

ECON3150/4150 Spring 2015

Lecture 3&4 - The linear regression model

Siv-Elisabeth Skjelbred

University of Oslo

January 29, 2015

- Chapter 4 in S&W
- Section 17.1 in S&W (extended OLS assumptions)

Overview

In this lecture we will:

- Clear up some things from previous lecture
- Start on the linear regression model with one regressor

Standard deviation and standard error

- (σ) - Population standard deviation, the true standard deviation in the population.
- Sample standard deviation: (s) An estimator of the population standard deviation.
- s_y is the estimate of the population standard deviation for the random variable Y of the population from which the sample was drawn.
- Standard error of an estimator: An estimator of the standard deviation of the estimator.
- $SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}} = s_y / \sqrt{n}$ is the standard error of sample mean, which is an estimator of the standard deviation of \bar{Y} .
- The sample mean is an estimator of the population mean.

Standard deviation and standard error

Population parameter	Sample statistic
μ = Population mean	\bar{Y} = Sample estimate of population mean
σ = Population standard deviation	s = sample standard deviation, estimator of σ
$\sigma_{\bar{Y}}$ = Standard deviation of \bar{Y}	$SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}}$ = Standard error of \bar{Y} estimator of $\sigma_{\bar{Y}}$.

Standard error

Standard error

$$SE(\bar{Y}) = \frac{s_y}{\sqrt{n}}$$

- The standard error of the sample average says something about the uncertainty around the estimate of the mean.
- It is an estimate of how far the sample mean is likely to be from the population mean.
- The standard error falls as the sample size increases as the extent of chance variation is reduced.
- The standard error is used to indicate uncertainty around the estimate.

Sample standard deviation

Sample standard deviation

$$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}$$

- The sample standard deviation is the degree to which individuals within the sample differs from the sample mean.
- The sample standard deviation will tend to the population standard deviation (σ_y) as the sample size increases.
- The sample standard deviation is used to describe how widely scattered the measurements are.

T-statistic

- A normally distributed variable (X) can be made standard normal by:

$$Z = \frac{X - \mu}{\sigma/\sqrt{n}}$$

- In practice we rarely know the population standard deviation required to calculate the standard normal variable Z . The alternative is to calculate the T variable:

$$T = \frac{X - \mu}{s_x/\sqrt{n}}$$

- Note that the t-distribution also depends on the assumption that X is normal distributed.
- The sample average is normally distributed whenever:
 - X_i is normally distributed.
 - n is large (CLT).

The simple linear regression model

Definition of the simple linear regression model

Goals of regression models:

- "Estimate how X effects Y "
- "Explain Y in terms of X "
- "Study how Y varies with changes in X "

For example:

Explained (y)	Explanatory (x)
Wages	Education
Grade	Hours of study
Smoke consumption	Cigarette tax
Crop Yield	Fertilizer

Can we write this in an econometric model?

The econometric model

Econometric model

An equation relating the dependent variable to a set of explanatory variables and unobserved disturbances, where unknown population parameters determine the ceteris paribus effect of each explanatory variable.

The econometric model must:

- Allow for other factors than X to affect Y
- Specify a functional relationship between X and Y
- Captures a ceteris paribus effect of X on Y

The simple linear regression model

The simple linear regression model can in general form be written as:

$$Y = \beta_0 + \beta_1 X + u$$

- It is also called the bivariate linear regression model.
- The econometric model specifying the relationship between Y and X is typically referred to as the population regression line
- u : is the error term (some books use e or ϵ instead) and represents all factors other than X that affects Y .
- β_0 : Population constant term/intercept.
- β_1 : Population slope parameter, the change in Y associated with a one unit change in X .

The simple linear regression model

If we increase X by Δ then:

Before:

$$Y = \beta_0 + \beta_1 X + u$$

After:

$$Y + \Delta Y = \beta_0 + \beta_1(X + \Delta X) + (u + \Delta u)$$

Difference:

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

As long as Δu is zero β_1 measures the effect of a unit change in X on Y .

Example

If we are interested in the effect of education on wages the model is:

$$\text{wages} = \beta_0 + \beta_1 \text{years of education} + \text{other factors}$$

here β_1 measure the ceteris paribus effect (holding all other factors constant) of one more year of education, that is:

$$\beta_1 = \frac{\text{change in wages}}{\text{change in education}} = \frac{\Delta \text{wages}}{\Delta \text{education}}$$

Linear in parameters

$$Y = \beta_0 + \beta_1 X + u$$

The SLRM is linear in parameters (β_0 and β_1).

- The SLRM is named linear due to being linear in parameters.
- Linear in parameters simply means that the different parameters appear as multiplicative factors in each term.
- The above model is also linear in variables, but this does not need to be the case.
- In chapter 5 we will cover when X is a binary variable.
- In chapter 8 we will cover X and Y being natural logarithms as well as other functional forms of X .
- In chapter 11 we cover Y being binary

Terminology

The variables X and Y have several different names that are used interchangeably:

Left side (Y)	Right side (X)
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

Terminology

- The intercept β_0 and the slope β_1 are the **coefficients** or **parameters** of the population regression line.
- Another name for the population regression line is the population regression function.

Estimating the simple linear regression model

- The simple linear regression model is a model where the dependent variable is a linear function of a single independent variable, plus an error term.
- It aims at describing the relationship between the dependent variable and the independent variable in a population.
- But how do we estimate this line?

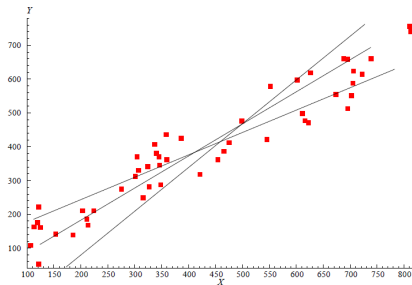
Estimating the simple linear regression model

- Need a sample of size n from the population
- Model: $y_i = \beta_0 + \beta_1 x_i + u_i$ where i is observation i
- u_i is the error term for observation i

Ordinary Least Squares

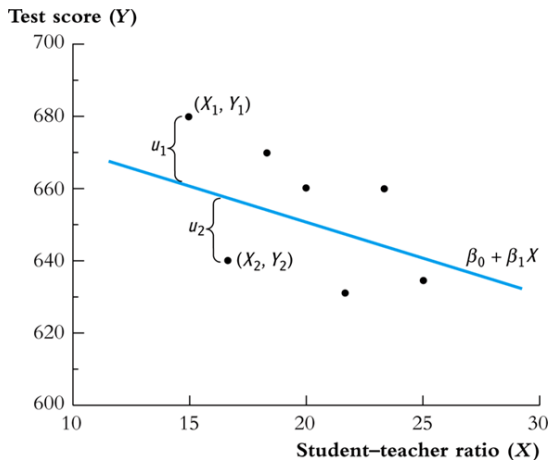
- Ordinary Least Squares (OLS) is a method for estimating the unknown parameters in a linear regression model.
- The method is to minimize the sum of squared errors.
- The OLS estimator chooses the regression coefficients so that the estimated regression line is as close as possible to the observed data.
- Under the assumptions that we will discuss later OLS is the most efficient estimator of the linear population regression function.

Ordinary Least Squares



- OLS finds the line that is closest to the observed data by minimizing the squared errors.

Ordinary Least Squares



- The error is the vertical distance between the regression line and the observation
- The value given by the regression line is the predicted value of Y_i given X_i .

Deriving the Ordinary Least Squares

- Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be some estimators of β_0 and β_1 .
- These estimators predict Y_i and the prediction mistake (\hat{u}_i) is defined as $Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$.
- The sum of squared prediction mistakes over all n observations is:

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Deriving the Ordinary Least Squares

OLS finds the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of the squared prediction mistakes (the OLS estimator).

- 1 Use calculus to obtain the following first order conditions for this minimization problem.

$$-2 \sum_{i=0}^n (Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_i) = 0 \quad (1)$$

$$-2 \sum_{i=0}^n X_i (Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_i) = 0 \quad (2)$$

- ② Solve the first order conditions for the unknown $\hat{\beta}_0$ and $\hat{\beta}_1$

The resulting OLS estimates of β_0 and β_1 are

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (3)$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (4)$$

Properties of OLS

- Given $\hat{\beta}_0$ and $\hat{\beta}_1$ we can obtain the predicted value \hat{Y}_i for each observation.
- By definition each predicted value is on the OLS regression line.
- The OLS residuals (\hat{u}_i) is the difference between Y_i and its predicted value.
- If \hat{u}_i is positive(negative) the line underpredicts (overpredicts) Y_i .
- The ideal case is $\hat{u}_i = 0$, but in most cases this is not true.

Properties of OLS

By the definition of \hat{u}_i and the first OLS first order condition the sum of the prediction error is zero:

$$\sum_{i=1}^n \hat{u}_i = 0 \quad (5)$$

The OLS residuals are chosen to make the residuals add up to zero.

Properties of OLS

By the second FOC:

$$\sum_{i=1}^n \hat{u}_i X_i = 0$$

The sample covariance between the independent variable and the OLS residuals is zero.

Properties of OLS

- The point (\bar{X}, \bar{Y}) is always on the regression line

Estimating OLS - example

A sample of ACT scores (American College testing) and GPA (Grade point average) for eight college students

Student	GPA	ACT
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25
7	2.7	25
8	3.7	30

Calculate:

$$\hat{GPA} = \hat{\beta}_0 + \hat{\beta}_1 ACT$$

Estimating OLS - example

Remember that:

$$\hat{\beta}_0 = \bar{Y} + \hat{\beta}_1 \bar{X}$$

$$ACT = \bar{x} = \frac{1}{n} \sum_i^n x_i = 25.875$$

$$GPA = \bar{y} = \frac{1}{n} \sum_i^n u_i = 3.2125$$

$$\hat{\beta}_0 \approx 3.2125 + \hat{\beta}_1 * 25.875$$

Estimating OLS - example

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Obs	GPA	ACT	1)	2)	1*2	$(x_i - \bar{x})^2$
			$(x_i - \bar{x})$	$(y_i - \bar{y})$		
1	2.8	21	-4.875	-0.4125	2.01	23.77
2	3.4	24	-1.875	0.1875	-0.35	3.52
3	3.0	26	0.125	-0.2125	-0.03	0.02
4	3.5	27	1.125	0.2875	0.32	1.27
5	3.6	29	3.125	0.3875	1.21	9.77
6	3.0	25	-0.875	-0.2125	0.19	0.77
7	2.7	25	-0.875	-0.5125	0.45	0.77
8	3.7	30	4.125	0.4875	2.01	17.02
Av.	3.2125	25.875		Sum:	5.8125	56.875

Estimating OLS - example

$$A\bar{C}T = \bar{x} = \frac{1}{n} \sum_i^n x_i = 25.875$$

$$G\bar{P}A = \bar{y} = \frac{1}{n} \sum_i^n u_i = 3.2125$$

$$\hat{\beta}_1 = 5.8125/56.875 \approx 0.1022$$

$$\hat{\beta}_0 \approx 3.2125 - (0.1022) * 25.875 \approx 0.5681$$

$$G\hat{P}A = 0.5681 + 0.1022A\hat{C}T$$

Interpretation of example

$$\hat{GPA} = 0.5681 + 0.1022ACT$$

- A person that has a one unit higher ACT score than another is predicted to have approximately 0.1 higher GPA.

Another interpretation example

Given a sample of 526 individuals the following OLS regression line can be estimated:

$$\widehat{wage} = -0.90 + 0.54educ$$

- Intercept indicates that a person with zero education pay 90 cents per hour to work.
- The regression line does poor at low levels of education because only 18 people have less than 8 years of education.
- β_1 indicates that one more year of education increase hourly wage by 54 cent.

Linear regression

You may have different goals with the regression. The goal can be:

- To describe the data in a scatterplot and no substantive meaning is attached to the regression line.
- To make forecasts, or predictions, of the value of Y for an entity not in the data set, for which we know the value of X .
- To estimate the causal effect on Y of a change in X .

Causality

In this course we will take a practical approach to defining causality:

Causality

A causal effect is defined to be the effect measured in an ideal randomized controlled experiment.

Ideal Randomized Controlled experiment

- Ideal: subjects all follow the treatment protocol.
- Randomized: subjects from the population of interest are randomly assigned to a treatment or control group (so there are no confounding factors).
- Controlled: Having a control group permits measuring the differential effect of the treatment.
- Experiment: The treatment is assigned as part of the experiment, subjects have no choice, so there is no "reverse" causality in which subjects choose the treatment they think will work best.

More about this will come in lecture 17.

Underlying assumptions of OLS

The OLS estimator is unbiased, consistent and has asymptotically normal sampling distribution if:

- 1 Random sampling.
- 2 Large outliers are unlikely.
- 3 The conditional mean of u_i given X_i is zero.

Random sample

- As covered extensively in the lecture 2, the observations in the sample must be i.i.d.
- We will address the failure of random sampling assumption under time-series analysis.

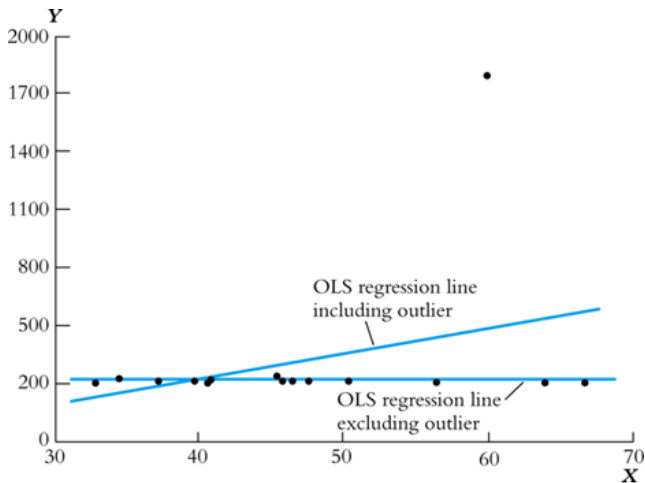
Outliers

- Large outliers are unlikely when X_i and Y_i have finite fourth moments.
item Outliers can arise due to:
 - Data entry errors.
 - Sampling from a small population where some members of the population are very different from the rest.

Outliers

- Outlying observations can provide important information by increasing the variation in explanatory variables which reduce standard errors.
- However, including them can dramatically change the estimated slope.
- When dealing with outliers one may want to report the OLS regression both with and without the outliers.

Outliers



Zero conditional mean

Assumptions

- 1 For simplicity we assume that $E(u) = 0$.
- 2 The average value of u does not depend on the value of X
 $E(u|X) = E(u)$.

Combining the two assumptions gives the zero conditional mean assumption $E(u|X) = 0$

Zero conditional mean

Example:

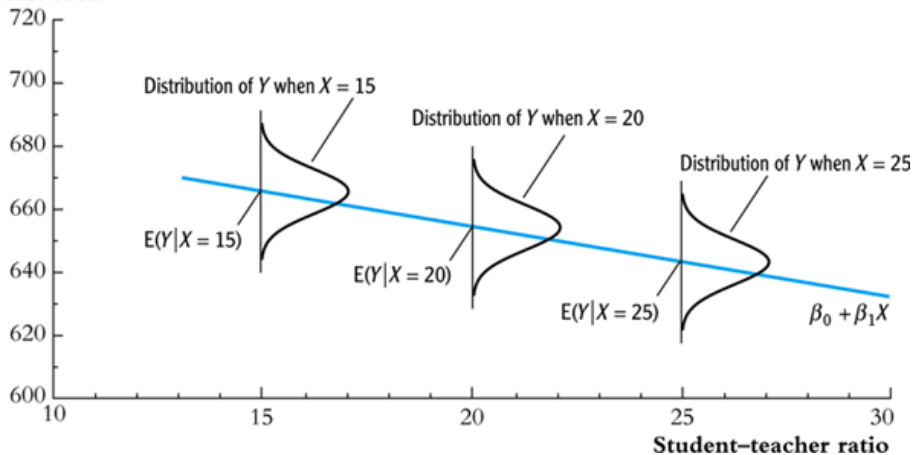
$$\text{wages} = \beta_0 + \beta_1 \text{educ} + u$$

- Ability is one of the elements in u .
- The zero conditional mean requires for example $E(\text{abil}|\text{educ} = 8) = E(\text{abil}|\text{educ} = 16)$.
- The average ability level must be the same for all education levels for the assumption to hold.

Zero conditional mean assumption

The conditional distribution of u_i given X_i has a mean of zero. I.e. the factors contained in u_i are unrelated to X_i

Test score



Efficiency of OLS

OLS is the most efficient (the one with the lowest variance) among all linear unbiased estimators whenever:

- The three OLS assumptions hold AND
- The error is homoskedastic.

Measures of fit

How well does the OLS regression line fit the data?

- The regression line can be divided into two elements: We can write that $y_i = \hat{y}_i + \hat{u}_i$
 - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$ is the systematic part - the part of Y explained by X.
 - \hat{u}_i is the unsystematic part, the part of Y not explained by X.

Total sum of squares

Total sum of squares

$$TSS = SST \equiv \sum_{i=0}^n (y_i - \bar{y})^2$$

- Total sum of squares, the sum of squared deviations of Y_i from its average
- SST is a measure of the total sample variation in the Y_i , that is it measures how spread out the Y_i are in the sample.

Explained sum of squares

Explained sum of squares

$$ESS \equiv \sum_{i=0}^n (\hat{y}_i - \bar{y})^2$$

- Explained sum of squares is the sum of squared deviations of the predicted value from its average.
- It measures the sample variation in \hat{Y}_i

Sum of squared residuals

Sum of squared residuals

$$SSR \equiv \sum_{i=0}^n \hat{u}_i^2$$

- Sum of squared residuals measures the sample variation in \hat{u}

Terminology

There is no uniform agreement about the abbreviations or the names of the measures of fit:

- Some textbooks denote total sum of squares SST and not TSS
- Some textbooks use SSE instead of ESS
- Stata used model sum of squares to denote ESS.
- Some say regression sum of squares for explained sum of squares.
- Some say error sum of squares instead of sum of squared residuals.

Decomposition of TSS

TSS

$$TSS = ESS + SSR$$

R-squared

- The regression R^2 is the fraction of the sample variance of Y_i explained by X_i .

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

- The R^2 ranges between 0 and 1
- $R^2 = 0$ - none of the variation in Y_i is explained by X_i
- $R^2 = 1$ - all the variation is explained by X_i , all the data points lie on the OLS line.
- A high R^2 means that the regressor is good at predicting Y_i (not necessarily the same as a "good" regression)

Standard error of the regression

The standard error of the regression (SER) is an estimator of the standard deviation of the regression error u_i . It measures the spread of the observation around the regression line.

$$SER = s_{\hat{u}} \text{ where } s_{\hat{u}}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{n-2}$$

Example

```
1 . reg wage educ
```

Source	SS	df	MS
Model	1179.73204	1	1179.73204
Residual	5980.68225	524	11.4135158
Total	7160.41429	525	13.6388844

Number of obs = 526
F(1, 524) = 103.36
Prob > F = 0.0000
R-squared = 0.1648
Adj R-squared = 0.1632
Root MSE = 3.3784

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.5413593	.053248	10.17	0.000	.4367534	.6459651
_cons	-.9048516	.6849678	-1.32	0.187	-2.250472	.4407687

Interpretation of coefficients

The interpretation of the coefficients depend on the unit of measurement.

- In the wage education example the unit of measurement for education is years of education while for wage it is dollar.
- Thus one more year of education is predicted to give \$0.54 higher wages.

Change the unit of measurement

- If the dependent variable is multiplied by the constant c , the OLS intercept and slope estimates are also multiplied by c .
- If the independent variable is multiplied by some nonzero constant c , then the OLS slope coefficient is divided by c .

Unbiasedness of OLS

- It can be shown that $E(\hat{\beta}_1) = \beta_1$ (appendix 4.3) thus β_1 is unbiased.
- This means that the sampling distribution of the estimator is centered about the value of the population parameter.
- Note: this is a property of the **estimator** and says nothing about whether an estimate for a given sample is equal to the population parameter.
- The crucial assumption is that $E(u_i|X_i) = 0$ which we will come back to in next lecture.

Variance of the OLS estimators

- How far can we expect $\hat{\beta}_1$ to be away from β_1 on average?
- The variance is necessary to choose the most efficient estimator.

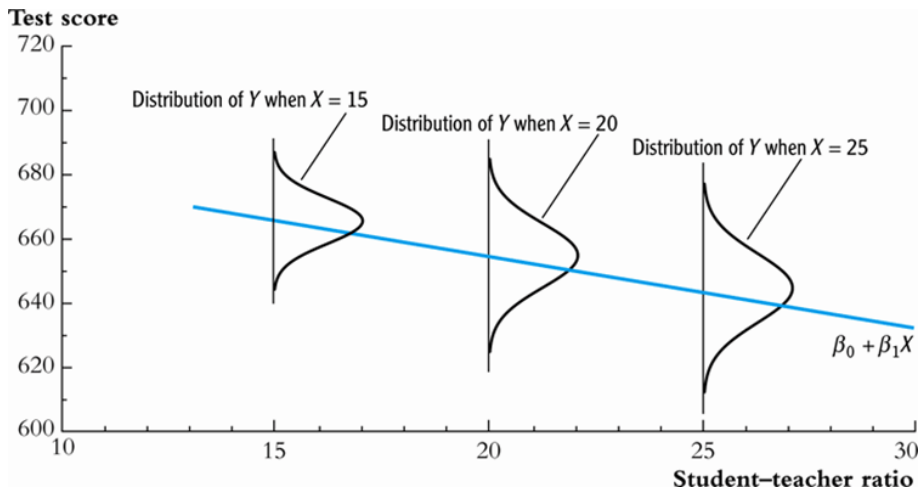
Homoskedasticity

Homoskedasticity

The error u has the same variance given any value of the explanatory variable, in other words: $\text{Var}(u|x) = \sigma^2$

- Homoskedasticity is not required for unbiased estimates.
- But it is an underlying assumption in the standard variance calculation of the parameters.
- To make the variance expression easy the assumption that the errors are homoskedastic are added.

Homoskedasticity



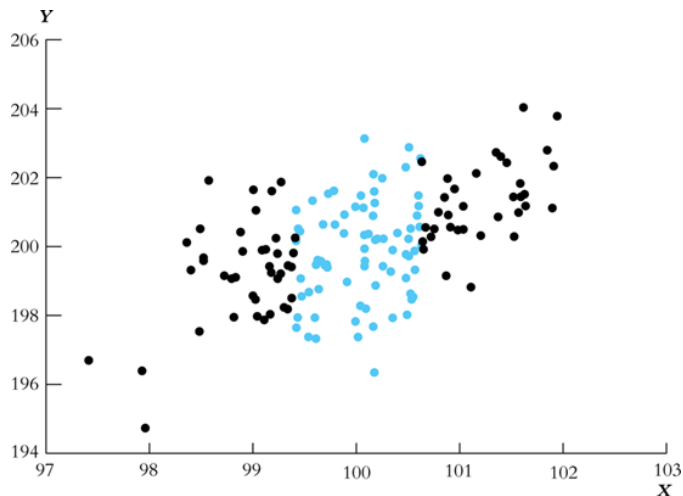
The figure illustrates a situation where the errors are heteroskedastic, the variance of the error increases with X .

Variance of the OLS estimators

It can be shown (appendix 4.3) that:

$$\text{Var}(\hat{\beta}_1) = \frac{1}{n} \frac{\text{var}[(X_i - \mu_x)u_i]}{[\text{var}(X_i)]^2} = \frac{\text{var}[(X_i - \mu_x)u_i]}{\sum_{i=1}^n (x_i - \bar{X})^2}$$

Variance of the OLS estimators

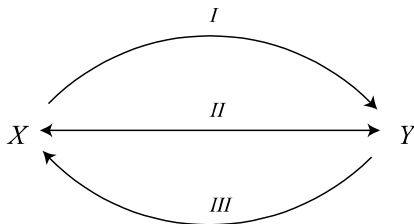


The larger the variance of X , the smaller the variance of $\hat{\beta}_1$.

Note on caution

- At this point you should be careful of reading too much into the regressions we do.
- We have not covered the grounds for establishing causal effects yet.

Regression and causality



Three possible theoretical causal relationships between X and Y .

- The regression is causal if I is true and II and III are not true.
- II is joint causality.
- III is reversed causality.
- Spurious correlation occurs if a third variable causes both Y and X .

Next week:

- Hypothesis testing - how to test if the slope is zero
- Confidence intervals for the slope