ECON3150/4150 Spring 2015

Lecture 3 - The linear regression model

Siv-Elisabeth Skjelbred

University of Oslo

Last updated: January 20, 2016

- Sections 3.3-3.5
- Chapter 4 in S&W
- Section 17.1 in S&W (extended OLS assumptions)

Overview

These lecture slides covers:

- Test statistics
- Confidence intervals
- Means comparison
- Introduction to the linear regression model with one regressor

Hypothesis testing

Steps in hypothesis testing:

1. Choose a desired significance level.
2. Perform a hypothesis test.
    a) Compute the test statistic
    b) Identify the critical value of the test-statistic

Test statistic

In order to test a null hypothesis against an alternative we need to choose a test statistic.

- A test statistic is a single measure of some attribute of a sample used in statistical hypothesis testing.
- The test statistic should quantify behavior, within the sample, that would distinguish the null from the alternative.
- The computed test statistic is compared to a critical value.

The critical value

- The critical value is a cutoff value, if the test statistic is more extreme than the critical value, then the null hypothesis is rejected.
- If the test statistic is not as extreme as the critical value we fail to reject the null.
- The critical value is defined by the area under the probability density function.

The test statistic

- For a normally distributed variable the test statistic is given by:

$$Z = \frac{Y - \mu_Y}{\sigma_Y}$$

- And it can compared to the critical value found in the normal distribution table.
- It requires the population distribution of Y.

The test statistic

Sample variance

The sample variance is an unbiased and consistent estimate of the population variance as long as the observations are i.i.d. and large outliers are unlikely. $(E(Y^4) < \infty)$

- The sample variance is an estimator for the population variance:

$$s_Y^2 = \hat{\sigma}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \text{"sample variance of Y"}$$

- The standard error is the estimator for the standard deviation:

$$SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}} = \frac{s_Y}{\sqrt{n}}$$

The t-statistic

T-statistic of sample average

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})} = \frac{\bar{Y} - \mu_{Y,0}}{s_y/\sqrt{n}}$$

- The t-statistic is t-distributed whenever Y is normally distributed.
- The t-statistic has heavier tails than the normal distribution.

Large sample distribution of the t-statistic

- When n is large s_Y^2 is close to σ_Y^2 with high probability.
- Thus the distribution of the t-statistic is well approximated by the standard normal distribution. (CLT)
- Thus under the null hypothesis t is approximately distributed N(0,1) for large n.

T-test for a population mean

Using the t-statistic for hypothesis testing:

1) Compute the t-statistic (t^{act})
2) Compute the degrees of freedom (v), which is n-1
3) Look up the critical value of your desired significance level (t^c) (Table 2, page 805)
4) Reject the null hypothesis if:
    - Two sided test: $|t^{act}| > t^c_{\alpha/2,v}$
    - Right-tailed test ($H_1 : \mu_y > \mu_{y0}$): $t > t^c_{\alpha,v}$
    - Left-tailed test ($H_1 : \mu_y < \mu_{y0}$): $t < -t^c_{\alpha,v}$

Note: Two-sided $t_{0.05,v}$ equals the one sided $t_{0.025,v}$

Example t-test

200 college graduates are asked about their wage. Mean wage in the sample is \$ 22.64 and the sample standard deviation is \$ 18.14. Is this evidence for or against the hypothesis that college graduates earn on average \$ 20 an hour?

$2.06 > 1.96$ the null hypothesis is rejected at a 5% significance level.

degrees of freedom $(n - 1)$	5% t-distribution critical value
∞	1.96

The p-value

P-value

The p-value is the probability of obtaining a test statistic, by random sampling variation, at least as adverse to the null hypothesis value as is the statistic actually observed, assuming that the null hypothesis is correct.

- The probability that we would observe a statistic at least as large as the sample average computed if the null hypothesis is true.
- The smaller the p-value the more unlikely it is to obtain the calculated statistic by random sampling if the null hypothesis is true.
- Assuming that the null is true you would obtain the a difference at least as large as the one observed in p% of studies due to random sampling error.

The p-value

Let $\bar{Y}^{act}$ denote the value of the sample average actually computed in the data set at hand then:

$$\text{p-value} = Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$$

When $\bar{Y}$ is (approximately) normally distributed we can standardize it using: $Z = \frac{X-\mu}{\sigma}$ which gives:

$$\begin{aligned}\text{p-value} =& Pr_{H_0}\left(|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}|\right) \\ =& 2\phi\left(-|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}|\right)\end{aligned}$$

where ϕ is the standard normal cumulative distribution function.

P-value for population mean

P-value when distribution is unknown

$$p - value = Pr_{H_0}(|t| > |t^{act}|) = 2\phi(-|t^{act}|)$$

$$\begin{aligned}
\text{p-value} &= Pr_{H_0}\left(|\frac{\bar{Y} - \mu_{Y,0}}{\hat{\sigma}_{\bar{Y}}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\hat{\sigma}_{\bar{Y}}}|\right) \\
&\cong 2\phi\left(-|\frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})}|\right) \\
&= Pr_{H_0}\left(|\frac{\bar{Y} - \mu_{Y,0}}{\frac{s_Y}{\sqrt{n}}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\frac{s_Y}{\sqrt{n}}}|\right)
\end{aligned}$$

$\cong$ probability under normal tails.

- When n is large t is approximately distributed N(0,1) (CLT) thus the distribution of the t-statistic is approximately the same as $(\bar{Y} - \mu_{Y,0})/\sigma_{\bar{Y}}$.

P-value and T-statistics

Looking at the formula you should recognize:

$$\frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})} = t$$

as the usual t-statistics. Thus the p-value is: $Pr_{H_0}[|t| > |t^{act}|]$. And due to the central limit theorem t is approximately distributed N(0,1) for large n.

P-value for population mean

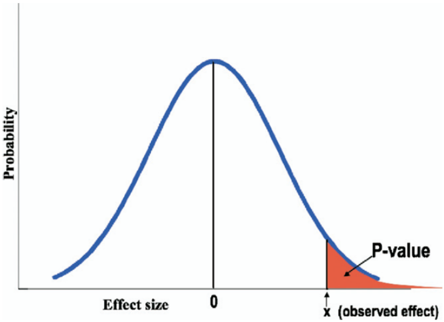


Figure: Graphical depiction of the definition of a (one-sided) p-value. The curve represents the probability of every observed outcome under the null hypothesis. The p-value is the probability of the observed outcome (x) plus all "more extreme" outcomes, represented by the shaded "tail area".

Rejection rules

Reject the null hypothesis if:

- If $|t^{act}| > t^c$
- If p-value $<$ desired significance level

What significance level?

When n is small

- The p-value calculations conducted is based on the assumption that the statistic is approximately normal (CLT and large n).
- When n is small the standard normal distribution can be a poor approximation to the distribution of the t-statistic.
- The exact distribution of the t-statistic depends on the distribution of Y and it can be very complicated.
- If the population distribution is normally distributed the student t distribution can be used for hypothesis testing.
- However, it is rare that economic variables are normally distributed.

Comparing means from two populations

Examples of questions one may ask:

- Are white applicants more likely to be called in for a job interview than African Americans?
- Do men earn more than women?
- Do people with a college degree earn more than those without?

The answer to all these questions involve comparing means of two different population distributions.

Comparing means from two populations

- Two types of tests for whether two sample means are the same
  - Unpaired: We have two separate sets of independent and identically distributed samples. T-test compares the means of the two groups of data to tests whether the two groups are statistically different.
  - Paired: A sample of matched pairs of similar units or one group of units that has been tested twice. The two measurements generally are before and after a treatment intervention. The test is calculated based on the difference between the two sets of paired observations.
- Both assume that the analyzed data is from a normal distribution.

The method chosen also requires to you to specify the relationship between the variance of the two samples.

- Pooled variance: the variance for the first population is about the same as that of the other population.
- Separate variance: The variances are unequal.

Comparing means from two populations

Let m denote men and w denote women. The null hypothesis is that men and women in the population we investigate have the same mean earnings, i.e. $d_0 = 0$

$$H_0 : \mu_m - \mu_w = d_0 \text{ v.s. } H_1 : \mu_m - \mu_w \neq d_0$$

- Estimate the means: $\bar{Y}_m - \bar{Y}_w$ is an estimator for $\mu_m - \mu_w$
- Calculate the standard error

$$SE(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}} \text{(due to CLT, two independent RNV)}$$

- Calculate t-statistic, p-value or confidence interval: $t = \frac{\bar{Y}_m - \bar{Y}_w - d_0}{SE(\bar{Y}_m - \bar{Y}_w)}$

Comparing means in Stata

Using the auto.dta data-set for independent samples and assuming unequal variance:

```
. ttest price, by(foreign) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Domestic |      52    6072.423    429.4911    3097.104    5210.184    6934.662
 Foreign |      22    6384.682    558.9942    2621.915     5222.19    7547.174
---------+--------------------------------------------------------------------
combined |      74    6165.257    342.8719    2949.496    5481.914      6848.6
---------+--------------------------------------------------------------------
    diff |           -312.2587    704.9376               -1730.856    1106.339
------------------------------------------------------------------------------
    diff = mean( Domestic) - mean( Foreign)                      t =  -0.4430
Ho: diff = 0                     Satterthwaite's degrees of freedom =  46.4471

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.3299         Pr(|T| > |t|) = 0.6599          Pr(T > t) = 0.6701
```

Exercise

The scores of a random sample of 8 students on an econometrics test are as follows: 60,62,67,70,72,75,78.
Test to see if the sample mean is significantly different from 65 at the 5% level. Report the t and p-values.

The simple linear regression model

Definition of the simple linear regression model

Goals of regression models:

- "Estimate how X affects 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

Simple linear regression

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.

Simple linear regression

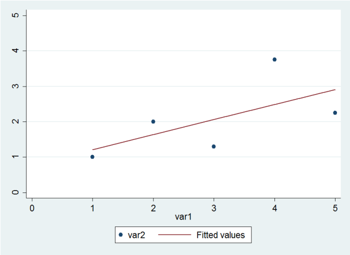
The variables X and Y have several different names that are used interchangeably:

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

Simple linear regression

In simple linear regressions, the predictions of Y when plotted as a function of X form a straight line.

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

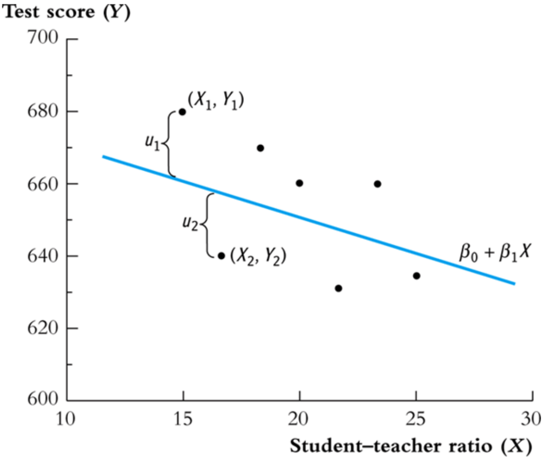



```
scatter var2 var1 , xlabel(0(1)5) ylabel(0(1)5) || lfit var2 var1
```

Simple linear regression

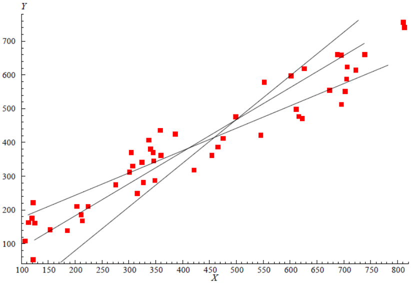
- Linear regression consists of finding the best-fitting straight line through the points.
- The best-fitting line is called a regression line.
- The best fitting line is the regression line and consists of the predicted score on Y for each possible value of X.
- The best fitted line is the one that minimizes the sum of squared errors.

Errors



- The error is the horizontal distance between the regression line and the observation
- The value given by the regression line is the predicted value of Y given X.

Errors



- Which line is closest to the observed data?

Estimating the simple linear regression model

Model:

$$Y_i = \beta_0 + \beta_1 x_i + u_i$$

- Need a sample of size n from the population.
- i is observation number i.
- u_i is the error term for observation i.
- β_0 is the intercept.
- β_1 is the slope parameter.

Ordinary Least Squares

- The method of finding the "best fitted line" by minimizing the sum of squared errors is called Ordinary Least Squares (OLS).
- The OLS estimator chooses the regression coefficients so that the estimated regression line is as close as possible to the observed data.
- OLS thus estimates the unknown parameters β_0 and β_1 assuming a linear regression model.
- Under the assumptions that we will discuss later OLS is the most efficient estimator of the linear population regression function.

Assumptions

- Random sample.
- Large outliers are unlikely.
- Zero conditional mean.
- Linear in parameters.

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.

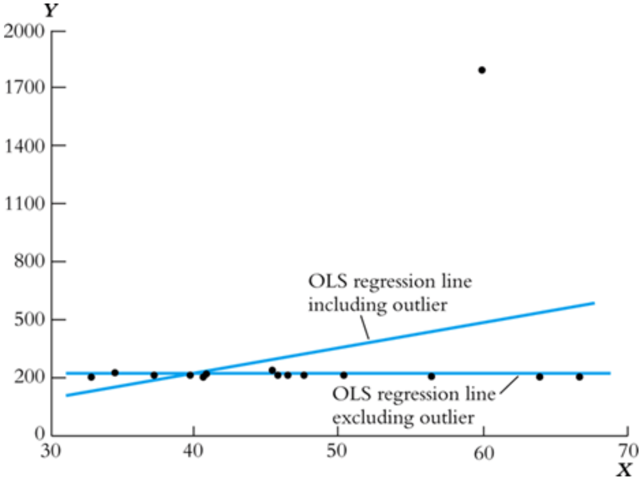
Outliers

- An outlier is an observation with large residuals.
- Large outliers are unlikely when X_i and u_i have final fourth moments.
- 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. (sample peculiarity)

OLS and outliers

The least squares method is not robust to outliers, one or several observations can have undue influence on the results.

- Conclusions that hinge on one or two data points must be considered extremely fragile and possible misleading.
- May be an idea to run the regression both with and without the outliers.
- In the presence of outliers that do not come from the same data generating process as the rest of the data OLS may be biased an inefficient.


Y
2000
1700
1400
1100
800
500
200
0
30
40
50
60
70
X
OLS regression line
including outlier
OLS regression line
excluding outlier

Zero conditional mean

1. $E(u) = 0$ The expected value of the error term is zero.
2. $E(u|x) = E(u)$ The expected value of the error term is independent of X.

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

Zero conditional mean

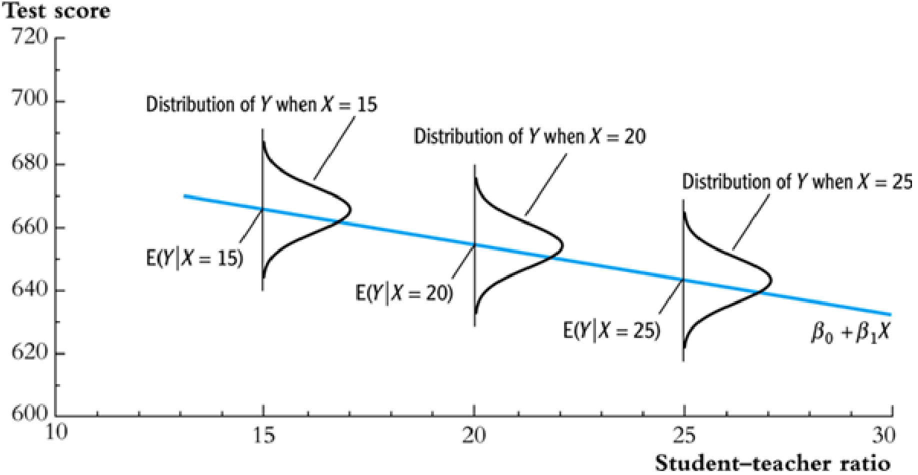
Example:

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

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

Zero conditional mean

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



Zero conditional mean example

The relationship between class attendance and grades can be modeled as:

$$grade_i = \beta_0 + \beta_1 Attend_i + u_i$$

The key is that u contains all the variables other than Attend that help determine your grade.
For the ZCM assumption to hold we need:

$$E(u|Attend = 19) = E(u|Attend = 5)$$

to hold.

- Can you list some of the variables in u?
- Is it likely that the ZCM holds?

Linear in parameters

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

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

- 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.

Homoskedasticity

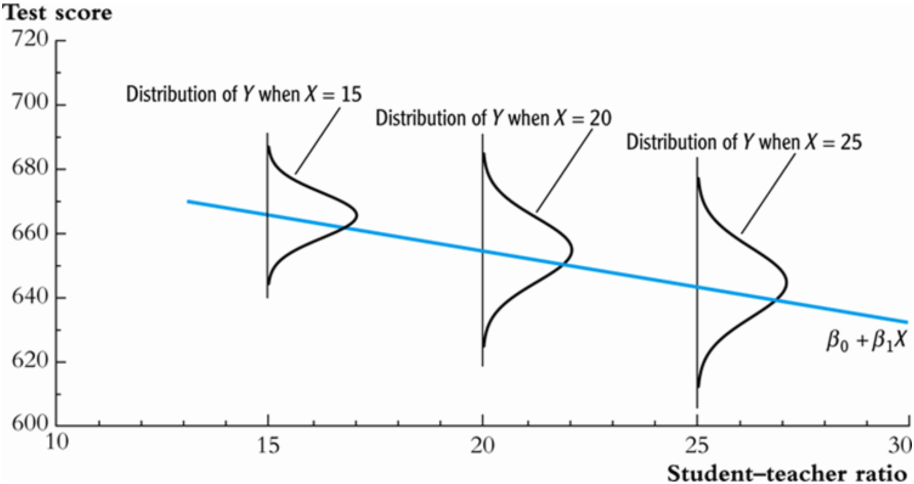
Standard OLS requires that errors are homoskedastic:

Homoskedasticity

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

- 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.
- If errors are not homoskedastic they are heteroskedastic.

Heteroskedasticity



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

Heteroskedasticity

What do we do:

- Run OLS but correct the standard errors.
- Run something other than OLS.

Summary

We have learned that:

- How to standardize a normally distributed variable.
- That the t-statistic is necessary when the population standard deviation is unknown.
- The sample average is normally distributed whenever:
  - X_i is normally distributed.
  - n is large (CLT).
- Means comparison
- The assumptions of OLS

Summary

The classical approach to testing hypothesis is:

- Choose a significance level, the convention is 5% and find the critical value.
  - The null hypothesis is rejected if the absolute value is less than the critical value (two sided test)
  - The null hypothesis is rejected if the p-value is smaller than the desired significance level.
- If the null hypothesis is true the statistic will lie within the two critical values (positive and negative value) with 100*(1-α)% of the time of random samples.