ECON 4003 Econometrics I

Empirical Exercise 3.1

By Duong Trinh

University of Glasgow

Picture the Scenario

How much does EDUCATION affect WAGE RATES?

Dataset: **cps5_small.dta**

- from the 2013 Current Population Survey (CPS)
- contains 1,200 observations

  on hourly wage rates, education, and other variables

(a1) Summary statistics and histogram for WAGE

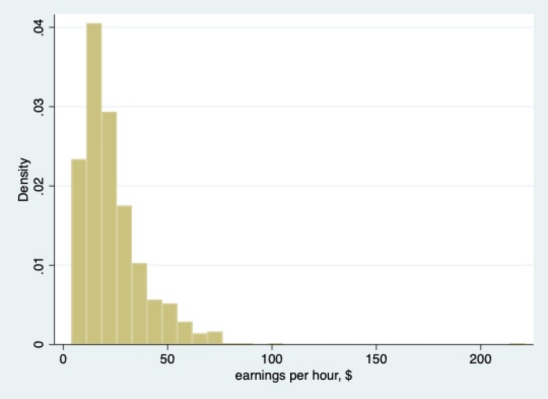
Half of the sample earns an hourly wage of more than \$19.30 per hour, with the average being \$23.64 per hour. The maximum earned in this sample is \$221.10 per hour and the least earned in this sample is \$3.94 per hour.

variable	n	min	25th pct	median	mean	75th pct	max
wage	1200	3.94	13.00	19.30	23.64	29.80	221.1

Percentiles

(a1) Summary statistics and histogram for WAGE

The observations for *wage* are **skewed to the right** indicating that most of the observations lie between the hourly wages of 5 to 50, and that there is a smaller proportion of observations with an hourly wage greater than 50.



(a2) Summary statistics and histogram for EDUCATION

25% of the people had up to 12 years of education.

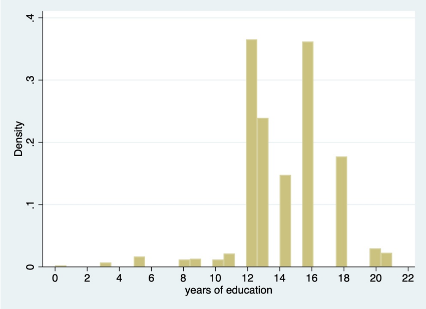
variable	n	min	25th qtile	median	mean	75th qtile	max
educ	1200	0	12.0	14.0	14.2	16.0	21.0

Percentiles

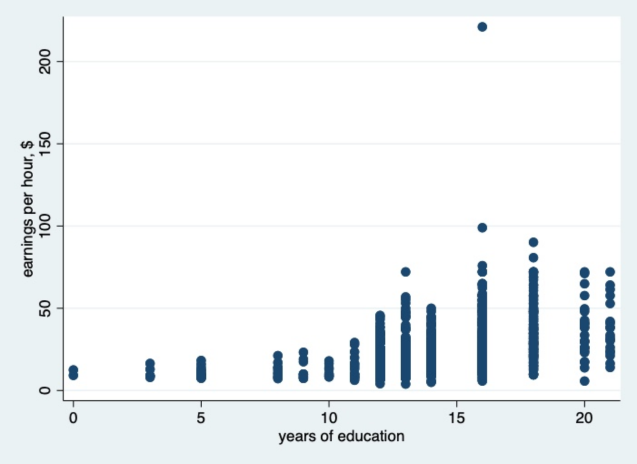
(a2) Summary statistics and histogram for EDUCATION

The spike at 12 years of education describes those who finished their education at the end of high school. There are a few observations at less than 12, representing those who did not complete high school.

Spikes at 13 and 14 years are people who had 1 or 2 years at college. Spike at 16 years describes those who completed a 4-year college degree, while those at 18 and 21 years represent a master's degree, and further education such as a PhD, respectively.



(b1) Scatterplot



There appears to be a *positive* relationship between wage and education.

association

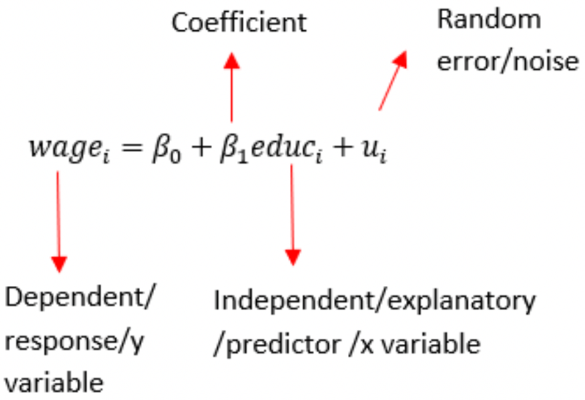
(b2) Covariance and Correlation

The sample covariance and correlation also suggest a *positive* relationship between wage and education.

$$\text{Cov}(\textit{wage}, \textit{educ}) = 20.029 \qquad \text{Corr}(\textit{wage}, \textit{educ}) = 0.455$$

association

(c) Linear regression model



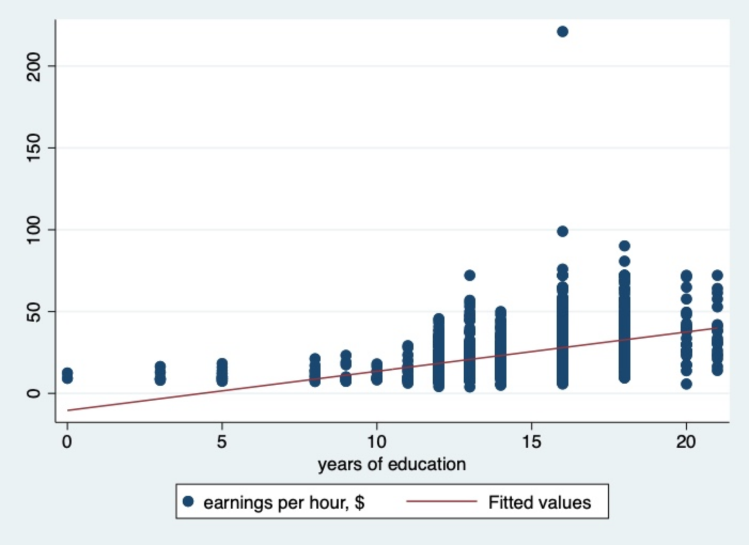
(c1) Estimation results

$$\underset{(se)}{\widehat{wage}} = -\underset{(1.9624)}{10.4000} + \underset{(0.1354)}{2.3968} \cdot educ \qquad R^2 = 0.2073 \quad SER = 13.553$$

- $\hat{\beta}_1$: The slope estimate 2.3968 suggests that an extra year of education *is associated with* an increase in hourly wage rate by \$2.3968 *on average*.

- $\hat{\beta}_0$: The intercept estimate -10.4 represents the estimated average hourly wage rate of a worker with no years of education. It should not be considered meaningful as it is not possible to have a negative hourly wage rate.

(c2) Scatterplot with regression line



(d) Linear regression with new variable lnwage

Log-Level model:

$$lnwage_i = \gamma_0 + \gamma_1 . educ_i + \epsilon_i$$

Estimation results:

$$\underset{(se)}{\widehat{lnwage}} = \underset{(0.070)}{1.5968} + \underset{(0.0048)}{0.0988} \cdot educ \qquad R^2 = 0.2557 \quad SER = 0.4847$$

- $\hat{\gamma}_1$: The slope estimate 0.0988 suggests that an extra year of education *is associated with* an increase in hourly wage rate by 9.88% *on average*.
- Brief proof:

$$\Delta log(wage) = \gamma_1 \Delta educ \implies \% \Delta wage = 100 \gamma_1 \Delta educ$$

$$\text{since} \quad \Delta log(x) = log(x_1) - log(x_0) \approx \frac{x_1 - x_0}{x_0} = \frac{\Delta x}{x} = \frac{\% \Delta x}{100}$$

log-level

(e) Linear regression for subgroups

Dependent variable: Log of hourly wage				
	(1)	(2)	(3)	(4)
	Male	Female	White	Black
Years of education	0.095	0.114	0.099	0.090
	(0.006)	(0.008)	(0.005)	(0.017)
Constant	1.724	1.272	1.603	1.638
	(0.088)	(0.113)	(0.074)	(0.239)
N	672	528	1095	105
R^2	0.263	0.297	0.260	0.210
SER	0.481	0.470	0.487	0.452

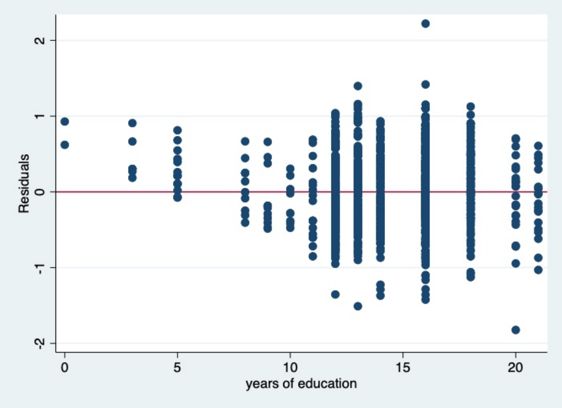
Note: Standard errors are in parenthesis.

(e) Linear regression for subgroups

The percentage increase in hourly wage associated with an extra year of education:

- is larger for female workers (11.4% per year) than for male workers (9.5% per year) on average.
- is larger for white workers (9.9% per year) than for black workers (9% per year) on average.

(g) Least squares residuals

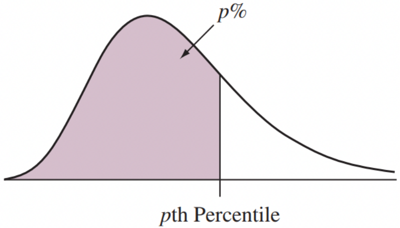


As *educ* increases, the spread of the residuals also increases, suggesting that the error variance is larger for larger values of *educ* - a violation of assumption **SR.5 homoskedasticity**.

Percentiles - Definition

The p^{th} **percentile** is a value such that p percent of the observations fall below or at that value.

- The 50^{th} percentile is usually referred to as the **median** ($p = 50$): 50% of the observations fall below or at it and 50% above it.

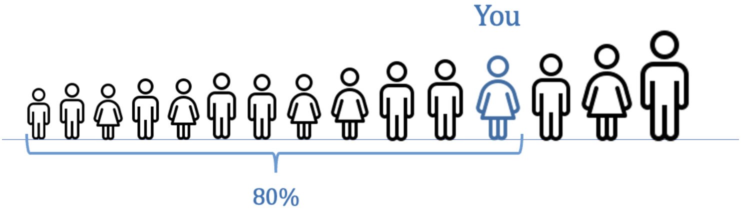


(a1)

Percentiles - Example

You are the fourth tallest person in a group of 15.

$\Longrightarrow$ 80% of people are shorter than or as high as you:

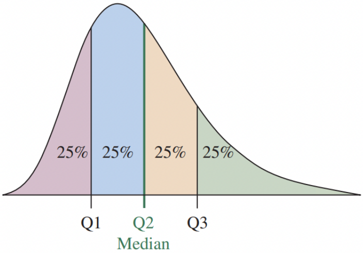


That means you are at the 80^{th} percentile.

If your height is 1.75m then "1.75m" is the 80^{th} percentile height in that group. (a1)

Percentiles - Quantiles

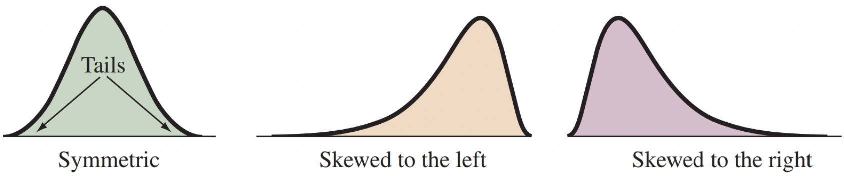
Three useful percentiles are the **quartiles**. The quartiles split distribution into four parts, each containing one quarter (25%) of the observations.



- The **first quartile** has $p = 25$, so it is the 25^{th} percentile.
- The **second quartile** has $p = 50$, so it is the 50^{th} percentile, which is the median.
- The **third quartile** has $p = 75$, so it is the 75^{th} percentile.

(a1)

Skewed Distribution - Definition



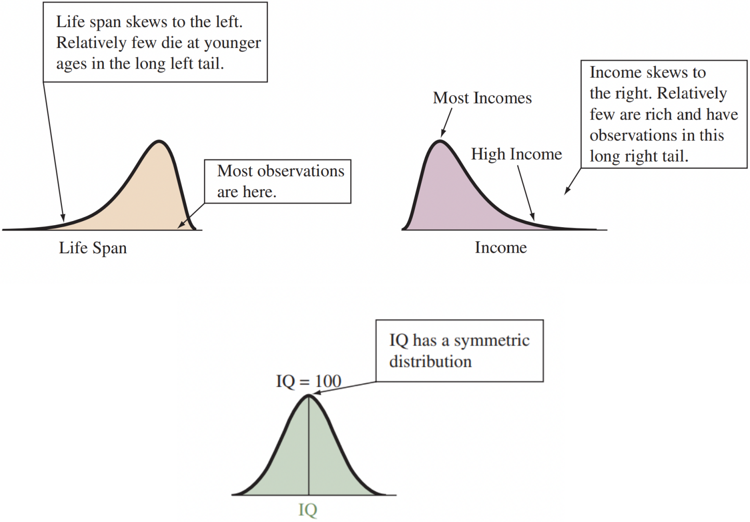
Curves for Distributions Illustrating Symmetry and Skew

To **skew** means to stretch in one direction.

- A distribution is *skewed to the left* if left tail is longer than right tail.
- A distribution is *skewed to the right* if right tail is longer than left tail.
- A left-skewed distribution stretches to the left and A right-skewed distribution stretches to the right.

(a1)

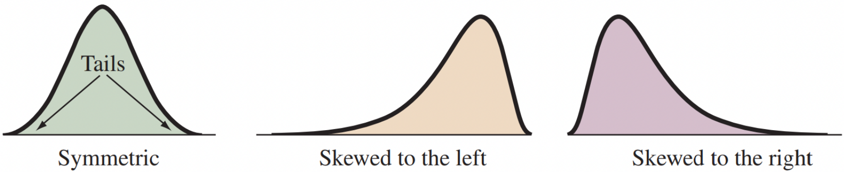
Skewed Distribution - Example



Skewness

Skewness measures **the degree and direction of asymmetry**.

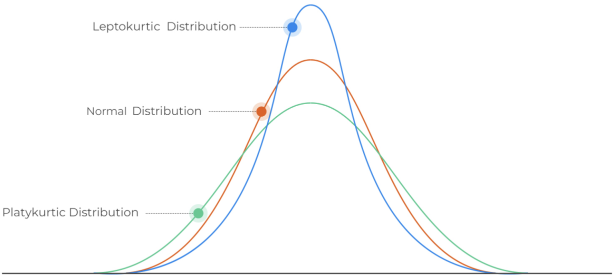
$$\text{skew}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}$$



- A *symmetric* distribution has a skewness of *0*.
- A *left-skewed* distribution has a *negative* skewness.
- A *right-skewed* distribution has a *positive* skewness.

(a1)

Kurtosis

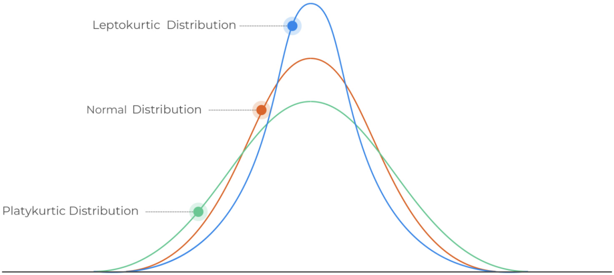


Kurtosis is a measure of **the heaviness of the tails** of a distribution.

$$\text{Kurt}[X] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4},$$

(a1)

Kurtosis (cont.)



Kurtosis is a measure of **the heaviness of the tails** of a distribution.

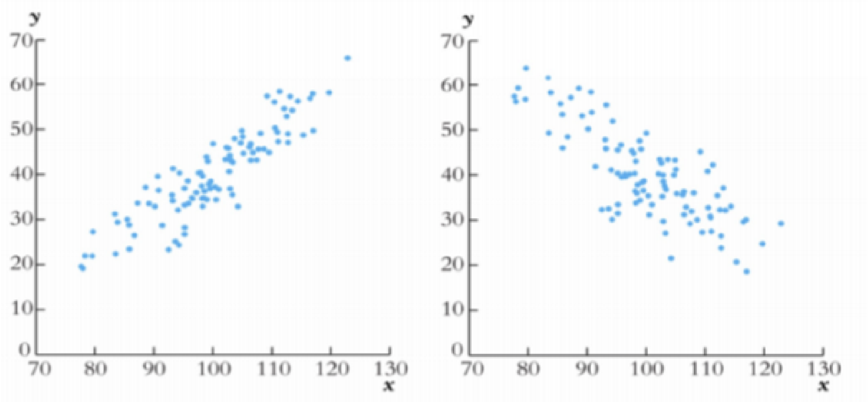
- A normal distribution has a kurtosis of 3.
- Heavy tailed distributions will have kurtosis greater than 3.
- Light tailed distributions will have kurtosis less than 3.

(a1)

Association - Scatterplot

Looking for **trend** of the **association** between two quantitative variables:

- **Positive association**: As x goes up, y tends to go up.
- **Negative association**: As x goes up, y tends to go down.



(b1)

Association - Correlation

Summarizing **direction** and **strength** of the **linear** (straight-line) **association** between two quantitative variables.

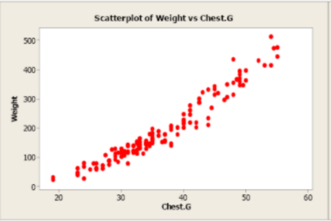
$$r == \frac{1}{n-1}\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) \tag{1}$$

Correlation coefficient r takes values between -1 and +1.

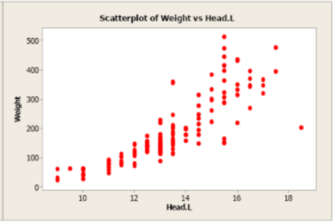
- **Direction**
  - $r > 0$ indicates a positive association
  - $r < 0$ indicates a negative association
- **Strength**
  - The closer r is to +/-1 the closer the data points fall to a straight line, and the stronger the linear association is.
  - The closer r is to 0, the weaker the linear association is.

(b2)

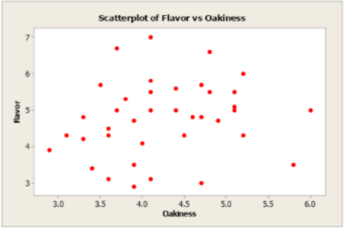
Association - Correlation (cont.)



Strong positive relationship
r = 0.96

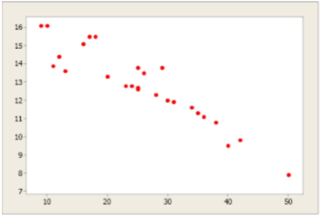


Moderate positive relationship
r = 0.67

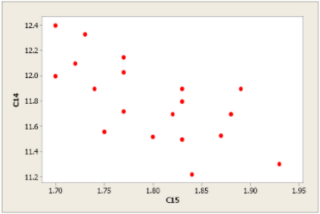


Very weak positive relationship
r = 0.07

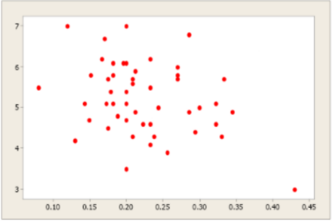
Association - Correlation (cont.)



Very strong negative relationship
r = -0.93



Moderately strong negative relationship
r = -0.67

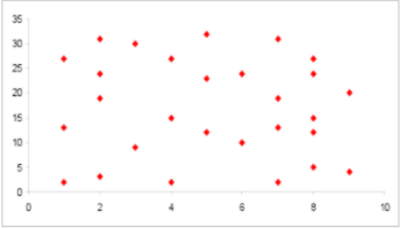


Very weak negative relationship
r = -0.13

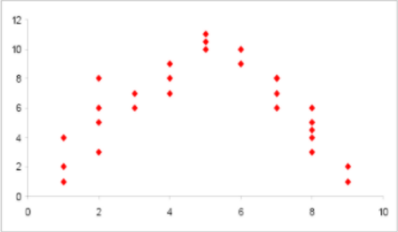
Association - Correlation (cont.)

(!) Correlation poorly describes the association when the relationship is curved (non-linear).

Plot 1



Plot 2



For this U-shaped relationship, the correlation is 0 (or close to 0), even though the variables are strongly associated. Ignoring the scatterplot could result in a serious mistake when describing the relationship between two variables.

Association - In Practice

1. When you investigate the relationship between two quantitative variables, always **begin with a scatterplot**. This graph allows you to look for patterns (both linear and non-linear).

2. The next step is to quantitatively describe the strength and direction of the linear relationship (an approximate straight-line relationship) using the **correlation coefficient "r"**.

3. Once you have established that a linear relationship exists, you can take the next step in model building.

Functional Forms Involving Logarithms

Constant unit change/ Constant percentage change/ Constant elasticity?

Interpret Slope Coefficient Estimates

Model	Interpretation of $\widehat{\beta}_1$
Level-level $Y_i = \beta_0 + \beta_1 X_i + u_i$	An increase in X by 1 unit is associated with a change in Y by $\widehat{\beta}_1$ units on average
Log-level $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$	An increase in X by 1 unit is associated with a change in Y by $(100 \times \widehat{\beta}_1)\%$ on average
Level-log $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$	An increase in X by 1% is associated with a change in Y by $(\widehat{\beta}_1/100)$ units on average
Log-log $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$	An increase in X by 1% is associated with a change in Y by $\widehat{\beta}_1\%$ on average

Functional Forms Involving Logarithms (cont.)

Why **logarithmic transformation**?

- Meaningful interpretation: reasonable, consistent with economic theories.
- Yields a distribution that is closer to normal $\Longrightarrow$ better for inference purpose.

(d)