Quantitative Methods 2

Quantitative Methods 2 (University of Melbourne)

Quantitative Methods 2

What is Statistics? (L1)

Statistics

How to use data to learn about phenomena.

Descriptive Statistics

Methods of organizing, summarizing and presenting data in ways that are useful, attractive and informative to the reader.

Inferential Statistics

Methods used to draw conclusions about a population based on information provided by a sample of the population.

Population – The set of all item of interest

Parameter – A descriptive measure of a population

Sample – A set of data drawn from the studied population

Statistic – A descriptive measure of a sample

Confidential level – The degree of certainty we have that our interval contains the value of the parameter

Significance level – The relative frequency of a wrong conclusion

Estimate – An approximate value of a parameter based on a sample statistic

Variable – Any characteristic of a population or sample

Data – Observations of the variable

Target population – The population about which we want to draw inferences

Sampled population – The actual population from which the sample has been drawn

Simple random sample – One in which each element of the population has an equal chance of appearing

Cluster sample – Choose groups or clusters at random from the population and take a census from these groups

Range – The difference between largest and smallest observations

Variance – A measure of variability of a numerical data set

Deviation – Difference between an observation and the mean of the set of data it belongs to

Variable Types

1. Quantitative (numerical or interval) – numerical data observes real numbers

   e.g. income, student marks, prices...etc

2. Qualitative (categorical or discrete or nominal) – nominal data observes categorical or qualitative

   e.g. marital status: married=1, single=0, divorced=2

3. Ordinal (ranked) – ordinal data observes ordered qualitative data

e.g. poor=1, fair=2, good=3, very good=4, excellent=5

Data Types

1. Quantitative (numerical or interval)

– values are real numbers

- all calculations are valid

- data may be treated as ordinal or nominal

2. Qualitative (categorical or discrete or nominal)

– values are the arbitrary numbers that represent categories

- only calculations based on the frequencies of occurrence are valid

- data may not be treated as ordinal or numerical

3. Ordinal (ranked)

– values must represent the ranked order of the data

- calculations based on an ordering process are valid (averages are often misleading)

- data may be treated as nominal but not as numerical

Sampling and Non-Sampling Errors

1. Sampling error:

-difference between statistic and parameter due to random process

-a legitimate difference / expected

-depends on sample size

-trade off sample size (cost)

2. Non-Sampling error:

-more serious

-regardless of sample size

-due to bad sampling design, methods

-errors in data acquisition (recording of incorrect responses)

-non-response error (when responses are not obtained from some members of the sample)

-selection bias (when some members of the target population cannot possibly be selected for inclusion in the sample)

Graphical Description

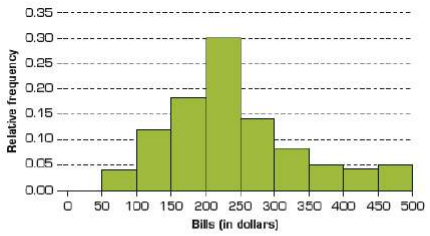
1. Histogram:

-a good way to visualize how the data is distributed

-shows how a variable is distributed

-data is divided into groups and the relative frequency of data for each group is calculated

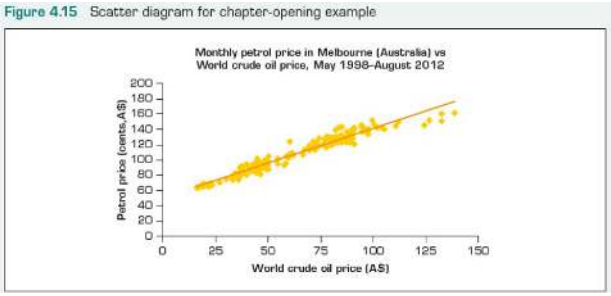
4.2 Relative frequency histogram of electricity bills with equal class width



2. Scatter plots:

-a good way to show the relationship between two variables

Figure 4.15 Scatter diagram for chapter-opening example



- Linear relationship: one in which two variables move proportionately

- Positive relationship: a relationship where the variables move in same directions

- Negative relationship: a relationship where the variables move in opposite directions to each other

3. Bar chart / Pie chart:

-often used for qualitative (categorical) variables in vertical bars or subdivided circle sectors

4. Line charts:

-often used for showing trends in a time series variable

Data Types

1. Cross Section

   Data measured across a population (or a sample) at one point in time

   e.g. a sample of personal characteristics of UNIMELB students' at year 2012

2. Time Series

   Data measured on the same variable at different points of time

   e.g. annually, quarterly, monthly or even daily, most macroeconomics and financial data

   1) governments want to know future values of interest rates, unemployment rates

   2) housing industry economists must forecast mortgage interest rates or demand for housing

3. Panel (Combination of time series & cross section)

Data is on the sample of individuals over several time periods

e.g. HILDA

Measures of Central Location

1. Mean:

Population mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Sample mean

$$\bar{x} = \frac{x_1 + x_2 \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. Median

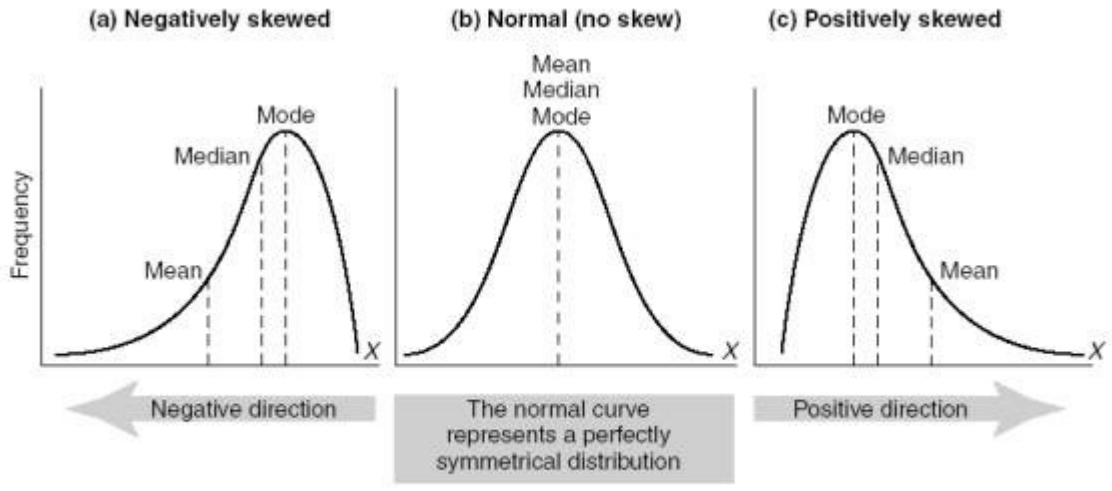
- middle value, if n is odd
- average of the two middle values, if n is even
- median may be preferred if the data is highly skewed, or with ordinal data

3. Mode

- the most frequently occurring value in a set of data

Relationship between mean, median and mode

1. Skewness: The degree to which a graph differs from a symmetric graph
2. Skewed to the right: positively skewed
3. Skewed to the left: negatively skewed



Dispersion Measures

Sample Variance

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

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Sample **standard** deviation: $s = \sqrt{s^2}$

Population **standard** deviation: $\sigma = \sqrt{\sigma^2}$

Interpretation f **SD**

$$P(\mu - 2\sigma \le x \le \mu + 2\sigma) \sim 0.95$$

under normal approximation

Question

It is well known that the logarithm of income in the population is well approximated by a normal distribution. Let $x = \ln(Income)$ and $\mu_x = 8 \;\&\; \sigma_x = 0.5$

Find the income range for the 95% of population around the mean

$$\mu_x - 2\sigma_x \le x \le \mu_x + 2\sigma_x \Rightarrow 7 \le x \le 9$$
$$\Rightarrow \exp(7) \le Income \le \exp(9)$$
$$\Rightarrow 1096 \le Income \le 8103$$

Approximately, 95% of people belong to the income range 1096 to 8100.

Coefficient of Variation

Sample

$$CV = \frac{s}{\bar{x}}$$

Population

$$CV = \frac{\sigma}{\mu}$$

Questions

e.g. Are rates of return for Trust A and Trust B is 27% and 15%. Standard deviation are 16.74% and 9.97%.

1) decision made based on better average rate of return, then Trust A

2) decision made based on the basis of risk (SD), then Trust B

3) decision made based on both higher return and lower risk, then Trust A

$CV_A = \frac{16.74}{27} = 0.62$

$CV_B = \frac{9.97}{15} = 0.665$

Review of Statistical Inference (L2)

Random Variable (rv) – A function that assigns a numerical value to each simple event in a sample space

Expected value – The sum of all possible values a random variable can take times the corresponding probabilities

Discrete random variable – A random variable that can assume only a countable number of values (finite or infinite)

e.g. the number of telephone calls received in a given hour, the number of customers served at a hotel on a given day

Continuous random variable – A random variable that can assume an uncountable number of values (any value in interval) *e.g. the exact winning time men's 100m dash Olympic 2016*

Best estimator – consistency, unbiased and minimum variance

Probability Density Function (pdf)

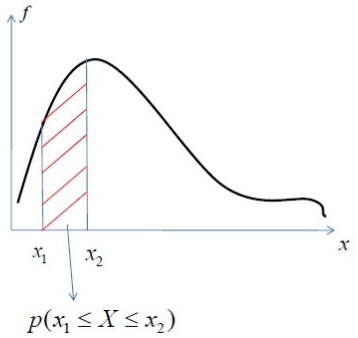
- A function f(x) such that:

1) f(x) is non-negative

2) the total area under f(x) is 1

3) the area under f(x) between the line x=a and x=b gives the probability that the value of x is between a and b

Probability Density Function for a continuous random variable



Sampling Distribution

- A relative frequency distribution of various values of the sample mean using a number of samples

- An estimator is a random variable because it can take different values over different samples

- Central Limit Theorem:

$$\bar{x} \xrightarrow{\text{in Distribution}} N\left(\mu, \frac{\sigma^2}{n}\right)$$

with increase in sample size n, distribution of sample mean ($\acute{X}$) becomes more and more like a *normal distribution* with mean equal to μ and variance equal to $\frac{\sigma^2}{n}$

- Usually sample size more than 30 is considered good enough for normal approximation

Normal Distribution

- The most important continuous distribution

- It is defined by two parameters μ and σ^2

- The curve is bell-shaped and symmetric around its mean

- Describes many phenomena which occur both in nature and in business

- Increasing μ shifts the curve to the right and decreasing μ shifts it to the left

- Larger value of σ widen the curve and smaller one narrow it

Standard Normal Distribution

- A normal distribution with μ=0 and σ^2=1

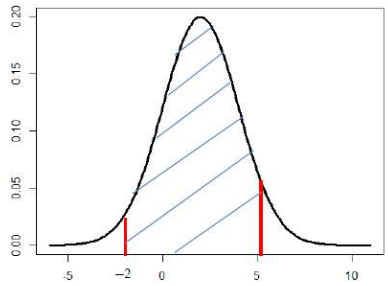
- The transformation $Z=\frac{X-\mu}{\sigma}$ makes things a lot easier since if X~N (μ,σ^2) then $Z=\frac{X-\mu}{\sigma}$ ~ N (0,1)

- Probabilities for standard normal distribution has been tabulated

- If μ and σ of a normally distributed random variable are known, this can always transform the probability statement about X into a probability statement about Z

A Question

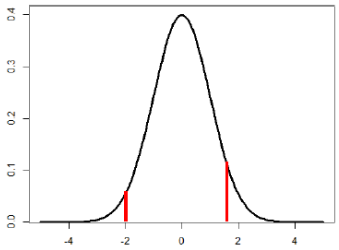
What is the probability of $-2 \le X \le 5$ when $X \sim N(2,4)$?



Calculating Probabilities

Perform $z_2 = \frac{x_2-\mu}{\sigma} = \frac{5-2}{2} = 1.5$

$z_1 = \frac{x_1-\mu}{\sigma} = \frac{-2-2}{2} = -2$



$$p(x_1 \le X \le x_2) = p(\frac{x_1-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{x_2-\mu}{\sigma}) = p(z_1 \le Z \le z_2)$$

$$p(-2 \le X \le 5) = p(Z \le 1.5) - p(Z \le -2) = 0.933 - 0.023 = 0.91$$

Hypothesis Testing

Step1:

Set up the null (H_o) and alternative (H_1) hypotheses for the question at hand.

Step2:

Determine the appropriate test statistic and its sampling distribution (criterion for decision).

Step3:

Specify the value of α (significance level).

Step4:

Define the decision rule (construct critical region, i.e. what values of the test statistic will reject H_o)

Step5:

Calculate the value of the test statistic

Step6:

Make a decision to answer the question (reject H_o or not), and interpret the result in light of the question in words.

Type I & II Errors

- When accepting or rejecting hypotheses, following cases might happen:

	H_0 is **true**	H_0 **is false**
Reject H_0	Type I Error P(Type I Error)=α Size of the test	Correct Decision P(Reject H_0 | H_0 false)=1-β Power of the test
Do not reject H_0	Correct Decision P(Correct Decision)=1-α	Type II Error P(Type II Error)=β

- The probability of making a Type II error increases as the probability of making a Type I error decreases

- The power of the test decreases as the level of significance is reduced

- The probability of making Type I error and α are the same

t-Distribution

- When σ^2 is unknown, sample variances s^2 is used which makes things slightly different

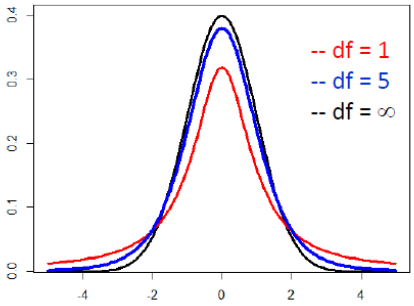
- It can be shown that $\frac{\acute{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ where t_{n-1} is a t-distribution with (n-1) degrees of freedom

- Like normal distribution there a table for t-distribution that gives the probabilities under the curve

- The t-distribution looks like a standard normal distribution which is symmetric around mean of zero, bell-shaped but has fatter tails

- The shape of the t-distribution depends on the degrees of freedom

- As the degrees of freedom become larger, the t-distribution approaches the standard normal distribution



Example

Are mean salaries from commerce Graduates below $50,000?

Sample information: $\acute{x}$ = $48,918, n = 50 and s = 6271

Solution

Step1: Null & Alternative hypothesis

H_o: $\mu = 50000$

H_A: $\mu < 50000$

Step2: Test statistics

$$\frac{\acute{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

Step3: Level of significance

α = 0.05

Step4: Decision Rule

Null is rejected if $t \leq t_{0.05,49} = -1.68$

Step5: Value of test statistics

$$t = \frac{\acute{x}-\mu}{s/\sqrt{n}} = \frac{48918-50000}{6271/\sqrt{50}} = -1.22$$

Step6: Conclusion

Since t > -1.68 do not reject null hypothesis

There is not sufficient evidence to conclude that mean salaries of commerce graduates are below $50,000

P-value

- The probability of observing a test statistic at least as extreme as the one computed given that the null hypothesis is true.

- The smallest value of α that would lead to rejection of the null hypothesis.

- Reject null hypothesis if p-value ≤α (level of significance)

e.g. for t-distribution example this implies: p-value = Pr(t_{49} ¿ -1.22) = 0.114

As the p-value is above the α of 0.05, again conclude to NOT reject the null hypothesis

Calculation of the p-value

If H_A: $\mu > \mu_0$: p-value = $P(Z > z_0)$ (right tail test)

If H_A: $\mu < \mu_0$: p-value = $P(Z < -z_0)$ (left tail test)

If H_A: $\mu \neq \mu_0$: p-value = $2P(Z > z_0)$ if $z_0 > 0$

= $2P(Z < -z_0)$ if $z_0 < 0$

Describing the p-value of a test

If ***p-value < 0.01***: overwhelming evidence (highly significant)

If ***p-value < 0.05***: strong evidence (significant)

If ***0.05 < p-value < 0.10***: weak evidence (not significant)

If ***p-value > 0.10***: no evidence (not significant)

- a small p-value indicates that there is ample evidence to support the alternative hypothesis

- a larger p-value indicates that there is a little evidence to support the alternative hypothesis

Testing two populations independent samples (L3)

- Dependent samples: if the values in one sample affect the value in other sample

- Independent samples: is the values in one sample reveal no information about those of the other sample

Comparing two populations

- Independent samples of quantitative (interval) variables

- Aim is to construct a test of the difference in population means

Population1 parameters: μ_1 and σ_1^2	**Population2** parameters: μ_2 and σ_2^2
Sample size n_1	Sample size n_2
Statistics: $\acute{x}_1$ and s_1^2	Statistics: $\acute{x}_2$ and s_2^2

Null & Alternative Hypothesis

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$ Two-Tailed Test

$\mu_1 - \mu_2 < 0$ Left-Tailed Test

$\mu_1 - \mu_2 > 0$ Right-Tailed Test

Known Population Variance Case

If σ_1^2 and σ_2^2 are known and both populations are normally distributed or sample sizes are big enough

Z-statistic

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Unknown Population Variance Case

If σ_1^2 and σ_2^2 are unknown but both populations are normally distributed or sample sizes are big enough

t-statistic (v degree of freedom)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_v \qquad \nu = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\left[\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}\right]}$$

Unknown but Equal Variance Case

If $\sigma_1^2 = \sigma_2^2$ are unknown but both populations are normally distributed or sample sizes are big enough

t-statistic (it combines both samples to produce a single estimate of the population variance)

s^2 is pooled variance estimate, the weighted average of two sample variances

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} \qquad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Correlation Coefficient (L9)

Coefficient of Correlation

- measures the strength and direction of the linear relationship between two numerical variables

- population coefficient of correlation is denoted by ρ:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \text{where} \quad \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- coefficient values range between -1 and 1

→ if **ρ = -1** perfect negative linear association (every point falls on a straight line)

→ if **ρ = 1** perfect positive linear association (every point falls on a straight line)

→ if **ρ = 0** no linear association

- sample coefficient of correlation:

$$r = \frac{s_{xy}}{s_x s_y} \quad \text{where} \quad s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

- t-test hypothesis testing of the coefficient of correlation (ρ) can be conducted to determine whether Y and X are linearly related

Testing Correlation

- test whether any linear relationship exists between two variables or not

The hypotheses are:

H_0: **ρ = 0** (no linear relationship)

H_A: **ρ ≠ 0** (a linear relationship exists)

H_A: **ρ > 0** (a positive linear relationship exists)

H_A: **ρ < 0** (a negative linear relationship exists)

Testing Correlation Coefficient

The test statistic is:

$$t = \frac{r - \rho}{s_r} \quad \text{where} \quad s_r = \sqrt{\frac{1-r^2}{n-2}}$$

$$\Rightarrow if\ \rho = 0 \Rightarrow t = r\sqrt{\frac{n-2}{1-r^2}}$$ with d.f = n – 2, provided variables are distributed as bivariate normal

Hypothesis Testing

Q: Are age and internet use linearly related?

Step1

H_o: ρ = 0

H_A: ρ ≠ 0

Step2:

If two series are bivariate normal quantitative variables, then t statistic.

Step3:

α = 0.05

Step4:

Recall n = 300, so reject null if t > $t_{\alpha/n,\ n-2} = t_{0.025,\ 298}$ = 1.972 or if t < - $t_{\alpha/n,\ n-2}$ = -1.972

Step5:

r = -0.75, so t = $\frac{r-\rho}{s_r} = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0.75-0}{\sqrt{\frac{1-(-0.75)^2}{300-2}}}$ = - 19.58

Step6:

Since t is well below the lower tail critical values, we reject the null hypothesis. The two variables are linearly related.

Spearman Ranked Correlation Coefficient (r_s)

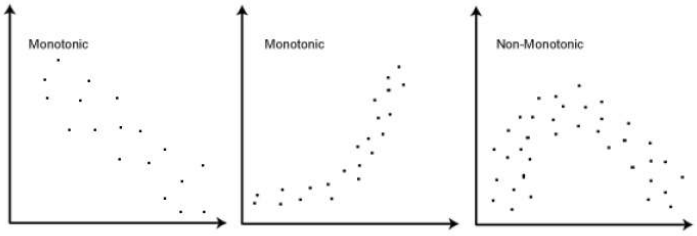
- constructed in the same manner as the Pearson Correlation Coefficient (r), bur uses the ranks of data instead of original numbers

- a measure of how two ordinal variables or two numerical but non-normal variables are related

$r_s = \frac{s_{ab}}{s_a\, s_b}$ where a and b refer to the data on x and y after it has been ranked

Monotonic Relationships

- always increasing or decreasing, it doesn't have to be linear



Spearman Rank Test

- test whether a relationship exists between variables in cases where

→ at least one variable is ranked, or

→ both variables are numerical but the normality requirement is not satisfied

- a nonparametric test based on ranking of data (test whether there is a relationship or not and sign of the relationship)

The hypotheses are:

H_0: **ρ_s = 0** (no monotonic relationship)

H_A: $\boldsymbol{\rho_s \neq 0}$ (a monotonic relationship exists)

H_A: $\boldsymbol{\rho_s > 0}$ (an increasing relationship exists)

H_A: $\boldsymbol{\rho_s < 0}$ (a decreasing relationship exists)

If $n \leq 30$:

Test statistic is simply r_s with exact critical values (under H_0 of $r_s = 0$) tabulated for one tailed tests in Critical values of the Spearman Rank Correlation Coefficient table. For two tailed tests – use critical values of α/2.

If $n > 30$:

Test statistic is

$$z = \frac{r_s - 0}{1/\sqrt{n-1}} = r_s\sqrt{n-1}$$ where z is approximately standard normal distributed

Hypothesis Testing

Q: Is a person's age and their likelihood of buying the new product mono?

Step1

H_o: $\rho = 0$

H_A: $\rho \neq 0$

Step2:

One of variables is ranked data so not normally distributed and sample size less than 30, so use the Spearman Rank Correlation Coefficient without normal approximation.

Step3:

$\alpha = 0.05$

Step4:

$n = 7$ and using α/2 for two-tailed test, Reject null if $r_s > 0.786$ or if $r_s < -0.786$

Step5:

From the table, $r_s = 0.655$

Step6:

Since $r_s < 0.786$ and $r_s > -0.786$, do not reject null hypothesis. There is no evidence of a relationship between age and the reported likelihood of buying a new product in this small sample.

Correlation does not Imply Causation

- if we estimate a correlation between two variables, X and Y, this just tells us that there is a statistical relationship

- we cannot say that necessarily X causes Y

- there are other reasons that we might see correlations:

a) Y causes X

b) X causes Y and Y causes X

c) There may be some other factor of factors that cause both X and Y

Correlation Vs Causation

1) **only Y causes X**

e.g. X = number of police in different regions

Y = crime rates in these regions

2) **X causes Y** and **Y causes X**

e.g. X = Quantity demanded for a product

Y = Price of the product

1) **a third factor might causes both X and Y**

e.g. X = Grades at the university

Y = Higher Salary later in jobs

Z = Higher IQ or better discipline

Simple Regression (L10)

Regression

Regression analysis – A technique that estimates the relationship between variables and aids forecasting

Simple linear regression – A regression equation with only one independent variable

$\mathbf{y = \beta_0 + \beta_1 x + \varepsilon}$

y: dependent variable, response variable, regressand, observable variables

x: independent variable, explanatory variable, regressor, observable variables

$\boldsymbol{\beta_0}$**:** y-intercept, regression parameters, regression coefficients

$\boldsymbol{\beta_1}$: slope of the line, regression parameters, regression coefficients

ε: unobserved random error

- ε (error term)

→ there are other factors that affect y but they are not include in the model

→ errors coming from linear approximation

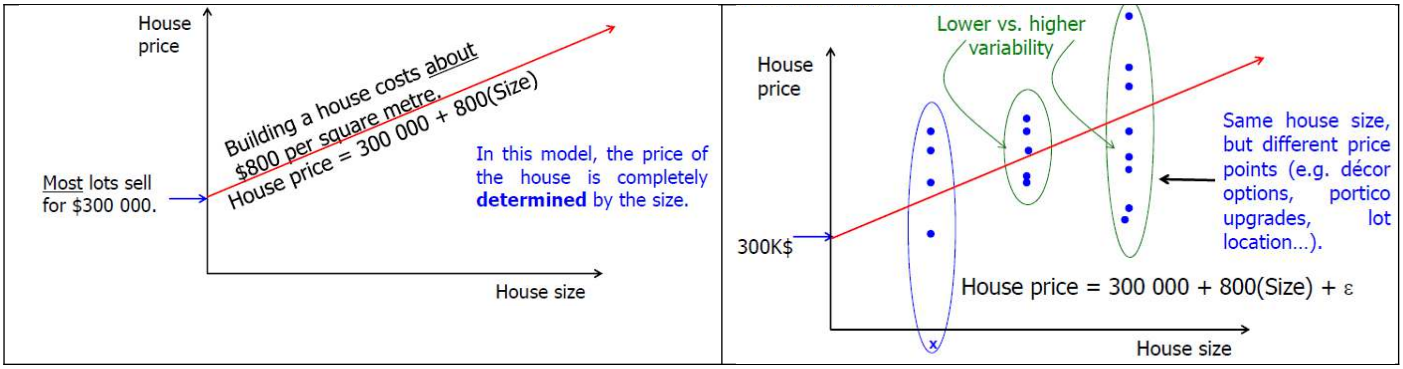
→ errors in measuring y

→ randomness in human behavior

→ $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$

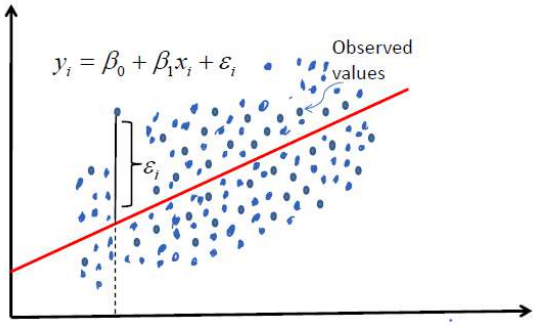
Examples: $y = 300000 + 800x + \varepsilon$

A model of the relationship between house size (independent variable) and house price (dependent variable) would be:	In real life, however, the house cost will vary even among the same size of house:



ε is the random error variable which is the difference between the actual selling price and the estimated price based on the size of the house. Its value will vary from house sale to house sale, even if the area of the house (x) remains the same due to other factors such as the location, age, décor etc of the house.

Regression from Population



Assumption:

- assumed model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

- the expected value of the error is zero $E(\varepsilon_i) = 0$ and

- same variance $Var(\varepsilon_i) = \sigma^2$

- uncorrelated errors $Cov(\varepsilon_i, \varepsilon_j) = 0$

- x is not correlated to ε

- (optional) $\varepsilon_i \sim N(0, \sigma^2)$

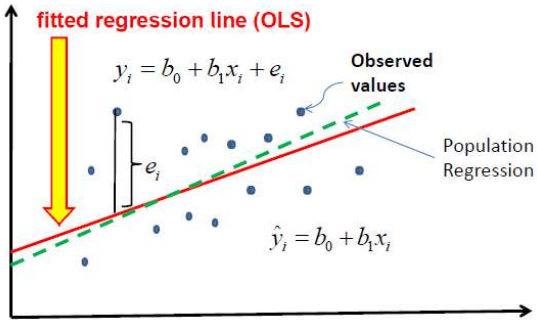
Regression from Sample

- main purpose in regression analysis is to learn about β_0 and β_1 for the population

- but what we usually have is data on x and y from a sample

- based on the samples, $y_i = b_0 + b_1 x_i + e_i \quad i = 1,\ldots,n$

- e_i is called residuals which is the differences between the predicted value of the dependent variable and its actual value



Least Square Principle

- aim is to find the best line through the observed sample, one way to do this is to find line with the least amount of total errors

- attempting to find b_0 and b_1 in a way that total difference between line and observed values of y is minimized

- defined fitted regression line as $\hat{y}_i$ **= b_0 + $b_1 x_i$** then error is **$e_i = y_i - (b_0 + b_1 x_i)$** where i = 1,...,n

Equations

Normal Equations

- we choose b_0 and b_1 to minimize the sum of squared errors:

$$Min_{b_0,b_1}\ S(b_0,b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

- first order conditions (F.O.C):

$$\frac{\partial S(b_0,b_1)}{\partial b_0} = 0$$
$$\frac{\partial S(b_0,b_1)}{\partial b_1} = 0$$

$$\begin{cases} \frac{\partial S}{\partial b_0} = -2\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)] = 0 \\ \frac{\partial S}{\partial b_1} = -2\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)] x_i = 0 \end{cases}$$

First Equations

- first equation:

$$\sum_{i=1}^{n} \underbrace{[y_i - (b_0 + b_1 x_i)]}_{e_i} = 0 \Rightarrow \sum_{i=1}^{n} e_i = 0$$

Point1: OLS residuals sum to zero

Point2: **$b_0 = \bar{y} - b_1 \bar{x}$**

$$\frac{1}{n}\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)] = 0 \Rightarrow \sum_{i=1}^{n} \frac{y_i}{n} - b_0 - b_1 \sum_{i=1}^{n} \frac{x_i}{n} = 0$$
$$\Rightarrow \bar{y} - b_0 - b_1 \bar{x} = 0 \Rightarrow b_0 = \bar{y} - b_1 \bar{x}$$

Second Equations

- second equation:

$$\sum_{i=1}^{n} \underbrace{[y_i - (b_0 + b_1 x_i)]}_{e_i} x_i = 0 \Rightarrow \sum_{i=1}^{n} e_i x_i = 0$$

- we can also write

$$\bar{x}\sum_{i=1}^{n} e_i = 0 \Rightarrow \sum_{i=1}^{n} e_i \bar{x} = 0$$

- therefore

$$\sum_{i=1}^{n} e_i x_i - \sum_{i=1}^{n} e_i \bar{x} = 0 \Rightarrow \sum_{i=1}^{n} e_i (x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)(x_i - \bar{x}) = 0$$

- by substituting $\mathbf{b_0 = \acute{y} - b_1 \acute{x}}$

$$\sum_{i=1}^{n} [y_i - (\bar{y} - b_1 \bar{x}) - b_1 x_i](x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n} [y_i - \bar{y} + b_1 \bar{x} - b_1 x_i](x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - \sum_{i=1}^{n} b_1 (x_i - \bar{x})(x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = b_1 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$b_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\mathbf{b_1} = \frac{S_{xy}}{S_{xx}}$$

OLS Estimator

- for deriving estimates for populations parameters, all we need to know OLS estimators (ordinary least squares)

$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

EViews Interpretation

Dependent Variable: INTERNET
Method: Least Squares
Date: 08/21/14 Time: 15:01
Sample: 1 300
Included observations: 300

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AGE	b_1 -0.435083	s_{b_1} 0.022218	-19.58260	0.0000
C	b_0 46.38335	s_{b_0} 1.201325	38.61014	0.0000

R-squared	0.562715	Mean dependent var	23.25000
Adjusted R-squared	0.561248	S.D. dependent var	5.708718
S.E. of regression	3.781362	Akaike info criterion	5.504690
Sum squared resid	4261.011	Schwarz criterion	5.529382
Log likelihood	-823.7035	Hannan-Quinn criter.	5.514571
F-statistic	383.4783	Durbin-Watson stat	1.931649
Prob(F-statistic)	0.000000		

S_E

Population model: $Internet_i = \beta_0 + \beta_1 AGE_i + \varepsilon_i$

Sample model: $Internet_i = b_0 + b_1 AGE_i + e_i$

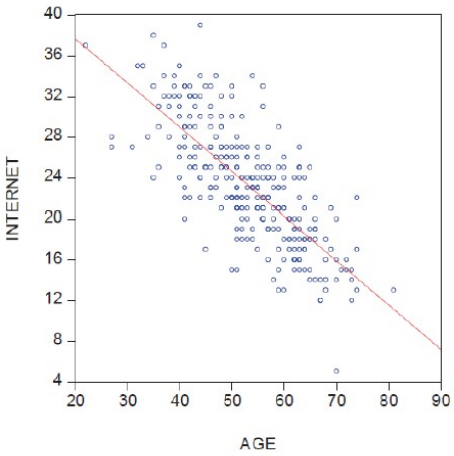
$\widehat{Internet}_i = \mathbf{b_0 + b_1 AGE_i}$

$\mathbf{= 46.38 - 0.435 AGE_i}$

b_0: y-intercept, where the line goes through the point where AGE = 0

b_1: the slope of the line, as age increases by one year, internet use falls by 0.453hours per week

Regression line



Properties of OLS Estimators

- we get different estimates if we took a different sample
- the estimated coefficients b_0 and b_1 are random variables and they each have sampling distributions
- hypothesis tests about coefficients can be conducted using knowledge of these sampling distributions

OLS

Under certain conditions:

1) OLS Estimator is unbiased
2) OLS Estimator has minimum variance among all the linear unbiased estimator
3) OLS Estimator is consistent

- All estimator with these property is called BLUE (best linear unbiased estimator)

4) If we add the normality of errors assumption then the OLS estimator has minimum variance among all unbiased estimators (linear and nonlinear)

Assumptions of OLS:

- assumed model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

- the expected value of the error is zero $E(\varepsilon_i) = 0$

- same variance $Var(\varepsilon_i) = \sigma^2$

- uncorrelated errors $Cov(\varepsilon_i, \varepsilon_j) = 0$

- x is not correlated to ε

- (optional) $\varepsilon_i \sim N(0, \sigma^2)$

Properties of OLS Estimators:

OLS Estimator for parameters:

$$b_0 = \bar{y} - b_1\bar{x} \qquad b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Under our assumptions b_0 and b_1 are normally distributed with mean β_0 and β_1 and following variances:

$$\sigma_{b_0}^2 = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \sigma_{b_1}^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Often σ^2 is not known:

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Hypothesis Testing

The hypotheses are:

H_0: $\boldsymbol{\beta_1 = D}$ where D is often zero

H_A: $\boldsymbol{\beta_1 \neq D}$

H_A: $\boldsymbol{\beta_1 > D}$

H_A: $\boldsymbol{\beta_1 < D}$

The Test Statistic:

$$\mathbf{t} = \frac{b_1 - \beta_1}{s_{b_1}}$$

Q: Does any linear relationship exist between AGE and INTERNET use?

Step1

H_o: $\beta_1 = 0$

H_A: $\beta_1 \neq 0$

Step2:

If OLS conditions hold, then use t statistic, when will be t distributed.

Step3:

$\alpha = 0.05$

Step4:

Recall n = 300, so reject null if $t > t_{\alpha/n,\, n-2} = t_{0.025,\, 298} = 1.972$ or if $t < -t_{\alpha/n,\, n-2} = -1.972$

Step5:

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{-0.435 - 0}{0.0222} = -19.58$$

Step6:

Since t is lower than -1.972, reject the null hypothesis. There is statistically linear relationship.

Confidence Interval

-an interval which we have a certain degree of confidence contains the value of the parameter

- confidence interval for $\boldsymbol{\beta_1}$

$$P\left(-t_{\alpha/2} < \frac{b_1 - \beta_1}{s_{b_1}} < t_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(b_1 - t_{\alpha/2} s_{b_1} < \beta_1 < b_1 + t_{\alpha/2} s_{b_1}\right) = 1 - \alpha$$

$$CI = [b_1 - t_{\alpha/2} s_{b_1},\ b_1 + t_{\alpha/2} s_{b_1}]$$

$$CI = [-0.435 - 1.972 \times 0.022,\ -0.435 + 1.972 \times 0.022]$$

$$CI = [-0.478,\ -0.392]$$

Multiple Regression (L11)

Multiple Regression

Multiple regression – A regression with more than one explanatory variable (as there are more than one important variable affecting the dependent variable)

- nearly all the results for the simple linear regression model is valid for multiple regression except:

→ interpretation of coefficients

→ degrees of freedom

→ assumption about the X's

Population model: $y_i = \beta_0 + \beta_1 x_{1,i} + \ldots\ldots \beta_k x_{k,i} + \varepsilon_i$

Sample model: $y_i = b_0 + b_1 x_{1,i} + \ldots\ldots b_k x_{k,i} + e_i$

where i = 1,....., n indexes the observations and have k different explanatory variables

Interpretation of the Coefficients

- the coefficients have the same interpretation as simple regression if we held other variables constant⇔

$$\beta_j = \left.\frac{\Delta y_i}{\Delta x_{ji}}\right|_{\text{all other variables are held unchanged}} = \frac{\partial y_i}{\partial x_{ji}}$$

Assumptions

1. Assumed model is $y_i = \beta_0 + \beta_1 x_{1,i} + \ldots\ldots \beta_k x_{k,i} + \varepsilon_i$
2. $E(y_i)\ \beta_0 + \beta_1 x_{1,i} + \ldots\ldots \beta_k x_{k,i} + \varepsilon_i = 0 \Leftrightarrow E(\varepsilon_i) = 0$
3. Homoskedastic errors: $Var(y_i) = Var(\varepsilon_i) = \sigma^2$
4. Serially uncorrelated errors: $Cov(\varepsilon_i, \varepsilon_j) = 0$ where $i \neq j$
5. $x_{k,i}$s are not random variables but are not exact linear function of the other explanatory variables
6. (optional) $\varepsilon_i \sim N(0, \sigma^2)$

OLS Estimation

- use computers and statistical software like EViews to estimate OLS because there are some very complicated formulae.

Estimation of Error Variance

- least square residuals are given by $e_i = y_i - \hat{y}_i = y_i - b_0 - b_1x_{1,I} - \ldots - b_kx_{k,i}$

- we can use the residuals to estimate error variance

$s_\varepsilon^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}$ where k+1 is the number of βs being estimated in the multiple regression model and n is the sample size

Standard error of estimate: $S\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$ where n=sample size and k=number of independent variables in the model

Properties

- under assumptions 1-5, OLS estimator of regression coefficients satisfy following properties:

→ it is consistent

→ it is BLUE: OLS estimator is unbiased and have minimum variance among all other linear estimators

→ if we assume normality of errors-6 as well, the OLS estimator have minimum variance among all unbiased estimators

→ the estimated coefficients are distributed as $b_k \sim t_{n-k-1}(\beta_k, s_{b_k}^2)$

Goodness of Fit

$y_i = \beta_0 + \beta_1x_{1,i} + \ldots\ldots \beta_kx_{k,i} + \varepsilon_i$

- y is determined by:

1) a systematic component (the economic model)

2) a random (unexplained) component, e

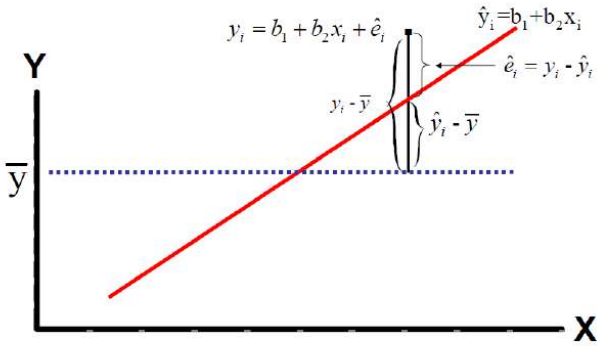
- the more one can explain changes in y by model the better, but before dealing with that issue we need a measure that can decompose variations in y into:

1) changes explained by the model

2) changes not explained by the model (e)

- one such measure is R^2

-Geometric representation of the decomposition:



$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}e_i^2$$

SST **SSR** **SSE**

SST (Total Sum of squares) – The total variation in y about its mean (measures the total variation in a variable)

SSR (Sum of squares for regression) – Measures the variation in a variable explained by the regression model

→ variation **explained** by the economic model (variation in about its sample average or sample mean)

SSE (Sum of squares for error) – The sum of squares of the absolute differences between actual and forecast values

→ variation that is **unexplained** by the economic model (variation in about its sample average, or sample mean, which is zero)

Note that by definition, we can write $y_i = \hat{y}_i + e_i$

then

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i+e_i-\bar{y})^2 \\ &= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}e_i^2 + 2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})e_i \\ &= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}e_i^2\end{aligned}$$

Zero

R^2 (coefficient of determination)

- The proportion of the variation in the dependent variable that is explained by the variation in the independent variable(s)

$$R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST} \qquad 0 \le R^2 \le 1$$

$$R^2 = \frac{Explained\ variation \in y}{Total\ variation \in y}$$

- It can be shown that $R^2 = r^2_{y\hat{y}}$

- If we only have one x variable, then $R^2 = r^2$, the square of the Pearson Correlation Coefficient

- There is not a particular 'test' for R^2 measures

- It provides a summary measure of the goodness of fit of model

- R^2 measures CANNOT be directly compared across model for different dependent variable (different y's)

Adjusted R^2 (coefficient of determination adjusted for degrees of freedom)

- A measure of the relationship between the dependent and independent variables, adjusted to take into account of the number of independent variables

- if the number of independent variables k is large relative to the sample size n, the R^2 value may be unrealistically high

- to avoid creating false impression, the adjusted R^2 is often calculated

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-k-1)}{SS_y/(n-1)} = 1 - \frac{MSE}{s_y^2}$$

- if n is considerably larger than k, the actual and adjusted R^2 values will be similar

- if k is high relative to n (SSE ≫ 0), R^2 and adjusted R^2 will differ substantially

- always adj $R^2 \leq R^2$

- The adj R^2 measure may go up or down if we add another x variable. It depends on whether the added variable bears any relationship to our dependent variable y or not.

Testing Validity of a Model

- can test whether any of explanatory x variables are linearly related to dependent variable y with an F-test

$$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}$$

- F distribution with k and d.f = n–k–1 (as long as the required conditions for OLS hold)

- if the variation in y that is 'explained' by our x variables(SSR) is large relative to the variation that remains unexplained (SSE), then we reject the null that none of the x variables are related to our y variable.

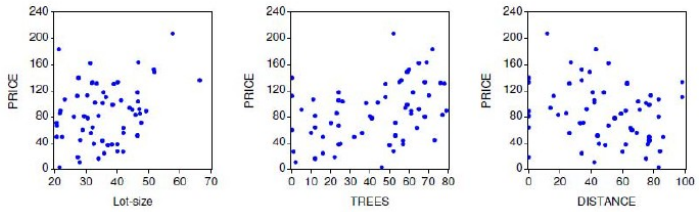
Examples

A Property Developer is considering the purchase of a tract of land for subdivision into holiday cottage properties.
She wants to know how much the subdivided lots will sell for, i.e. She wants to predict a sale price for each lot.
The developer has learnt from past experience ("theory") that sale prices are affected by:

(1) lot size,
(2) number of mature trees
(3) distance from lake.

She then gathers data for 60 recent sales of similar lots in a nearby area.

Least Square Principle



Price - in $000s **Lot Size** - in hundreds of square metres

Trees - count of mature trees **Distance** - in metres from the lake

EViews Interpretation

Dependent Variable: PRICE
Method: Least Squares
Date: 08/28/14 Time: 10:18
Sample: 1 60
Included observations: 60

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LOT_SIZE	0.699904	0.558855	1.252389	0.2156
TREES	0.678813	0.229306	2.960292	0.0045
DISTANCE	-0.378361	0.195237	-1.937961	0.0577
C	51.39122	23.51650	2.185326	0.0331

R-squared	0.242472	Mean dependent var	86.71333
Adjusted R-squared	0.201890	S.D. dependent var	45.04688
S.E. of regression	40.24353	Akaike info criterion	10.29212
Sum squared resid	90694.33	Schwarz criterion	10.43174
Log likelihood	-304.7635	Hannan-Quinn criter.	10.34673
F-statistic	5.974883	Durbin-Watson stat	2.174578
Prob(F-statistic)	0.001315		

Estimated regression equation: $\widehat{price_i}$ **= 51.39 + 0.70Lot_Size$_i$ + 0.68Trees$_i$ - 0.38Distance$_i$**

Each slope coefficient $b_j(j=1,.,k)$ is the average partial effect of that explanatory x_j variable on dependent variable y

$$\frac{\partial \hat{y}}{\partial x_j} = b_j \qquad \text{Example:} \qquad \frac{\partial \hat{y}}{\partial x_2} = \frac{\partial \widehat{\text{price}}}{\partial \text{trees}} = b_2 = 0.68$$

Interpretation: Holding lot size and distance fixed, increasing number of trees by 1 (one unit) is associated with $680 higher average selling price (0.68 of unit).

R^2 = 0.242, thus 24.2% of the variation in selling prices can be 'explained' by the three explanatory variables: lot size, mature trees and distance from lake.

R^2 = 0.242 and adj R^2 = 0.202, these are not too far apart. We have not overfitted the model here: just 3 independent (x) variables.

Hypothesis Testing

- The p-values in the EViews table of regression output above are for the following two-tailed test:

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

- this test employs the same t statistic used for simple regression with v = n - k - 1 for the general case

- to use it validly, the required conditions of OLS must hold

- **conclusion**:

  → the coefficient on lot size is not significantly different from zero, so we do not reject the null hypothesis as p-value > 0.10

  → the coefficient on trees is highly significantly different from zero (p-value <0.10)

  → coefficient on distance to lake is not significantly different from zero at 5% level, but it is very close (0.05 < p-value < 0.10), there is weak evidence of a relationship

- **interpretation**:

  → we can say that price of lots is positively related to the number of trees

  → we cannot say that price of lots is significantly related to lot size, as the coefficient is not significantly different from zero.

  → we can say that there is weak evidence that distance from the lake is negatively related to the price of lots

  → the coefficient on the constant term (b_0) is the predicted average selling price of a lot where all x variables are set to zero, thus the b_0 estimate doesn't always have a useful interpretation

Step1

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

H_A: At least one slope β_j is not equal to zero

Step2:

As long as the OLS conditions hold, then use F-statistic

Step3:

$\alpha = 0.05$

Step4:

Recall n = 60 and k=3, so reject null if $F > F_{\alpha, k, n-k-1} = F_{0.05, 3, 56} = 2.76$

We only reject the null if F-statistic is large, i.e. that SSR is large relative to SSE. Both SSR and SSE are always positive

Step5:

Eviews: F = 5.97

p-value = 0.0013

Step6:

We reject the null of all the slope coefficients equaling zero. The model is valid and has utility. The x variables help explain y.

NOTE: This F-test is very much related to the ANOVA F-test