15}$$

- Now, the problem becomes **"finding the proper expression for $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the SSE in Equation (16) is minimized.** Basically, this because a minimization problem.

Illustration of Residuals and Sum of Squared Errors

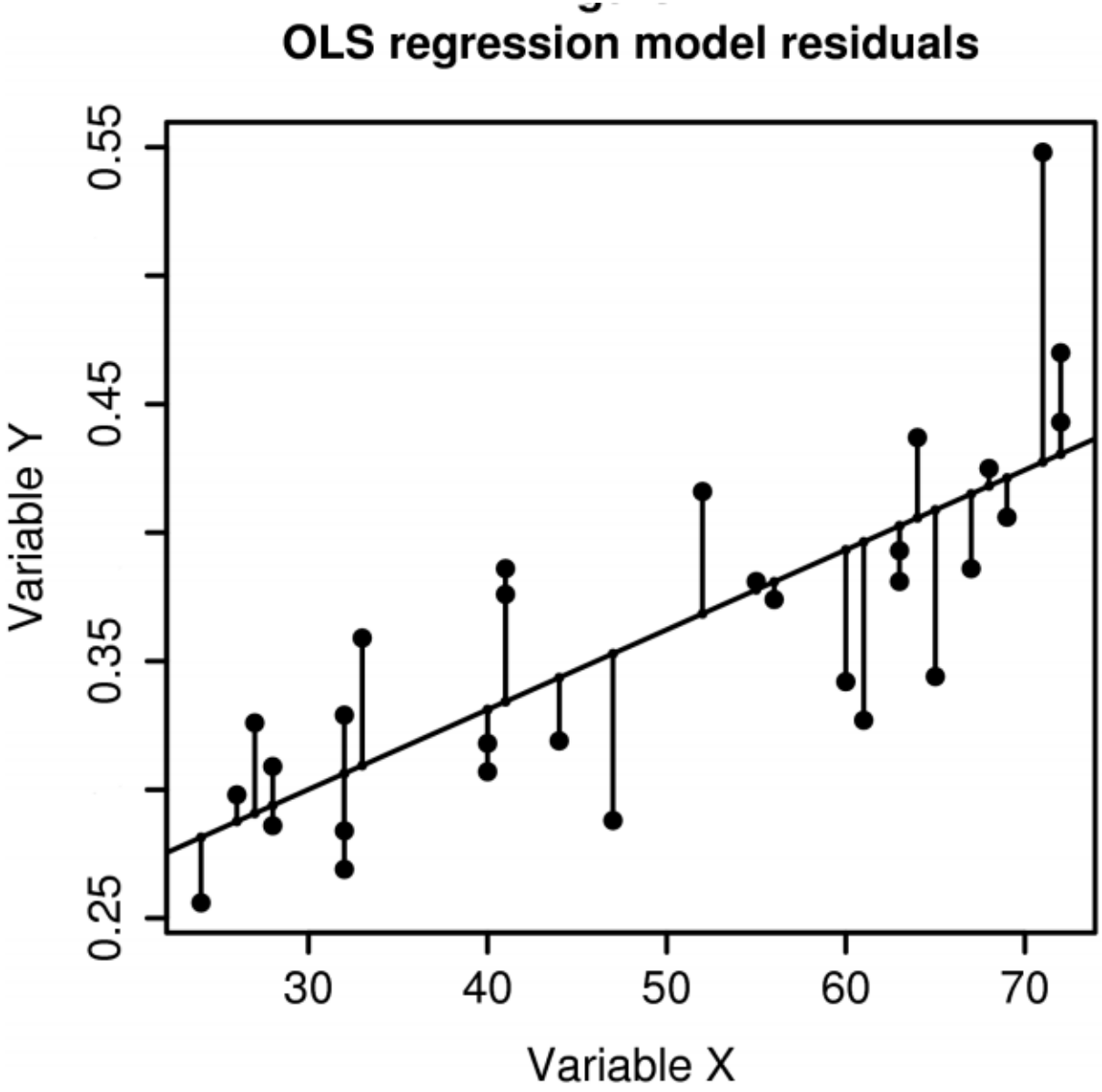
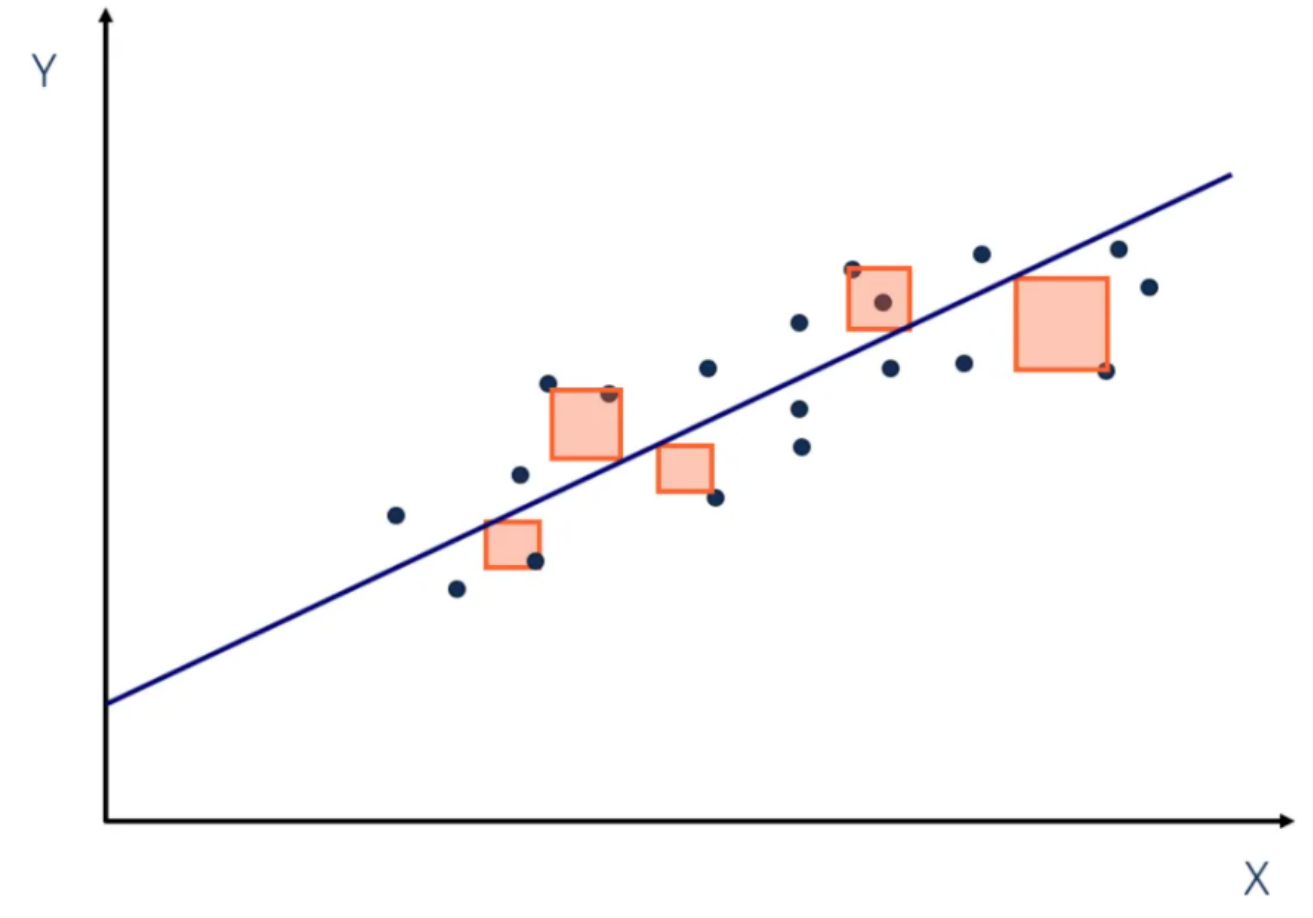


Illustration of Residuals and Sum of Squared Errors



The OLS Estimator

- The OLS estimator minimizes the **sum of squared errors/residuals (SSE)**:

$$\text{SSE} = \Sigma_{i=1}^{n} \epsilon_i^2 = \Sigma[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \tag{16}$$

- Now, the problem becomes "finding the proper expression for $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the SSE in Equation (16) is minimized. Basically, this because a **minimization problem**.

Estimating Model Parameters

- Let's try to solve this minimization problem:

  Let's use Z to denote SSE.

$$Z = \Sigma[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \tag{17}$$

So Z is a function of β_0 and β_1. As we learned in calculus, the value of Z is minimized **when the derivatives with respect to β_0 and β_1 are set equal to zero.**

$$\frac{\partial Z}{\partial \beta_0} = -2\Sigma(y_i - \beta_0 - \beta_1 x_i) = 0 \tag{18}$$

$$\frac{\partial Z}{\partial \beta_1} = -2\Sigma x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \tag{19}$$

Estimating Model Parameters

- Now we are going to rearrange the two expressions above to solve for β_0 and β_1.
- Rearranging Equation (18) gives us:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{20}$$

Then we plug in the above expression for β_0 in Equation (20). After rearranging Equation (19) we have:

$$\hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \tag{21}$$

Estimating Model Parameters

- The above derivation shows that the $\hat{\beta}_0$ and $\hat{\beta}_1$ that solves the minimization problem are expressed as the following:

$$\hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{Cov(X, Y)}{Var(X)} \tag{22}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{23}$$

- Note also that the unbiased estimator for σ^2 is: $\hat{\sigma}^2 = MSE = \frac{\Sigma(y - \hat{y})^2}{n-2}$

Assumptions and Properties of the OLS Estimator

- It is important that you keep in mind that two important assumptions are required if we want the OLS estimators as shown in Equations (22) and (23) to be unbiased and consistent:
  1. The error term ϵ has **zero mean**: $E(\epsilon) = 0$.
  2. The error term is **mean independent of x**: $E(\epsilon|x) = E(\epsilon)$.
- Combining the two assumptions above we have:

$$E(\epsilon|x) = E(\epsilon) = 0$$

- What do these two assumptions mean?
- Can you think of empirical cases where ϵ is NOT mean independent of x?

Assumptions and Properties of the OLS Estimator

- If it's assumptions are met, the OLS estimator satisfies some desirable properties:
  1. **Unbiased**: The expected values of $\hat{\beta}_0$ and $\hat{\beta}_1$ coincide with the true parameter: $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$.
  2. **Consistent**: $\hat{\beta}_0$ and $\hat{\beta}_1$ converges in probability to the population parameters β_0 and β_1 when the sample size n goes to infinity.

OLS is BLUE!

(if Gauss-Markov Assumptions are Met)

- Gauss-Markov Assumptions:
  1. Linear in parameters.
  2. Random sampling.
  3. No perfect collinearity.
  4. Zero conditional mean: $E(\epsilon|x_1, x_2, ..., x_k) = 0$.
  5. Homoskedasticity: $Var(\epsilon|x_1, x_2, ..., x_k) = \sigma^2$.
- When the above assumptions are met, the OLS estimator is the Best Linear Unbiased Estimator (**BLUE**). That is, it has lower sampling variance than any other unbiased estimator in the class of linear estimators for all possible values of the parameter.
- So we need not look for alternative unbiased estimators if all the above assumptions are met: none will be better than OLS.

Assumptions and Properties of the OLS Estimator

- However, as we will discuss later this course, we should always interpret the OLS estimator with some caution, because the assumptions are not always satisfied.
- Next week's lecture will address some of the scenarios where these assumptions do not hold.

Relationship between β_1 and Correlation ρ

- What's the relationship between β_1 and correlation ρ? Recall that:

$$\hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{Cov(X, Y)}{Var(X)} = \frac{Cov(X, Y)}{\sigma_X \sigma_X} \tag{24}$$

- And we know that the correlation between x and y is expressed as:

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{25}$$

- So we have:

$$\hat{\beta}_1 = \rho_{X,Y} \cdot \frac{\sigma_Y}{\sigma_X} \tag{26}$$

Assumptions and Properties of the OLS Estimator

$$\hat{\beta}_1 = \rho_{X,Y} \cdot \frac{\sigma_Y}{\sigma_X} \tag{27}$$

- Difference between β_1 and ρ:
- The slope coefficient tells us the **direction** of the association, but the correlation ρ tells us the **strength** of association, as it is the covariance between X and Y "standardized" by dividing it by the standard deviations of the two variables.

Assumptions and Properties of the OLS Estimator

- Why is the slope coefficient **NOT** a measure for the strength of association?
- Note that β_1 is sensitive to the **scale** at which the variables are measured.
- For example, if X=Average income, Y=infant mortality rate (country-level). Suppose $Y = a + bX$ when X is measured in dollars, then $Y = a + (1000b) \cdot X$ if X is measured in thousands of dollars.
- If convert the two variables to their z-scores: $z_x = \frac{x-\bar{x}}{s_x}$ and $z_y = \frac{y-\bar{y}}{s_y}$, then we know $\sigma_{z_x} = \sigma_{z_y} = 1$, so β_1 and ρ will be identical.

Other Methods for Finding Estimators

- Remember that we found the OLS estimators by solving the minimization problem (i.e. minimizing squared residual variance).
- There are also other methods for finding estimators of population parameters. For example, there is a method called the maximum likelihood estimator (**MLE**), which generates estimators by looking for the parameter values that maximize the likelihood of getting the observations in your sample given the parameters.
- The MLE estimator is one of the most popular technique for deriving estimators.

The MLE estimator

- Recall that if $x_1, x_2, ., x_n$ are an iid sample from a opulation with pdf or pmf $f(x|\theta_1, ..., \theta_k)$, the likelihood function is defined by:

  $$L(\theta|x) = I(\theta_1, ..., \theta_k|x_1, x_2, ., x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, ..., \theta_k).$$

- If the likelihood function is differentiable (in θ_i), possible candidates for the MLE are the values of $f(x|\theta_1, ..., \theta_k)$ that solve $\frac{\partial L(\theta|x)}{\partial \theta_i} = 0$.
- Note that the solutions to the first-order condition above are only possible candidates for the MLE. It may give a local, but not global, maximum (unless we can show that it is the only zero of the FOC).
- Also, the zeros of the first derivative locate only extreme points in the interior of the domain of a function. The **boundary** must be checked separately for extrema.

Example: Bernoulli MLE

- Keep in mind that the likelihood function is defined by:

  $L(\theta|x) = I(\theta_1, ..., \theta_k|x_1, x_2, ., x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, ..., \theta_k)$.

- Let $X_1, X_2, ..., X_n$ be iid Bernoulli(p) sample. Then the likelihood function is written as:

  $L(\theta|x) = I(\theta_1, ..., \theta_k|x_1, x_2, ., x_n) = \prod_{i=1}^{n} p_i^x (1-p)^{1-x_i} = p^y (1-p)^{n-y}$,

  where $y = \Sigma X_i$.

- It's much easier to differentiate the log likelihood:

  $logL(\theta|x) = ylogp + (n-y)log(1-p)$.

- Differentiating the above log likelihood and setting it to zero give the solution: $\hat{p} = y/n$. This is the MLE of p.

Example: Normal MLE

- Let $X_1, X_2, ..., X_n$ be iid Normal sample, with both μ and σ^2 unknown. Then it can be shown that the MLE estimators for these two parameters are:

  $\hat{\mu}_{MLE} = \bar{x}$

  $\hat{\sigma^2}_{MLE} = \frac{\Sigma_{n=1}^{n}(x_i - \bar{x})^2}{n}$.

- Note that the MLE estimator for σ^2 is a *biased* estimator.

Example: Linear Regression MLE

- Let the data be n iid observations with one explanatory variable (X) and one outcome variable (Y). Suppose the model is that:

  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

- The likelihood function is:

  $L(\beta_0, \beta_1, \sigma^2 | y, x) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} exp(-\frac{1}{2\sigma^2}\Sigma_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2)$.

- It can be shown that when we assume that the error term is normally distributed, than the OLS estimators for β_0 and β_1 and the MLE estimators will be identical.

- The MLE estimator for σ^2 is the in-sample mean squared error. It is a biased estimator of σ^2.

Going back to OLS: An Example

- Let's look an example. Suppose we have two variables: Average number of years of education in an occupation (X=educat) and Average income in an occupation (Y=income). The dataset looks like this:

	educat	income			
1	13.11	12351			
2	12.26	25879			
3	12.77	9271			
4	11.42	8865			
5	14.62	8403			
6	15.64	11030			
7	15.09	8258			
8	15.44	14163			
9	14.52	11377			
10	14.64	11023			
11	12.39	5902			
12	12.3	7059			
13	13.83	8425			
14	14.44	8049			

Going back to OLS: An Example

- We want to model occupational income as a linear function of occupational education level:

$$income = \beta_0 + \beta_1 educat + \epsilon$$

.

- Our predicted income would be:

$$\hat{income} = \hat{\beta}_0 + \hat{\beta}_1 educat$$

.

- In other words, we regress *income* on *educat*.
- How can we find the estimates for β_0 and β_1?

Going back to OLS: An Example

- In STATA or R, we can generate the sample standard deviations for the two variables and the covariance of the two variables by the following commands:

```
sum income educat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
income	102	6797.902	4245.922	611	25879
educat	102	10.73804	2.728444	6.38	15.97

```
. corr income educat, covariance
(obs=102)
```

	income	educat
income	1.8e+07	
educat	6691.13	7.44441

- From this, we know that $\bar{X} = 10.73804$, $\bar{Y} = 6797.902$, $Var(X) = 2.728444^2 = 7.444$, $Var(Y) = 4245.922^2 = 18027854$, and $Cov(X, Y) = 6691.13$.

Going back to OLS: An Example

- So we can calculate $\hat{\beta}_1$ and $\hat{\beta}_0$ based on Equation (22) and (23):

$$\hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{Cov(X, Y)}{Var(X)} \tag{28}$$

$$= \frac{6691.13}{7.444} = 898.862 \tag{29}$$

mean predicted value

slope

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{30}$$

$$= 6797.902 - 898.862 \cdot 10.73804 = -2854.114 \tag{31}$$

Going back to OLS: An Example

- We can also calculate β_1 based on **correlation** ρ and Equation (27). We first obtain correlation between the two variables:

```
. corr income educat
(obs=102)
```

	income	educat
income	1.0000	
educat	0.5776	1.0000

- Then we calculate:

$$\hat{\beta}_1 = \rho_{X,Y} \cdot \frac{\sigma_Y}{\sigma_X} \tag{32}$$

$$= 0.5776 \cdot \frac{4245.922}{2.728444} = 898.844 \tag{33}$$

Going back to OLS: An Example

- Finally, we can run the regression directly:

```
. regress income educat
```

Source	SS	df	MS	Number of obs	=	102
				F(1, 100)	=	50.06
Model	607421404	1	607421404	Prob > F	=	0.0000
Residual	1.2134e+09	100	12133920.1	R-squared	=	0.3336
				Adj R-squared	=	0.3269
Total	1.8208e+09	101	18027855.6	Root MSE	=	3483.4

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educat	898.8129	127.0354	7.08	0.000	646.7782	1150.847
_cons	-2853.586	1407.039	-2.03	0.045	-5645.112	-62.05996

The Linear Regression Model: Hypothesis Testing

- Now, let's move from estimation to hypothesis testing.
- Our research questions are often about whether there is an association between the variation in Y and the variation in X.
- In the linear regression model, we test whether the coefficient on X, β_1, is significantly different from zero.
- Formally, our null and alternative hypotheses are:
  H_0: $\beta_1 = 0$ (Slope equals zero).
  H_1: $\beta_1 \neq 0$ (Slope differs from zero).

The Linear Regression Model: Hypothesis Testing

- Let's go back to $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Keep in mind that the two OLS estimates are calculated based on **sample** data.
- That is, we introduced randomness when we draw the sample from the population, because if our samples included different elements, our calculated $\hat{\beta}_0$ and $\hat{\beta}_1$ are likely to be different. Hence, we would like to know the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$.
- (More generally, the population does not have to exist. Randomness is always introduced as long as there is a (random) data generating process behind the data.)
- We know from an earlier point about the property of OLS estimator that the means of the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ are their corresponding true parameters (i.e. unbiasedness).

The Linear Regression Model: Hypothesis Testing

- What is the standard deviation of the sampling distribution? That is, what is the **standard error** of each parameter estimates?
- To calculate the standard error, we start with the **mean square error (MSE)**:

$$MSE = \frac{SSE}{n-2} = \frac{\Sigma(y-\hat{y})^2}{n-2} \tag{34}$$

- And we take the square root:

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}} \tag{35}$$

The Linear Regression Model: Hypothesis Testing

- The standard error of β_1 is a function of MSE. Specifically, it is:

$$se_{\beta_1} = \frac{s}{\sqrt{\Sigma(x - \bar{x})^2}} \tag{36}$$

- Note that $\Sigma(x - \bar{x})^2 = (n-1)s_x^2$.
- So the denominator increases (i.e. se_{β_1} decreases) when the **sample variance** of x (s_x^2) increases or when the sample size (n) increases.

The Linear Regression Model: Hypothesis Testing

- Now let's go back to hypothesis testing. The test statistic is:

$$t = \frac{\hat{\beta}_1 - 0}{se_{\beta_1}} \tag{37}$$

- When H_0 is true, t follows t-distribution with degree of freedom n-2 (we are using n observations to estimate 2 parameters). We compare the calculated t (based on our sample and the above formulas) with the t-distribution to obtain p-value.
- If p-value is smaller than our pre-determined α, we reject the null hypothesis, or we say that the estimated slope is significantly different from zero.

The Linear Regression Model: Hypothesis Testing

- We can also construct the confidence interval of β_1 (just as we did for the CI of mean) as: $\hat{\beta}_1 \pm t \cdot (se_{\beta_1})$.
- If this confidence interval does not contain zero, then we reject the null hypothesis. Otherwise we fail to reject the null hypothesis.

Example

- Let's look at the previous example of the relationship between occupational education level and occupational income.
- We have our estimate for β_1: $\hat{\beta}_1 = 898.862$. So how to construct a 95% confidence interval of β_1?
- First, from Equation (34) we know that we need SSE, which is $\Sigma(y - \hat{y})^2$. We let STATA calculate MSE by the following codes:

```
. gen income_predicted=-2854.114+898.862*educat

. gen residual=income-income_predicted

. gen residual_sq=residual^2

. egen SSE=total(residual_sq)

. gen MSE=SSE/(102-2)

. di MSE
12133920
```

Example

- So we have se_{β_1} as follows:

$$se_{\beta_1} = \frac{s}{\sqrt{\Sigma(x-\bar{x})^2}} = \frac{\sqrt{MSE}}{\sqrt{n-1}\cdot s_x} \tag{38}$$

$$= \frac{\sqrt{12133920}}{\sqrt{102-2}\cdot 2.728444} = 127.66902 \tag{39}$$

- The 95% confidence interval for β_1 would be (approximating t distribution by normal distribution):

$$898.862 \pm 1.96 \cdot 127.66902 = [648.6, 1149.1]$$

- The 95% confidence interval does not contain 0 (The P-value is smaller than 0.001 (significant on the 0.01 level)), so we reject the null hypothesis that the slope is zero.

Coefficient of Determination: R-squared

Coefficient of Determination: R-squared

- We know that regression describes a relationship between X and Y. We are trying to explain variations in Y by variations in X. So how much of the variations in Y can be explained by X?
- If we do not have information on X, then the total amount of variation in Y in the sample is called Total Sum of Squares (TSS):

$$TSS = \Sigma(y_i - \bar{y})^2$$

- That is, the grand mean is our best guess (without any additional information)!

Coefficient of Determination: R-squared

- We know that regression describes a relationship between X and Y. We are trying to explain variations in Y by variations in X. So how much of the variations in Y can be explained by X?
- If we do not have information on X, then the total amount of variation in Y in the sample is called Total Sum of Squares (TSS):

$$TSS = \Sigma(y_i - \bar{y})^2$$

- That is, the grand mean is our best guess (without any additional information)!
- We learned earlier that the sum of squared residuals from a regression model is called the Sum of Squared Errors (SSE):

$$SSE = \Sigma(y_i - \hat{y}_i)^2$$

Coefficient of Determination: R-squared

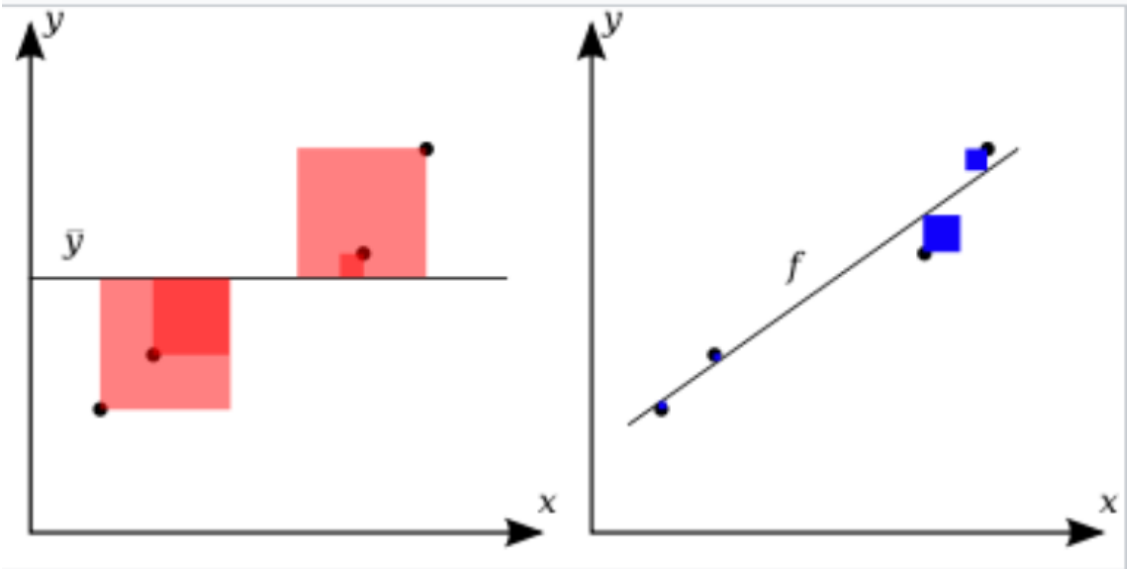
- By providing information on the X variable, we are able to reduce the amount of variation in Y from TSS to SSE. The proportion reduction in TSS is called **R-squared** (coefficient of determination):

$$R^2 = \frac{TSS - SSE}{TSS} \tag{40}$$

- It can be shown that $R^2 = \rho^2$ in the case of bivariate linear regressions.
- For example, earlier we know that the correlation between income and education in an occupation is 0.5776, so we know that the R^2 in this bivariate regression is $0.5776^2 = 0.3336$. We can also find this in the STATA output.

Coefficient of Determination: R-squared

- If we have no other variables, then the best we can do is to predict Y using the mean of Y. In this case, the variance of Y will be the "residual," and the R^2 is zero.
- The better the linear regression (on the right) fits the data in comparison to the simple average (on the left graph), the closer the value of R^2 is to 1.

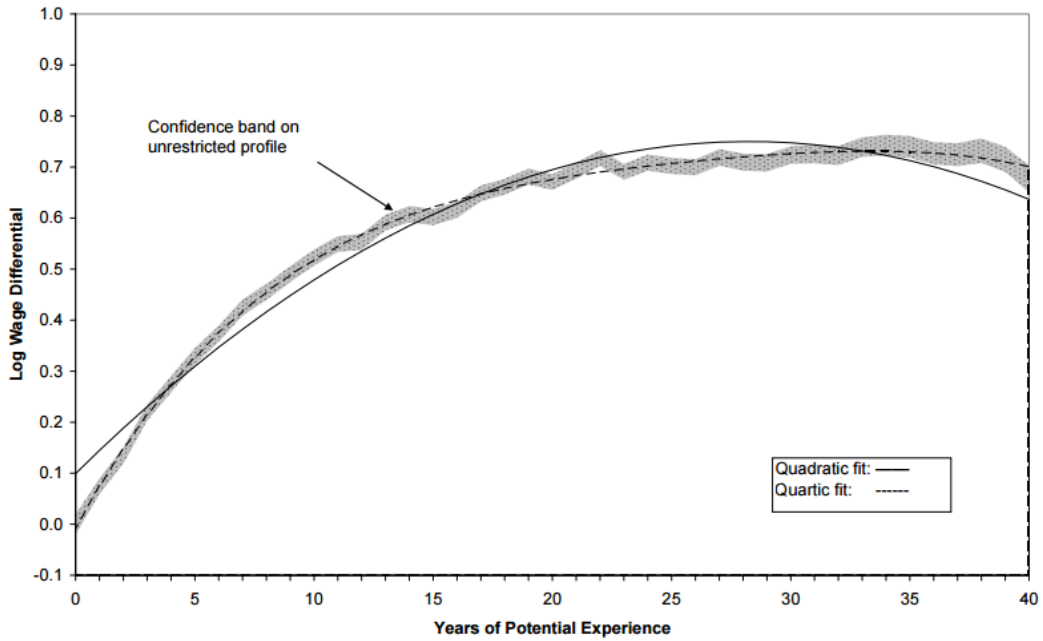


Some More Practical Thoughts on Linear Regression

Non-linear relationships?

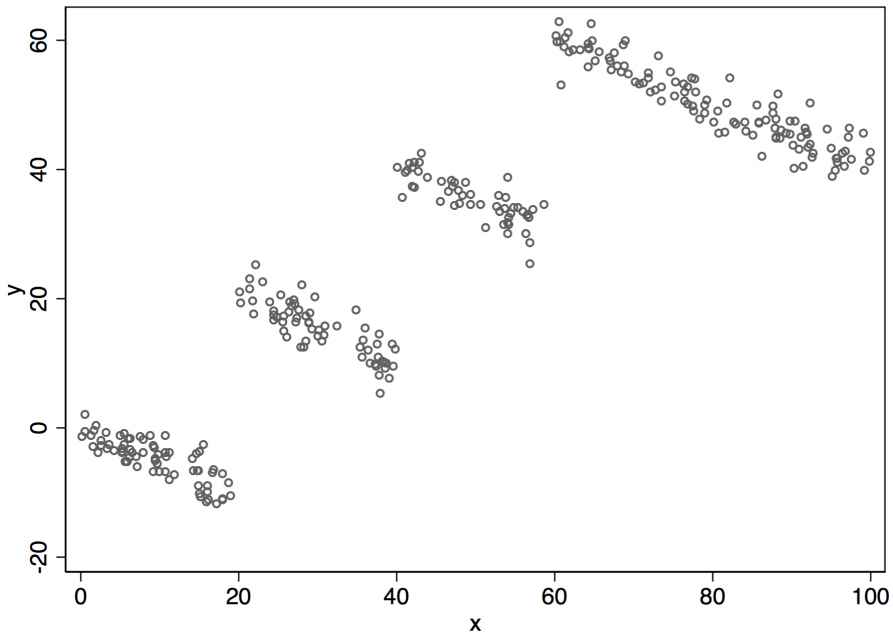
- What would you do when the relationship is non-linear in X? From Lemieux 2003:

Figure 11.8: Experience Profiles for Men, 1999-2001 CPS



Ecological Fallacy

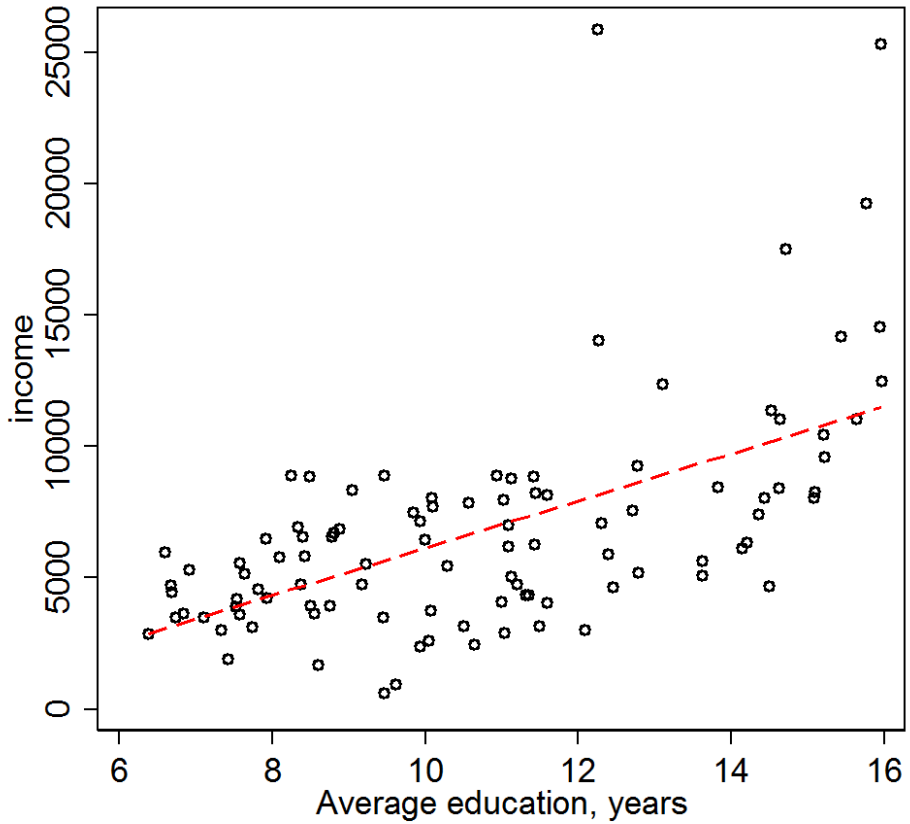
- Does the association between groups always consistent with the association within groups?
- Look at the following problem of an ecological fallacy:



- Ecological fallacy occurs when the relationship between two variables differs within subgroups compared to that observed for the aggregated data. (Simpson's paradox)

Sensitivity to Outliers

- Is the OLS estimator sensitive to outliers in the data?
- It's always a good idea to see the scatterplot of your data first:



Sensitivity to Outliers

- Now let's run the regression excluding those with income greater than 20,000 (potentially outliers):

```
. regress income educat if income<=20000
```

Source	SS	df	MS	Number of obs	=	100
				F(1, 98)	=	59.46
Model	415369013	1	415369013	Prob > F	=	0.0000
Residual	684601386	98	6985728.43	R-squared	=	0.3776
				Adj R-squared	=	0.3713
Total	1.1000e+09	99	11110812.1	Root MSE	=	2643.1

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educat	758.5689	98.37479	7.71	0.000	563.3473	953.7904
_cons	-1672.395	1082.481	-1.54	0.126	-3820.544	475.7534

Sensitivity to Outliers

- Some practical strategies to deal with outliers:
- The simple strategy is to drop them!
- But when data size is large and you have a number of variables, it is not always straightforward to detect outliers.
- Taking the log may help:

```
. gen logincome=log(income)

. regress logincome educat
```

Source	SS	df	MS	Number of obs	=	102
				F(1, 100)	=	42.94
Model	10.6141108	1	10.6141108	Prob > F	=	0.0000
Residual	24.7168977	100	.247168977	R-squared	=	0.3004
				Adj R-squared	=	0.2934
Total	35.3310085	101	.349811966	Root MSE	=	.49716

logincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educat	.1188136	.018131	6.55	0.000	.0828422	.1547849
_cons	7.384684	.200818	36.77	0.000	6.986266	7.783101

Sensitivity to Outliers

- Another strategy is the run the **quantile regression** with quantile set at the 50th percentile, which predicts conditional median instead of *conditional mean*:

less sensitive to outliers

```
. qreg income educat, quantile(0.5)
Iteration  1:  WLS sum of weighted deviations =  120436.36

Iteration  1: sum of abs. weighted deviations =  120611.73
Iteration  2: sum of abs. weighted deviations =  118824.04
Iteration  3: sum of abs. weighted deviations =  118301.94

Median regression                               Number of obs =        102
  Raw sum of deviations   143188 (about 5902)
  Min sum of deviations 118301.9                Pseudo R2     =     0.1738
```

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educat	672.5275	138.8062	4.85	0.000	397.14	947.9149
_cons	-876.0548	1537.412	-0.57	0.570	-3926.236	2174.126

Sensitivity to Outliers

```
. qreg income educat if income<=20000, quantile(0.5)
Iteration  1:  WLS sum of weighted deviations =  101796.81

Iteration  1: sum of abs. weighted deviations =  101643.05
Iteration  2: sum of abs. weighted deviations =  101303.44
Iteration  3: sum of abs. weighted deviations =  101239.31

Median regression                               Number of obs =        100
  Raw sum of deviations 123496.5 (about 5811)
  Min sum of deviations 101239.3                Pseudo R2     =     0.1802
```

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educat	650.4835	137.392	4.73	0.000	377.8336	923.1335
_cons	-705.6459	1511.812	-0.47	0.642	-3705.788	2294.496

```
. di 650.4835/672.5275
.96722216
```


PHILIP COLLA