Lecture 1
Economic Data and Simple Linear Regression

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

February 22, 2022

Lecture Outline

1. Economic Data
   - Taiwan's Macroeconomic Data
   - Taiwan's Microeconomic Data
   - Why Econometrics?
2. Simple Linear Regression
   - Least-Squares Minimization
   - Algebraic Properties of LS Estimation
   - Statistical Properties of LS Estimation

Introduction

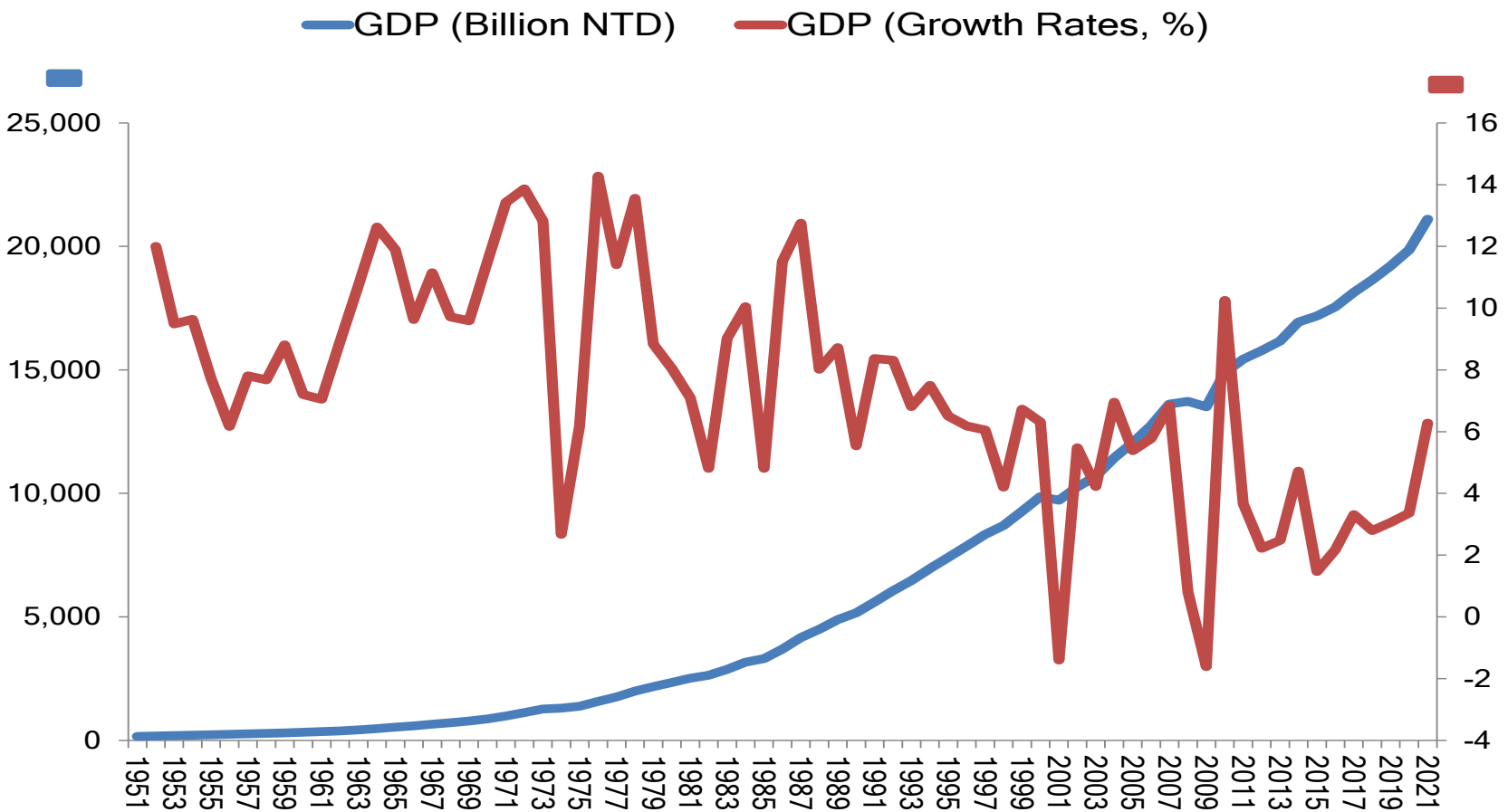
- Economic data are records of economic activities. Such data are usually compiled by government agencies (e.g. GDP and unemployment rates), collected from controlled experiments or surveys (e.g. Survey of Family Income and Expenditure), or recorded by some electronic systems (e.g. stock market transaction data).
- Internet activities, such as visiting a website, posting on a social media, shopping or booking online, and clicking through an on-line ad, also produce a large amount of data unintentionally. Such data are also known as "digital footprints".
- For some analysis, artificial data may be generated (simulated) computationally using certain algorithms or randomly re-arranged by some re-sampling methods.

- Economic data may be time series if they are recorded over a period of time, cross section data if they are recorded across different units (agents, households, firms, industries, or countries) at a particular time point, or panel data if they are recorded across different units over a period of time.
- Econometrics offers various statistical, mathematical and computational methods that can be used to establish (or analyze) economic relations based on economic data.
- Econometric analysis typically relies on numeric data; text documents are typically converted to numeric data (for example, using some text mining techniques) before they can be analyzed by econometric methods.

GDP and Unemployment Rates

- Taiwan's GDP data are collected and compiled by the DGBAS (Directorate General of Budget, Accounting and Statistics).
  - Annual data since 1951
  - Quarterly data since 1961
  - Seasonally adjusted, quarterly data since 1982
- Since Nov. 2014, all national income statistics have been calculated in accordance with the guideline of United Nations (2008SNA). In particular, the GDP growth rates are now computed using the chain-linked method.
- Taiwan's unemployment data are also collected by the DGBAS.
  - Monthly data since 1978
  - Seasonally adjusted, monthly data available from 2011

Taiwan's Annual GDP: 1951–2021

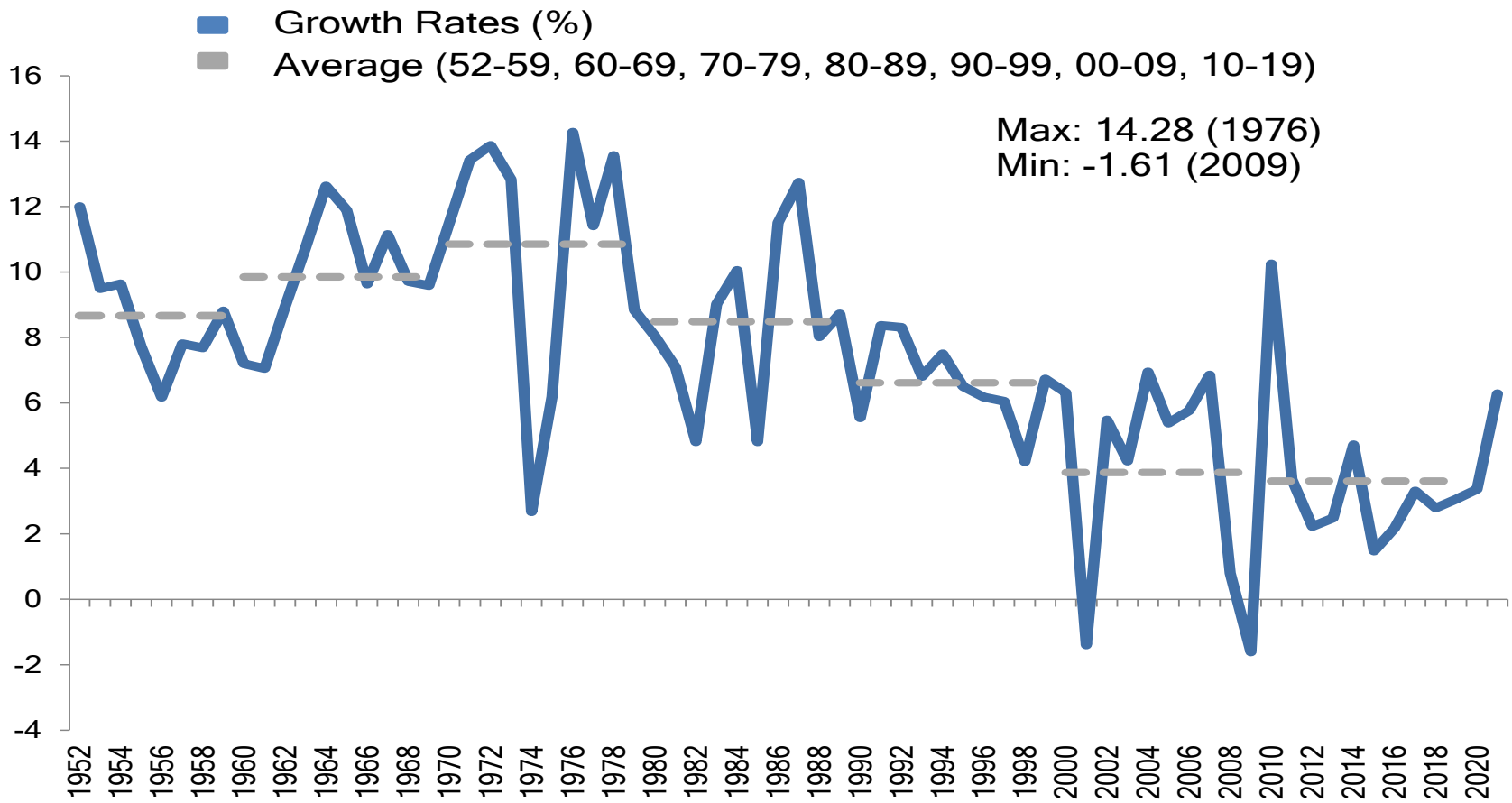


GDP and its growth rates (average: 7.31%; s.d.: 3.63%)

Summary Statistics of Annual GDP Growth Rates

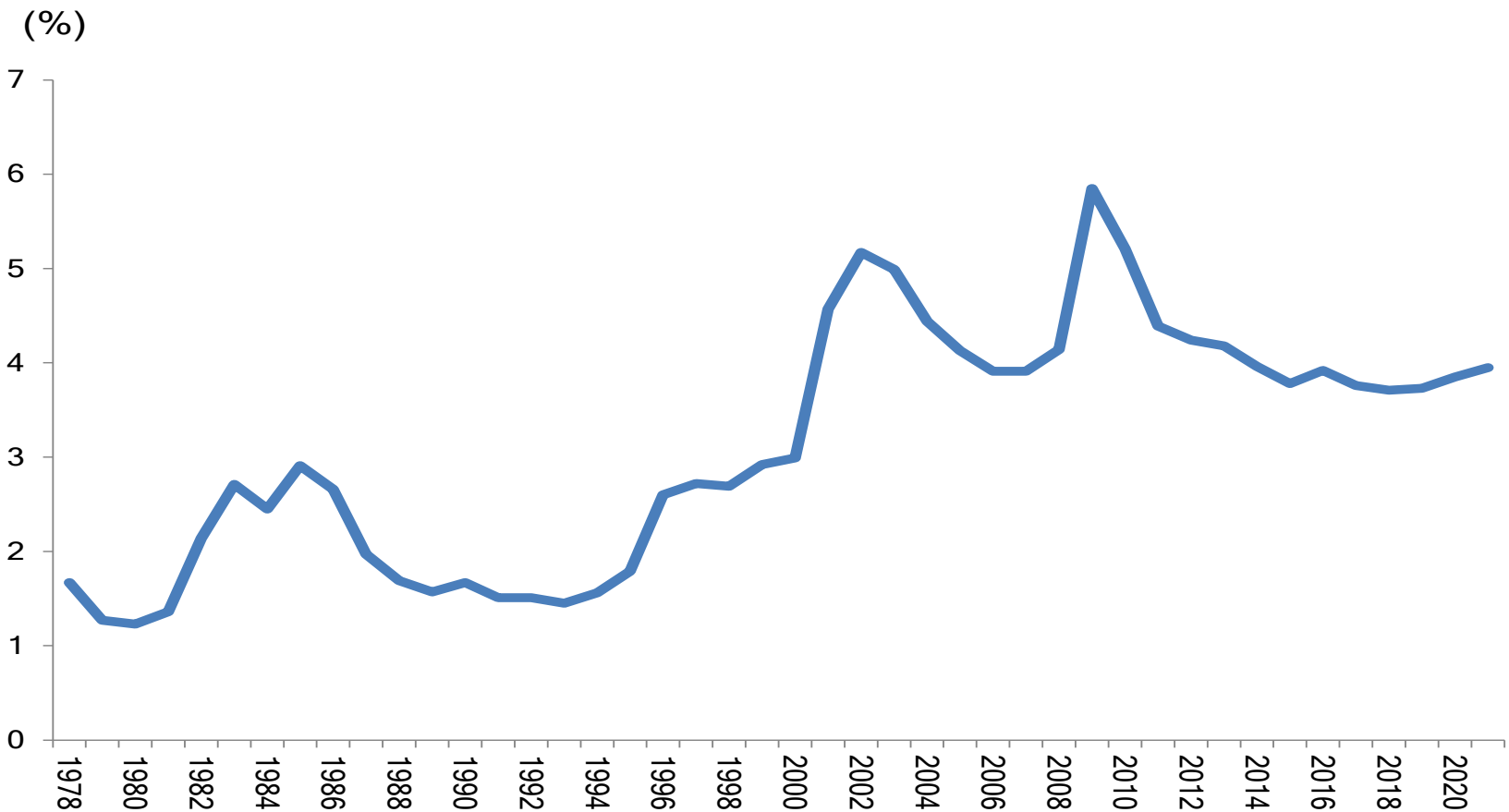
Period	Avg	S.d.	Max	Min
52–21	7.31	3.63	14.28	−1.61
52–59	8.67	1.76	12.00	6.17
60–69	9.85	1.83	12.63	7.05
70–79	10.86	3.83	14.28	2.67
80–89	8.48	2.57	12.75	4.81
90–99	6.62	1.26	8.37	4.20
00–09	3.88	3.34	6.95	−1.61
10–19	3.51	2.51	10.25	1.47

Taiwan's GDP Annual Growth Rates: 1952–2021



GDP and its growth rates (with 10-year averages)

Taiwan's Annual Unemployment Rates: 1978–2021

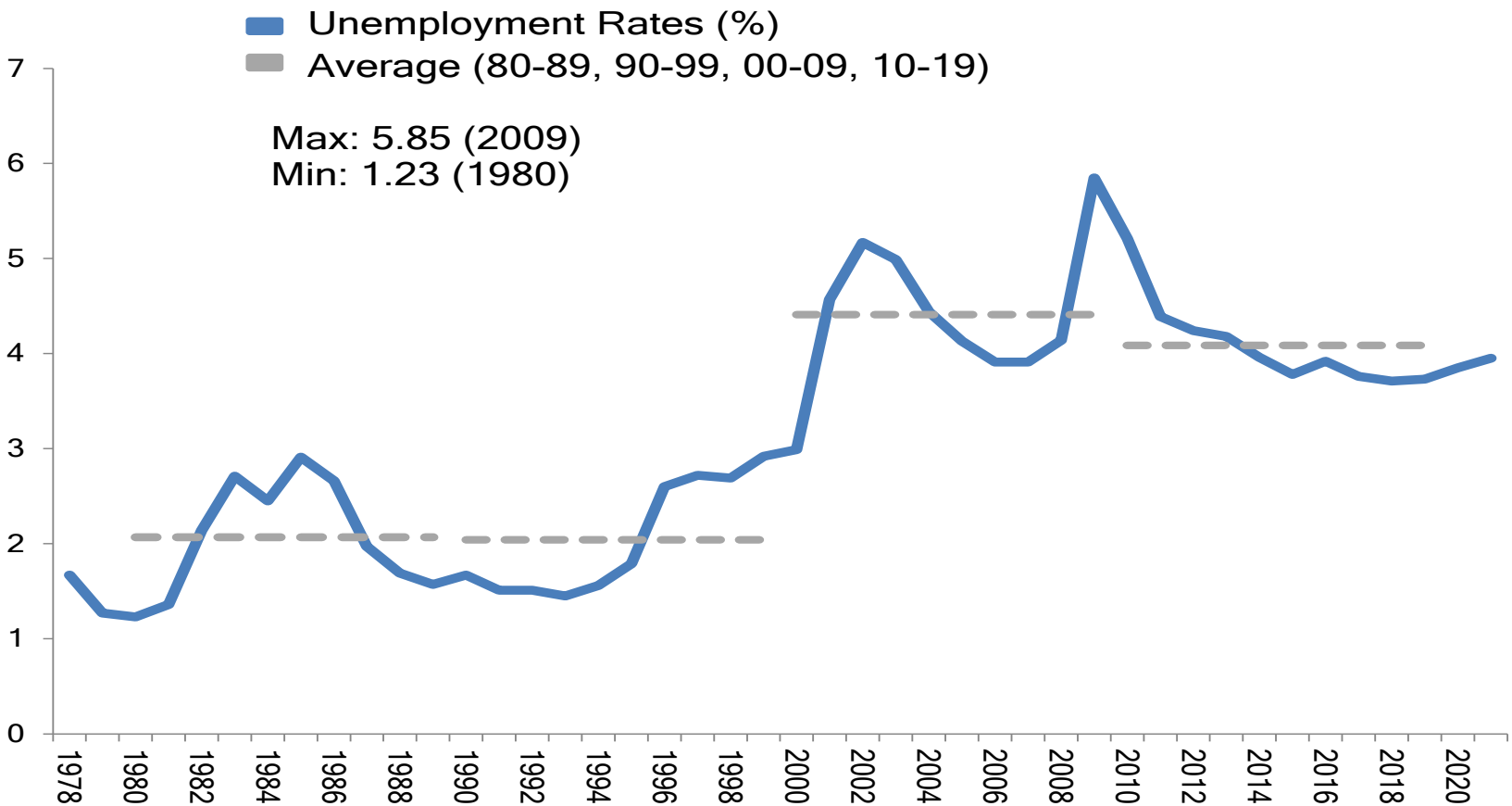


Unemployment rates (average: 3.11%; s.d.: 1.27%)

Summary Statistics of Annual Unemployment Rates

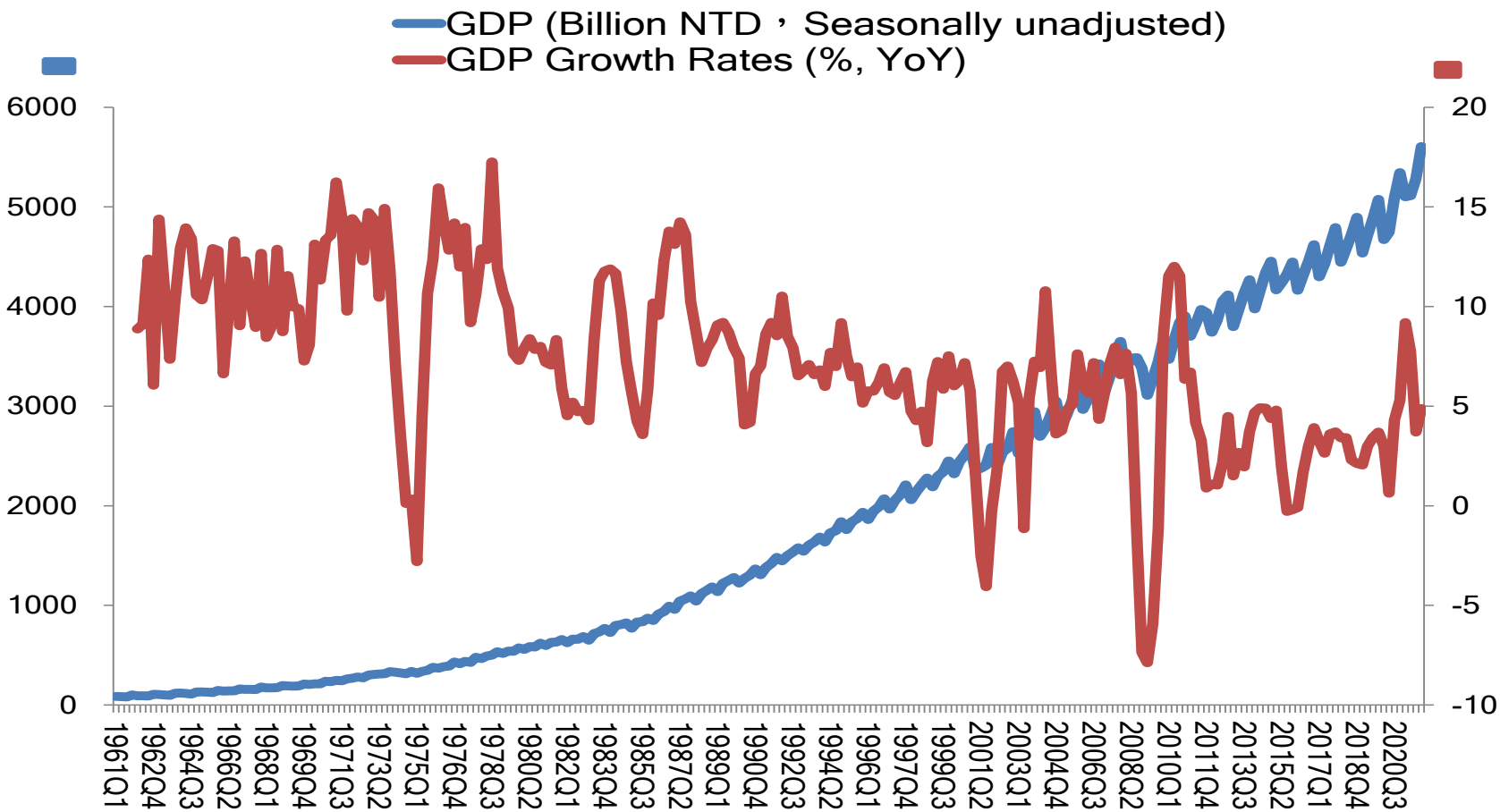
Period	Avg	S.d.	Max	Min
78–21	3.11	1.27	5.85	1.23
80–89	2.07	0.60	2.91	1.23
90–99	2.04	0.61	2.92	1.45
00–09	4.41	0.79	5.85	2.99
10–19	4.09	0.46	5.21	3.71

Taiwan's Annual Unemployment Rates: 1978–2021



Unemployment rates (with 10-year averages)

Taiwan's Quarterly GDP: 1961Q1–2021Q4



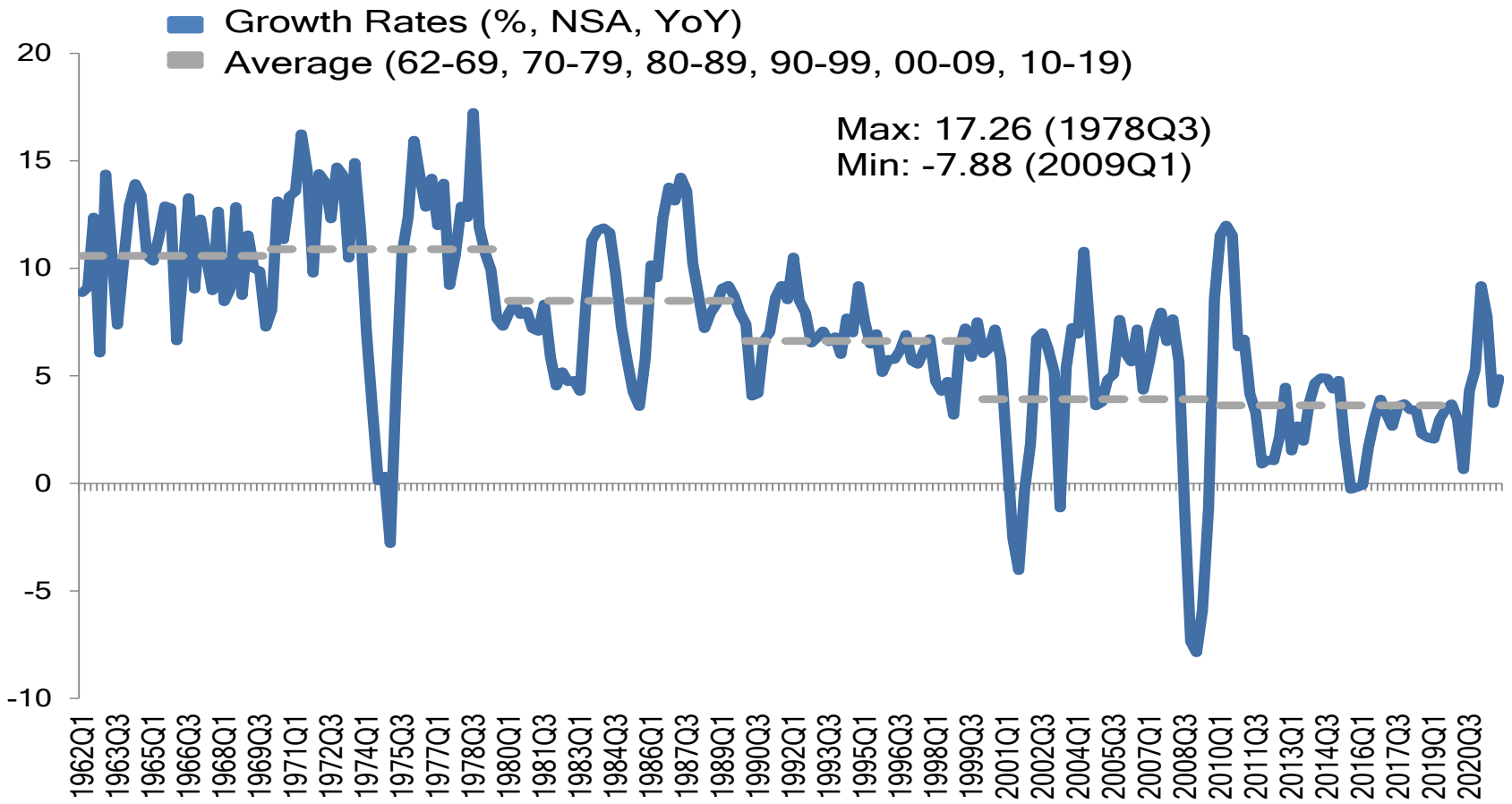
Seasonally unadjusted GDP and its YoY growth rates

(average: 7.17%; s.d.: 4.35%)

Summary Statistics of Quarterly GDP Growth Rates (YoY)

Period	Avg	s.d.	Max	Min
62–21	7.17	4.35	17.26	−7.88
62–69	10.59	2.23	14.39	6.05
70–79	10.89	4.54	17.26	−2.81
80–89	8.49	2.86	14.25	3.57
90–99	6.63	1.53	10.53	3.16
00–09	3.92	4.57	10.88	−7.88
10–19	3.60	2.83	12.02	−0.28

Taiwan's Quarterly GDP: 1962Q1–2021Q4



GDP YoY growth rates with 10-year averages

Manpower Utilization Survey

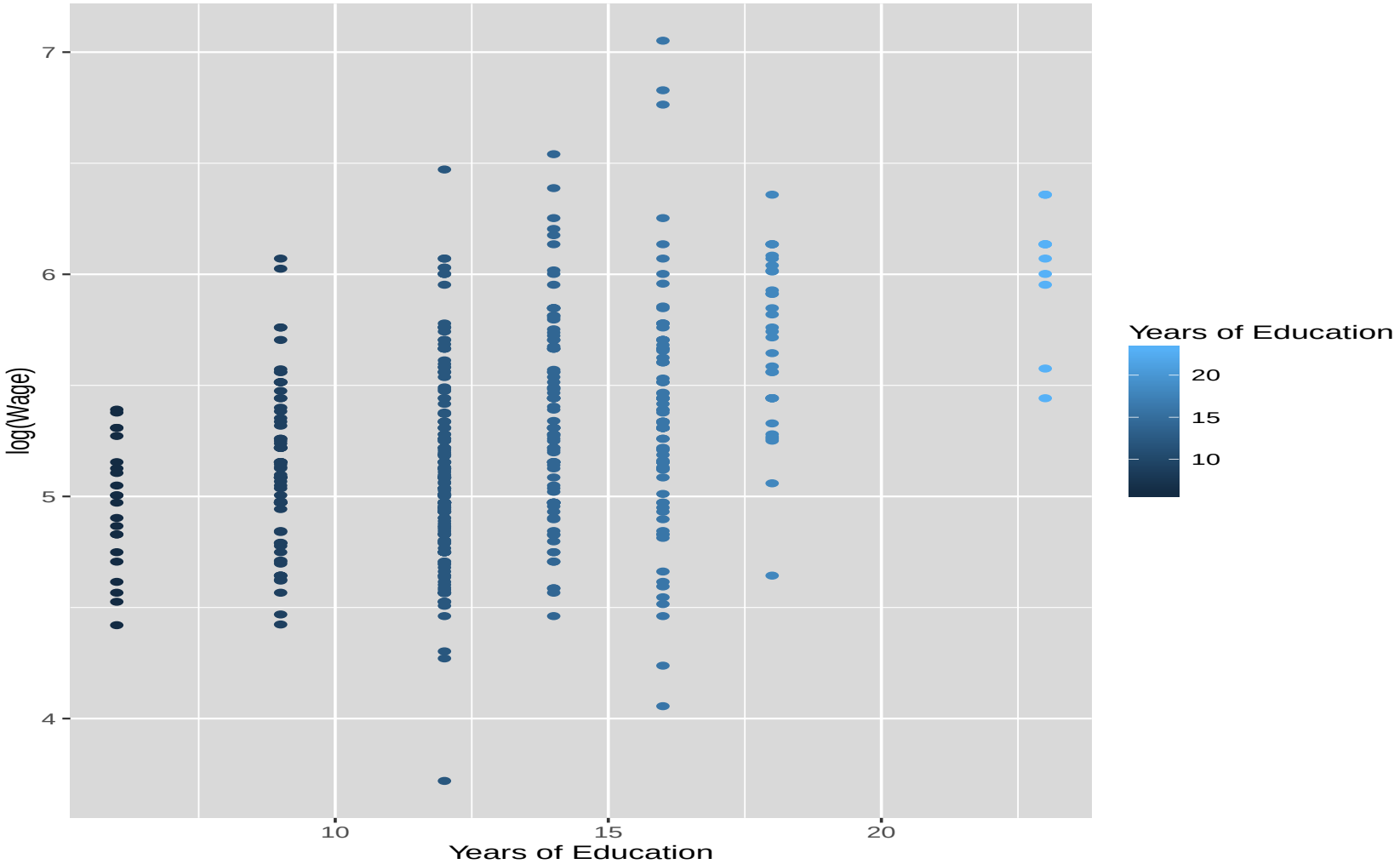
- Manpower utilization survey is conducted with Manpower survey in every May by the DGBAS.
  - Individuals above 15 in every household were surveyed.
  - Approximately 20,000 households, 60,000 individuals were surveyed each time.
- In 2010, there were 11,561 males and 9,348 females surveyed. The questions in the survey include: work status, working hours, earning, education level, etc.

Summary Statistics of log(wage)

		Avg	S.d.	Max	Min
Full Sample		5.13	0.44	7.46	2.82
Education	≤ 9	4.93	0.38	6.83	2.99
Years	10–12	5.00	0.38	6.87	2.82
	≥ 13	5.29	0.44	7.46	3.36
	≤ 5	4.98	0.36	6.54	3.18
Working	6–15	5.12	0.40	7.05	2.82
Experience	16–25	5.21	0.46	7.46	3.36
	≥ 26	5.15	0.49	6.99	2.99

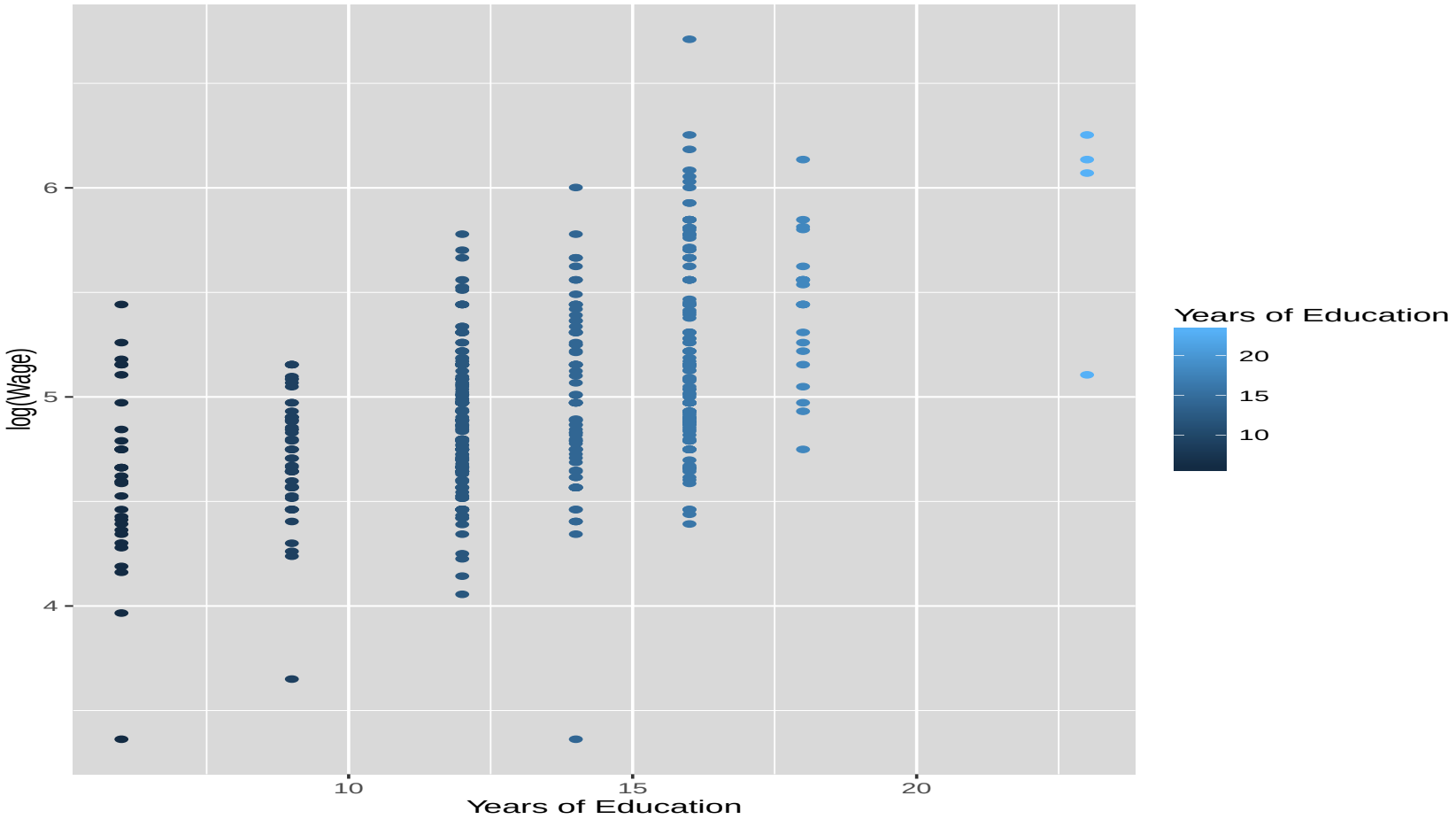
Note: Wage is real hourly wage in NTD; the base year is 2000.

Male Wage vs. Education Level: 2010



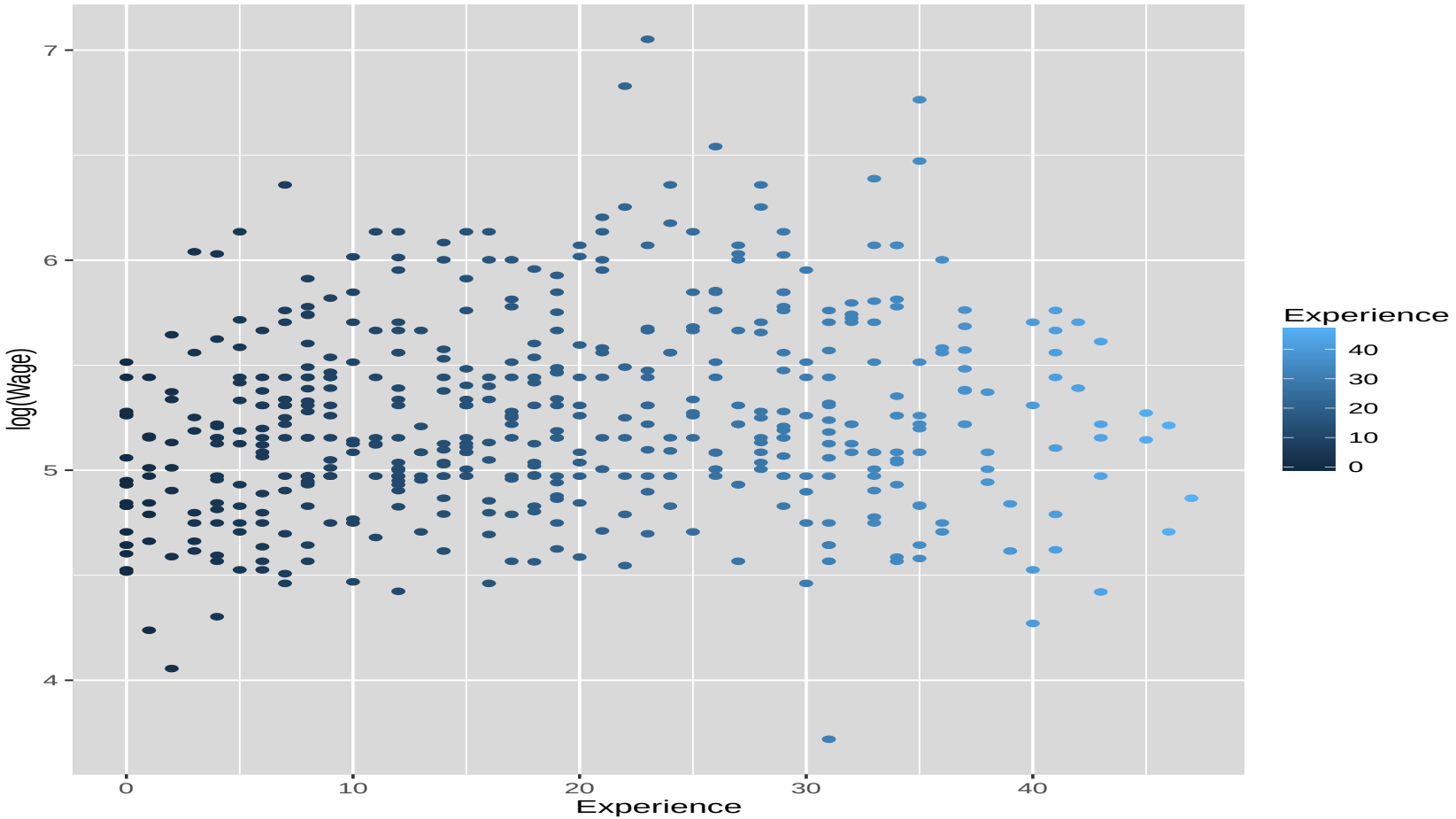
Partial sample of 500 observations

Female Wage vs. Education Level: 2010



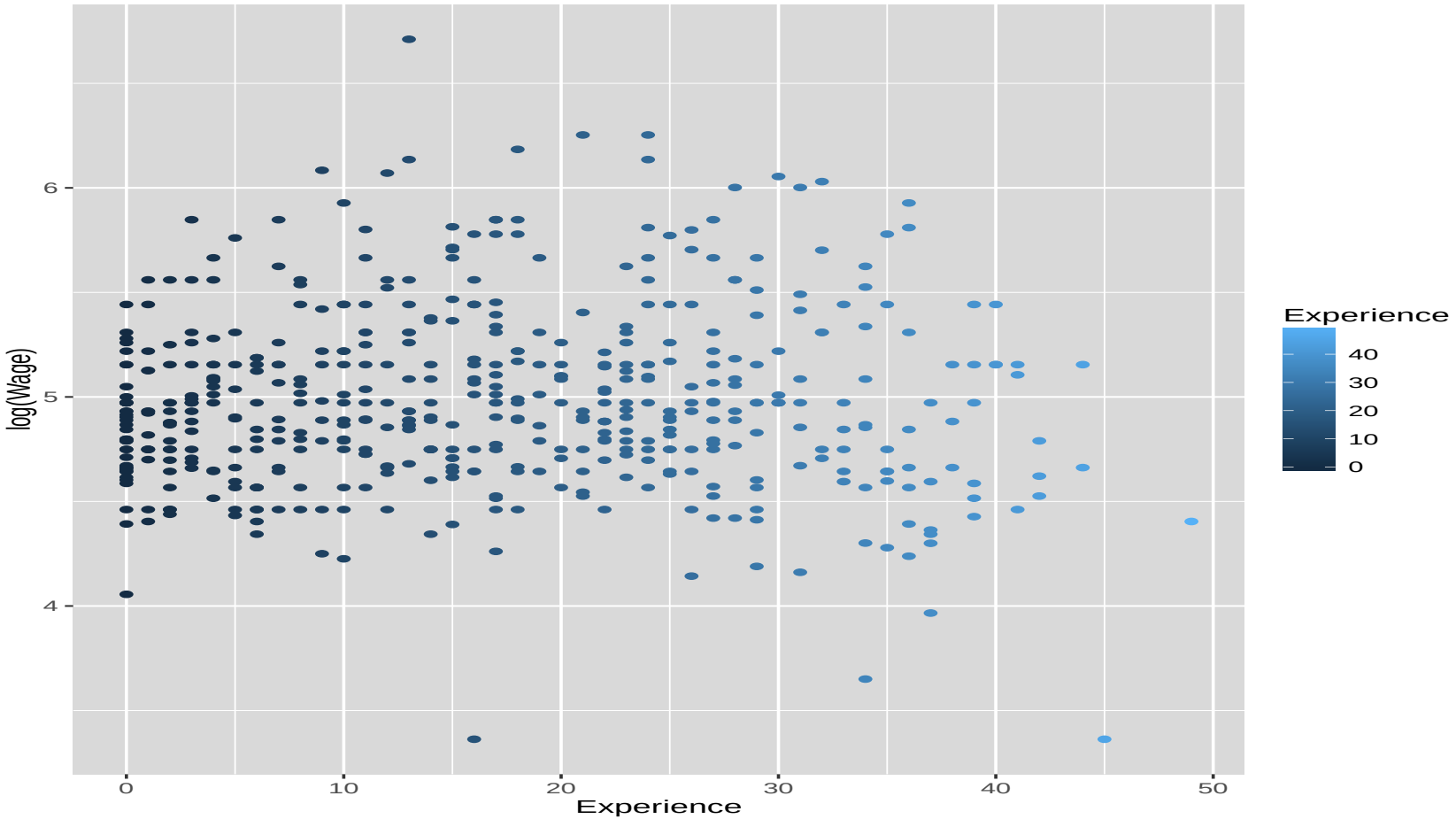
Partial sample of 500 observations

Male Wage vs. Working Experience: 2010



Partial sample of 500 obs; experience = age − education years − 6

Female Wage vs. Working Experience: 2010



Partial sample of 500 obs; experience = age − education years − 6

Why Econometrics

- One may examine data based on their summary statistics (e.g. mean, median, s.d. etc.). Yet, these statistics cannot tell us how one variable is related to other variables.
- Economists are interested in knowing the relations between economic variables. It is thus important to have an analysis of the variable of interest, conditional on other economic variables. This is precisely what econometrics does.
- The purpose of econometrics is to estimate economic relations based on some models, test economic theories and hypotheses, predict the future behavior of economic variables, and/or evaluate the effects of government policies and business programs.

Linear Specification

Given the variable of interest y, we are interested in analyzing the systematic part of y based on the information of another variable x. This systematic component is characterized by a function of x such that

$$y = f(x) + u,$$

where u is the error term, the behavior of y that cannot be explained by $f(x)$. Finding a proper f is a challenging task in econometric analysis. In practice, it is convenient to postulate f as the linear function:

$$f(x) = \beta_0 + \beta_1 x,$$

with unknown parameters β_0 and β_1. The task of finding f thus simplifies to determining proper values for β_0 and β_1.

Specifically, we write

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{systematic part}} + \underbrace{u(\beta_0, \beta_1)}_{\text{error}}.$$

where y is usually referred to as the dependent variable (regressand), x is referred to as the explanatory variable (regressor), and the error depends on the parameter values: $u(\beta_0, \beta_1) = y - (\beta_0 + \beta_1 x)$. Given the sample data (x_i, y_i), $i = 1, \ldots, n$, we have

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \ldots, n,$$

where $u_i = u_i(\beta_0, \beta_1)$ is the i^{th} error term. Our goal is to determine β_0 and β_1 based on the sample information of (x_i, y_i), $i = 1, \ldots, n$. This amounts to finding a line that "best" fits these sample data.

Least-Squares Minimization

The best fit of sample data can be obtained by minimizing the following sum of squared errors with respect to β_0 and β_1:

$$Q_n(\beta_0, \beta_1) := \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2;$$

this is known as the least-squares (LS) minimization problem. The first order conditions (FOCs) are:

$$\frac{\partial Q_n(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0,$$

$$\frac{\partial Q_n(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0,$$

which are two equations with two unknowns: β_0 and β_1.

Least-Squares Estimator

Solving the FOCs for β_0 and β_1 we obtain the ordinary least squares (OLS) estimators of β_0 and β_1 (verify!):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

where $\bar{y} = n^{-1} \sum_{i=1}^n y_i$ and $\bar{x} = n^{-1} \sum_{i=1}^n x_i$. The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are then obtained by plugging the sample data into these estimators.

Remark: The OLS method does not require any assumption on y_i and x_i, except that x_i cannot be a constant. Note that when x_i are a constant c, so is $\bar{x}$. In this case, the denominator of $\hat{\beta}_1$ is 0, and the OLS method breaks down.

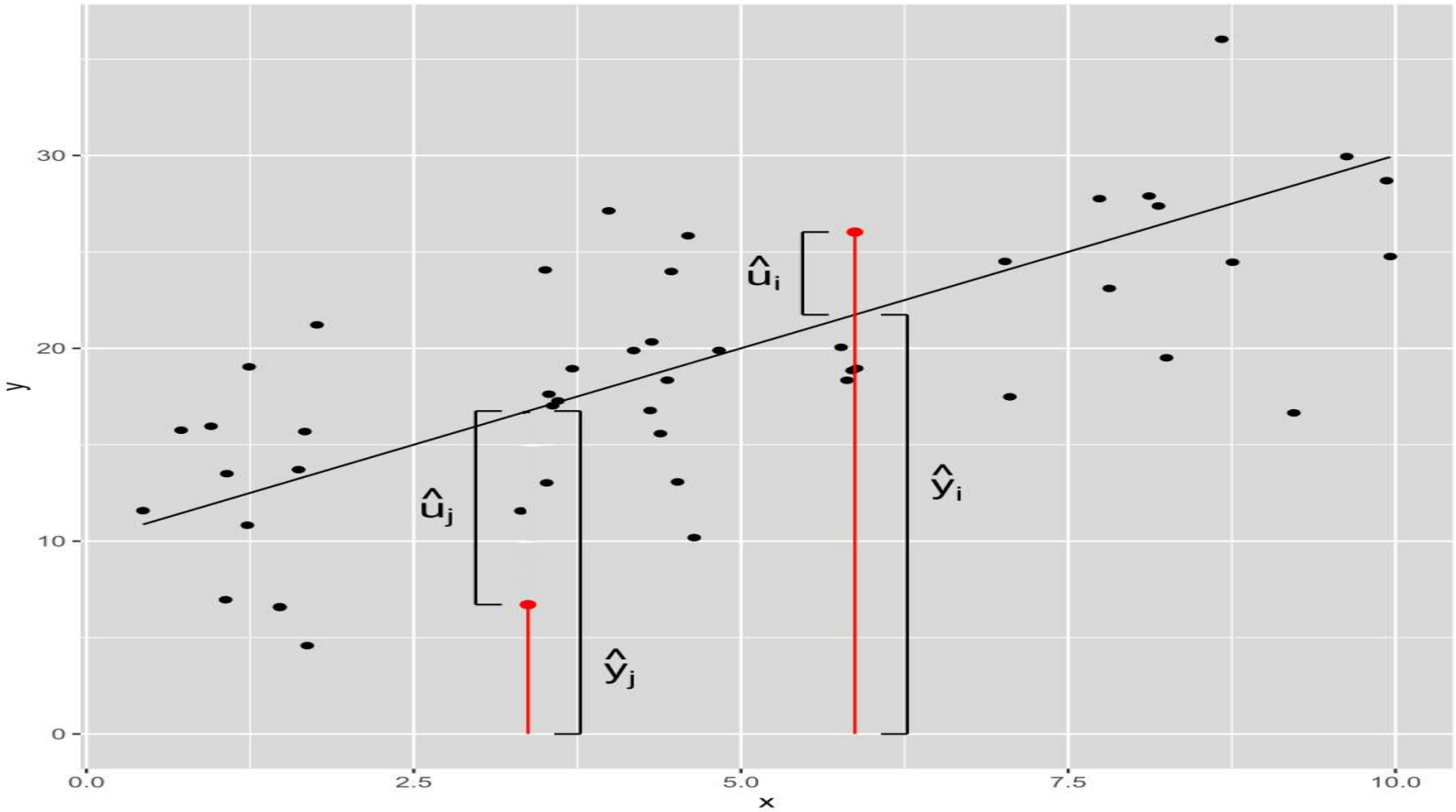
- Estimated regression line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

with the i^{th} fitted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

  - $\hat{\beta}_1 = \mathrm{d}\hat{y}/\mathrm{d}x$ is the slope of the estimated regression line, which predicts how much y would change when x changes by one unit.
  - $\hat{\beta}_0$ is the intercept of the estimated regression line, which predicts the level of y when $x = 0$.

- Residual: $\hat{u} = u(\hat{\beta}_0, \hat{\beta}_1) = y - \hat{y}$ is the error evaluated at $\hat{\beta}_0$ and $\hat{\beta}_1$. The i^{th} residual is $\hat{u}_i = y_i - \hat{y}_i$, the difference between the true value y_i and the fitted value $\hat{y}_i$.



Fitted regression line

Special Case 1

For the specification without the intercept: $y_i = \beta_1 x_i + u_i$, the LS criterion function is:

$$Q_n(\beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2.$$

The FOC is: $\sum_{i=1}^{n} (y_i - \beta_1 x_i) x_i = 0$, which yields the OLS estimator of β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

The estimated regression line is $\hat{y}_i = \hat{\beta}_1 x_i$ which has no intercept term. As such, this regression line must pass through the origin.

Special Case 2

For the specification without any regressor: $y_i = \beta_0 + u_i$, the LS criterion function is:

$$Q_n(\beta_0) = \sum_{i=1}^{n} (y_i - \beta_0)^2,$$

and the FOC is: $\sum_{i=1}^{n} (y_i - \beta_0) = 0$. The resulting OLS estimator of β_0 is

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}.$$

This shows that the sample average gives the "best" fit of y_i in the LS sense, when there is no other information. In this case, the estimated regression line is $\hat{y}_i = \bar{y}$, a horizontal line over x.

Other Minimization Programs

- Given the linear specification, LS minimization yields the "best" fit of data in the sense that the sum of squared errors is the smallest. This is not the only way to find the best fit of data, however.
- There are other minimization programs for data fitting. For example, minimizing the sum of absolute errors:

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|,$$

yields the least absolute deviation (LAD) estimators for β_0 and β_1. Different criterion functions lead to different fits of data and hence different regression lines.

A special case: For the specification $y_i = \beta_0 + u_i$, the LAD criterion function is:

$$\sum_{i=1}^{n} |y_i - \beta_0| = \sum_{i:y_i>\beta_0} (y_i - \beta_0) - \sum_{i:y_i<\beta_0} (y_i - \beta_0).$$

The FOC is $-\sum_{i:y_i>\beta_0} 1 + \sum_{i:y_i<\beta_0} 1 = 0$, so that the solution satisfies:

$$(\text{the number of } y_i > \beta_0) = (\text{the number of } y_i < \beta_0).$$

That is, the LAD estimator of β_0 is the sample median of y_i. Thus, the LAD regression line in effect describes the median behavior of y_i conditional on x_i. This result also shows that meidan (0.5 quantile) may be computed via an optimization program.

Algebraic Properties

- Plugging $\hat{\beta}_0$ and $\hat{\beta}_1$ into the FOC: $\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$, we have:

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^n \hat{u}_i = 0.$$

That is, the positive and negative residuals must cancel out. Also,

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = \sum_{i=1}^n \hat{u}_i x_i = 0,$$

which suggests that the sample covariance between x_i and $\hat{u}_i$ is zero.

- As $\sum_{i=1}^n \hat{u}_i = 0$, we can see:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i = \hat{\beta}_0 + \hat{\beta}_1 \left(\frac{1}{n}\sum_{i=1}^n x_i\right) + \frac{1}{n}\sum_{i=1}^n \hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 \bar{x},$$

which shows the estimated regression line must pass through $(\bar{x}, \bar{y})$.

Remarks

- These properties are algebraic consequences of the OLS method and hence "algebraic properties". These properties hold without any statistical assumptions on data.
- For the specification without the intercept: $y_i = \beta_1 x_i + u_i$, the residuals are such that

$$\sum_{i=1}^{n} x_i \hat{u}_i = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 x_i) = 0,$$

by the FOC. Yet, the sum of residuals,

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i),$$

is not necessarily zero.

Goodness of Fit

It is easy to verify that

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SST} = \sum_{i=1}^{n}(\hat{u}_i + \hat{y}_i - \bar{y})^2 = \underbrace{\sum_{i=1}^{n}\hat{u}_i^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SSE},$$

where SST denotes total sum of squares, SSR residual sum of squares, and SSE explained sum of squares. The goodness of fit of the estimated regression line is measured by the coefficient of determination, which is defined as the proportion of the total variation (SST) of y_i due to the variation of $\hat{y}_i$ (SSE):

$$R^2 = \text{SSE}/\text{SST} = 1 - \text{SSR}/\text{SST}.$$

Clearly, the larger the R^2, the better the regression line fits the data.

- $0 \leq R^2 \leq 1$: R^2 is based on sum of squares and hence cannot be negative. Also, as SSE $\leq$ SST, R^2 cannot exceed one.
- Extreme cases
  - $R^2 = 0$ if SSE is zero. This happens when $\hat{y}_i = \bar{y}$ for all i; that is, the estimated regression line is the horizontal line $\bar{y}$, and x has no explanatory ability at all.
  - $R^2 = 1$ if SSR is zero. This happens when there is perfect fit of the regression line; that is, the estimated regression line passes through all data points.
- R^2 from different regressions are comparable only when these regressions are for the same dependent variable y (so that they have the same SST but with different regressor x).

Example: Simple Wage Regressions

Taiwan's estimated wage models based on 2010 male data (11561 obs):

$$\widehat{\log(\text{wage})} = 4.5929 + 0.0494\ \text{educ}, \quad R^2 = 0.133$$

$$\widehat{\log(\text{wage})} = 5.1208 + 0.0059\ \text{exper}, \quad R^2 = 0.026$$

where educ and exper denote, respectively, the years of education and working experience. Note that the slope coefficient is

$$\frac{\mathrm{d}\log(\text{wage})}{\mathrm{d}x} = \frac{1}{\text{wage}}\frac{\mathrm{d}(\text{wage})}{\mathrm{d}x},$$

which is the predicted percentage change of wage when x changes by one unit. Thus, the first regression line predicts 5% wage increase for a male with one more year of education, and the second regression line predicts only 0.6% wage increase for a male with one more year of experience.

Classical Assumption I

The random variables y_i, $i = 1, \ldots, n$, are such that:

(i) $\mathbb{E}(y_i) = b_0 + b_1 x_i$ for some b_0 and b_1, where x_i are non-random;

(ii) $\text{var}(y_i) = \sigma_o^2$, $\text{cov}(y_i, y_j) = 0$ for $i \neq j$.

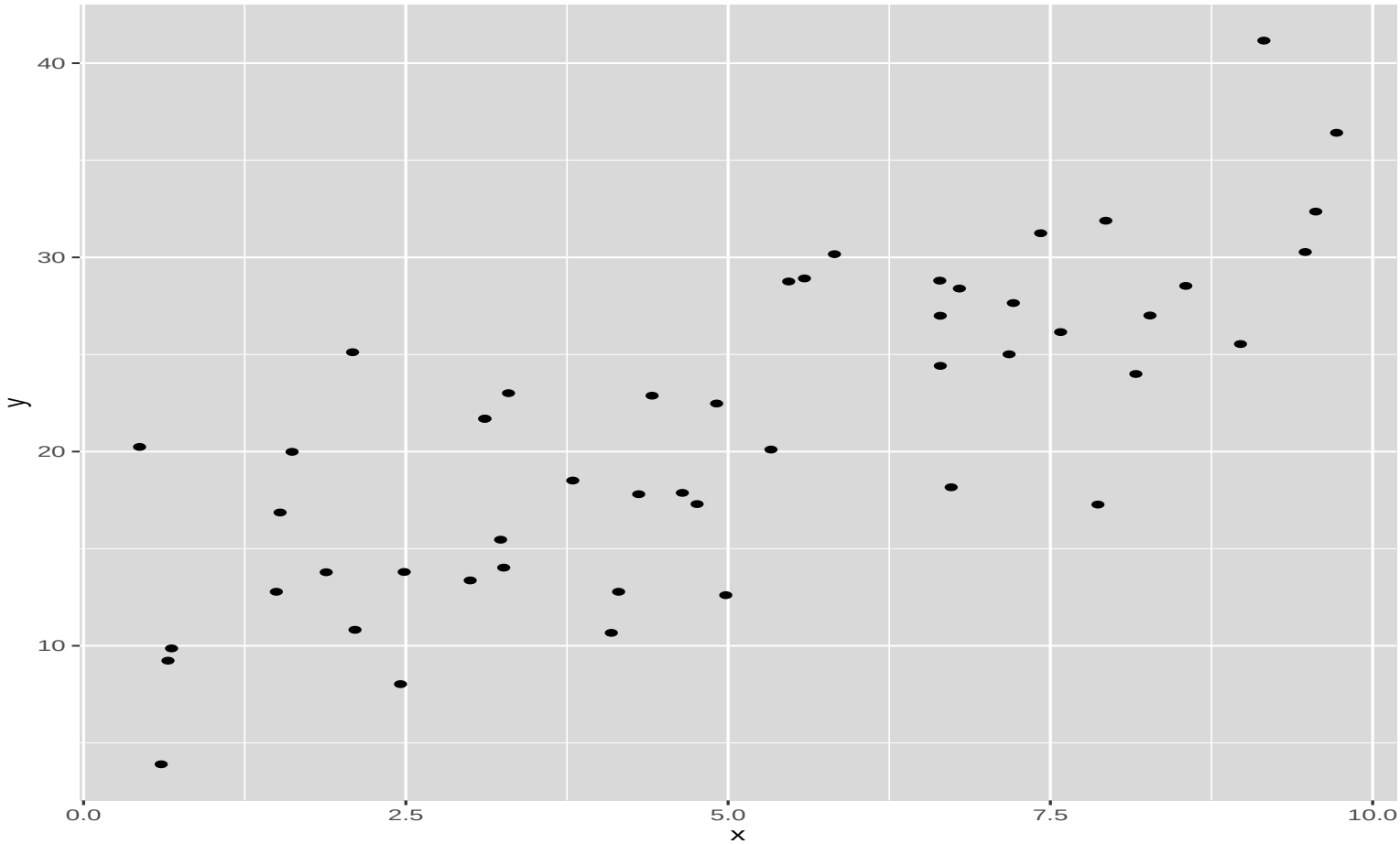
Letting ε_i denote the error $u_i(b_0, b_1)$, this assumption is equivalent to:

$$y_i = b_0 + b_1 x_i + \varepsilon_i,$$

with x_i non-random, $\mathbb{E}(\varepsilon_i) = 0$, $\text{var}(\varepsilon_i) = \sigma_o^2$, and $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.

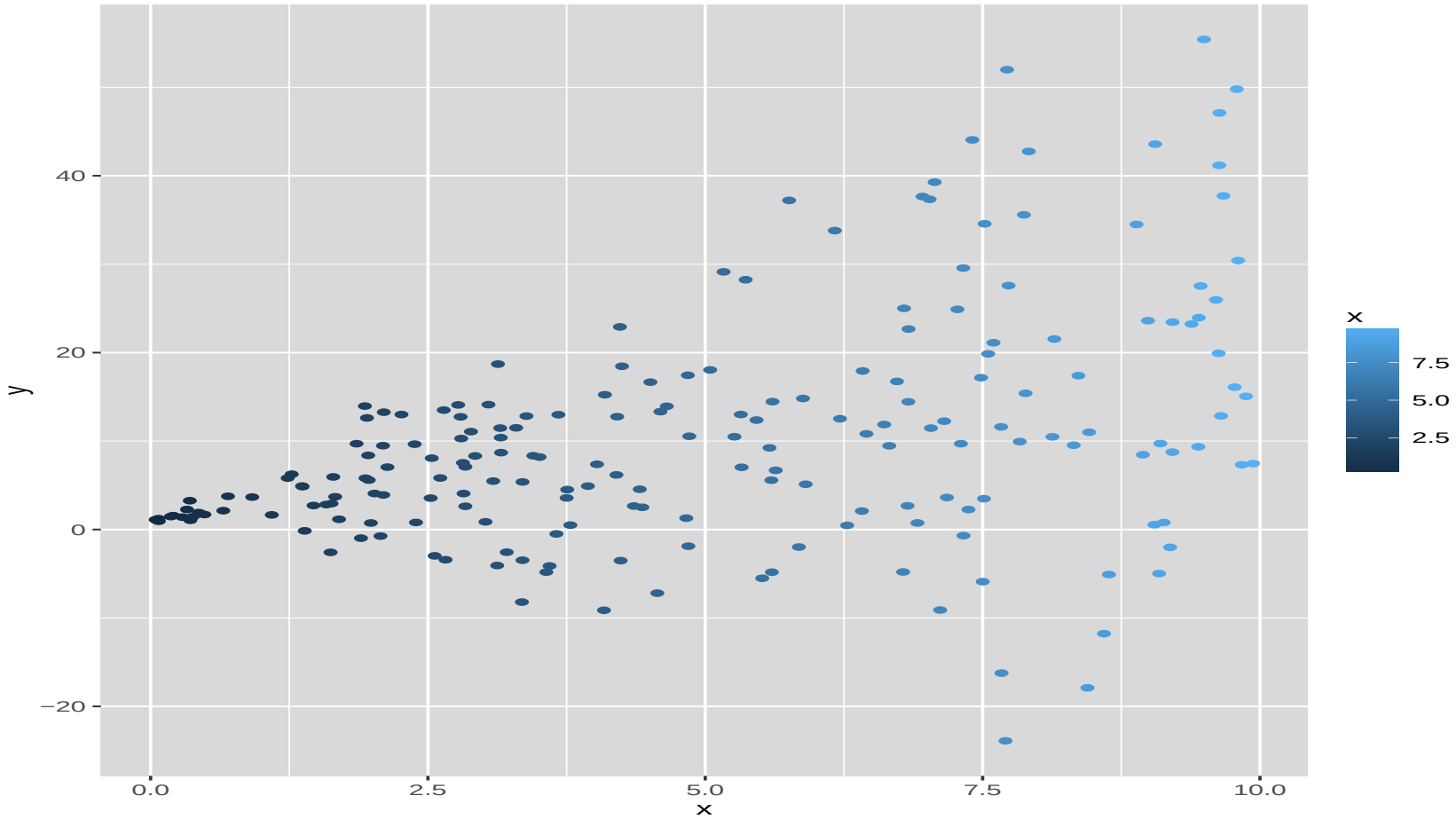
- The assumption (i) means the linear function, $\beta_0 + \beta_1 x$, is the correct specification for the mean function.
- The assumption (ii) requires y_i to be uncorrelated and have constant variance (homoskedasticity); when y_i have unequal variances, they are said to exhibit heteroskedasticity. As $\text{var}(y_i) = \text{var}(\varepsilon_i)$, σ_o^2 is also known as the error variance.

Example of Conditionally Homoskedastic Data



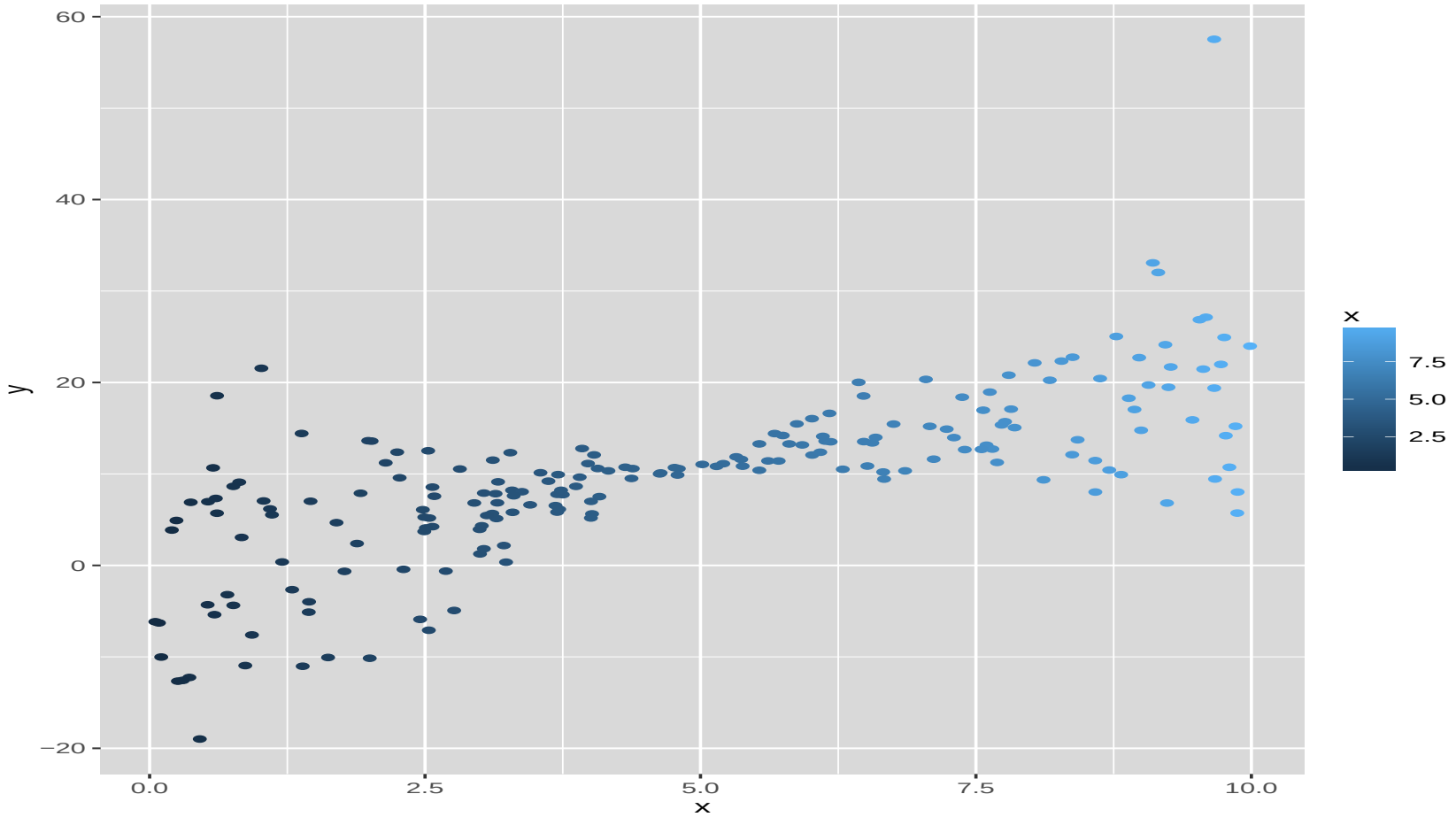
Simulated data: y with homogeneous variation across x

Example of Conditionally Heteroskedastic Data



Simulated data: y with heterogeneous variation across x

Example of Conditionally Heteroskedastic Data



Simulated data: y with heterogeneous variation across x

Unbiasedness of the OLS Estimators

Under Classical Assumption I(i), $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for, respectively, b_0 and b_1.

Note that $\sum_{i=1}^n (x_i - \bar{x}) = 0$, and hence $\bar{x} \sum_{i=1}^n (x_i - \bar{x}) = 0$,

$$\sum_{i=1}^n x_i (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})^2.$$

As x_i are non-random, we have

$$\mathbb{E}(\hat{\beta}_1) = \frac{\sum_{i=1}^n \mathbb{E}(y_i)(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (b_0 + b_1 x_i)(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = b_1.$$

Given $\mathbb{E}(y_i) = b_0 + b_1 x_i$, we have $\mathbb{E}(\bar{y}) = b_0 + b_1 \bar{x}$ and

$$\mathbb{E}(\hat{\beta}_0) = \mathbb{E}(\bar{y} - \hat{\beta}_1 \bar{x}) = b_0 + b_1 \bar{x} - \mathbb{E}(\hat{\beta}_1)\bar{x} = b_0.$$

Variance of the OLS Estimators

Under Classical Assumption I(i) and (ii),

$$\text{var}(\hat{\beta}_1) = \sigma_o^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \text{var}(\hat{\beta}_0) = \sigma_o^2 \frac{\sum_{i=1}^n x_i^2/n}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

It is easy to verify that

$$\begin{aligned}\text{var}(\hat{\beta}_1) &= \frac{\sum_{i=1}^n \text{var}(y_i)(x_i - \bar{x})^2}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2} = \sigma_o^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2} \\ &= \sigma_o^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}.\end{aligned}$$

We omit the details of deriving $\text{var}(\hat{\beta}_0)$.

Remark: $\text{var}(\hat{\beta}_1)$ would be smaller if x_i are more disperse (about $\bar{x}$). In this case, the estimated regression line is more stable and would not be affected much by a few x_i.

The OLS estimator of σ_o^2 is an average of squared residuals:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2,$$

and $\hat{\sigma}$ is also known as regression standard error. Note that the sum of squared residuals is divided by $n-2$, rather than n, because $\hat{u}_i$ must satisfy two FOCs of LS estimation and hence lose two degrees of freedom. The result below shows that $\hat{\sigma}^2$ is also an unbiased estimator for σ_o^2; the proof is omitted.

Unbiasedness of $\hat{\sigma}^2$

Under Classical Assumption I(i) and (ii), $\mathbb{E}(\hat{\sigma}^2) = \sigma_o^2$.

Replacing σ_o^2 in $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_0)$ with $\hat{\sigma}^2$, we obtain the following variance estimators:

$$\widehat{\text{var}(\hat{\beta}_1)} = \hat{\sigma}^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2},$$
$$\widehat{\text{var}(\hat{\beta}_0)} = \hat{\sigma}^2 \frac{\sum_{i=1}^n x_i^2/n}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

which are also unbiased for, respectively, $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_1)$. Their square roots are the standard errors for $\hat{\beta}_1$ and $\hat{\beta}_0$.

Example: Simple Wage Regressions

The estimated wage model based on Taiwan's 2010 male data (11561 obs): The dependent variable is log(wage), and the estimated parameters are:

$$\begin{array}{lll} 4.5929 & + 0.0494\,\text{educ}, & R^2 = 0.133, \\ (0.0156) & (0.0012) & \hat{\sigma} = 0.3971 \\ 5.1208 & + 0.0059\,\text{exper}, & R^2 = 0.026, \\ (0.0073) & (0.0003) & \hat{\sigma} = 0.4208 \end{array}$$

where the numbers in the parentheses are the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. These results show that the parameter estimates are quite precise because they have very small standard errors.