**Research Methodology – Statistical methods**

**Lubor Homolka**



**2018**

Contents

[Contents 2](#_Toc523489416)

[Introduction 3](#_Toc523489417)

[1 probability theory 4](#_Toc523489418)

[1.1 Random Variable 5](#_Toc523489419)

[2 Desing of the Quantitative research 8](#_Toc523489420)

[2.1 Power Sample Analysis 10](#_Toc523489421)

[3 regression analysis 12](#_Toc523489422)

[3.1 Types of Regression Model 12](#_Toc523489423)

[3.1.1 Type of dependent variable 13](#_Toc523489424)

[3.1.2 Data design 16](#_Toc523489425)

[3.1.3 Types of Independent Variables 17](#_Toc523489426)

[3.2 Log models 17](#_Toc523489427)

[3.3 Matrix Notation of the Regression model 19](#_Toc523489428)

Introduction

Validity of the theoretical prepositions is mainly based on comparing scientific prediction to the real empirical findings. Empirical findings can have various types. We generally distinguish between qualitative (outcomes of structured interviews with studied agents) and quantitative (results of an experimental trial or questionnaire survey) outcomes. The latter is predominantly used in the manufacturing engineering studies due to the methodological nature of the studied problems. This does not, however, mean that the former shall not be used or studied. This approach might even be preferable if the engineering topic is concerned about the managerial perspective.

This part of the study material titled “Statistical Analysis” approaches the problem through the quantitative lenses. It presents the topics of the data design, convenient methods and techniques designed to assess a validity of research hypothesis or questions. Content focuses on the methodological part of the issue. Several tools are presented to pinpoint to various possibilities and applications. Understanding of the content requires additional study and consultation.

# probability theory

This chapter outlines the foundation of the probabilistic theories applicable in the social sciences.

Probability is a mathematical measure with several definitions. It’s crucial to understand each of them to correctly interpret results which are derived upon their usage. We distinguish following probability types:

* Standard
* Classic (frequentist)
* Subjective

There is also the axiomatic definition of probability which list requirements of probability. If any measure follows these requirements (such as the measure is only defined in the range 0-1).

**Standard probability** is a measure of a chance that some event happens. This event must have finite number of possible outcomes with equal chance of occurrence. As a good example can serve a roll of a dice with 6 edges (finite number) where a chance of seeing number 1 equals to a chance of seeing number 6 as the outcome of the experimental roll. The clear advantage of the standard probability is that the probabilities can be assessed by theoretical reasoning. It’s unnecessary roll a dice 100 times to find out the probability of seeing number 1.

**Classical probability** is also called a **frequentist probability**. Probability of occurrence of some event is assessed by empirical means. This empirical analysis requires computation of relative frequencies on the sample with infinite number of repetitions (definition) or on a very large sample (practical view). Probability of the event or its outcome is computed as:

This probability type dominates in the current scientific works and teaching literature. All results which connect to the probability statement must be interpreted in terms of frequencies unless it’s stated that the analysis is based on the Bayesian reasoning. This fact tremendously complicates the interpretation of elementary terms such as p-value or confidence intervals.

Classical probability should be used in the experimental settings due to its nature which requires replications. The true probability cannot be found unless we have number of observations. We can, however, estimate the probability from the sample analysis.

**Subjective probability** is the last type of probability. This probability is the underlying principle of the Bayesian analysis. Bayesian approach is progressively developing field which can answer the same question as the classical frequentist approach. It uses different methodological perspective as well as tools, though.

This type of probability does not work with frequencies and does not require infinite number of observations. Instead, it expresses the degree of believe of the event to happen / to be true. Therefore, it is a convenient approach in situations where the study cannot be (easily) replicated as in the previous paragraph.

## Random Variable

Random variable is used to describe studied variable/event in a mathematical way. The outcome of the event is derived by random experiment (future value is not known to the experimenter). Random variable is a real function on the set of appropriately-defined inputs. Type of a random variable is linked to the outcome of the function. We distinguish between:

* Discrete (finite and small number of possible outcomes)
* Continuous (very large or infinite number of possible outcomes)

Distribution of the random variable is used to describe its behaviour. Every outcome of the discrete random variable, or interval of outcomes in case of continuous variables, is linked to the particular probability value.

Random variables can have a very narrow purpose and interpretation. “What is the probability of seeing *x* heads if we toss 10 unbiased coins?” is just a function of relative frequency. In many applications random variable is not directly linked to the units of underlying data. Consider continuous random variable called *t* which is also a test criterium of the one-sample t-test:

Let’s name the discrete random variable of the first example . We can use the binomial distribution to describe a behaviour of the random variable (how likely we will see all possible outcomes). Outcome of the random variable is denoted as and ranges from 0 to 10 (10 head out of 10 tosses). We can write the random variable mathematically as:

Value is a true probability of outcome being a positive result (in our example outcome is Head). It equals to 0.5 if the coin is unbiased. This value is often the object of your scientific work described in the research hypothesis. We can find this value only if we collect infinite number of observations. As this option is not viable, we need to settle with the estimate denoted as *p*.

The second case is the t random variable. This variable follows student’s distribution and can be written as:

where is a parameter called degrees of freedom computed as

It is not possible to compute a probability value of the point value of the continuous random variable. Contrary to the discrete case, where it’s possible to compute probability of seeing 5 head out of 10 tosses, it is not possible to compute a probability of . We only can compute a probability of having value 1.5 and lagrer/smaller/around.



Image 1 Density function of the students distribution. Source: Own

This limitation is one of the reasons why concept of p-value is not computed for the null hypothesis only but reflects oriented alternative hypothesis.

# Desing of the Quantitative research

A convenient methodological approach has to be set after the research hypothesis are defined. Let’s assume a simple design on the cross-sectional data.

Turning research hypothesis into the statistical hypothesis is done in **the first step**. If the scientific hypothesis postulates that:

*“New approach increases the efficiency by more than 10% compared to the old approach.”*

Two scientific hypotheses can be derived from the hypothesis above. The first one is so called null hypothesis H0 which claims that there is no positive effect. This is the statement which we want to refute. If the true effect is written as θ then:

H0: θ = 0.1

According to the classical approach θ is a fixed and value which can be found only on the infinite sample (population).

Null hypothesis contains the opposite of what is wanted to be achieved in the research. The alternative hypothesis HA negates the null hypothesis:

HA: θ > 0.1

An appropriate test is selected in the **second step**. Selection of the test has to reflect several conditions at one time, such as a sample size, or data type or data design. Generally speaking, parametric tests (such as t-test) do a better job in finding small effects (e.g. θ) than non-parametric. Non-parametric tests are, on the other hand, able to work with fewer data. Let’s finish the example by proposing three methods to verify the presence of positive effect.

1. Regression Analysis

The efficiency *yi* in the company *i* is explained by dummy factor x1 which takes values 0 when the old approach is used and 1 when the new approach is being applied and by control variable x2 which can be some relevant variable to the analysis which is not a main concern of the research. The mathematical model then can be written as:

After all coefficients are estimated we can infer about the effects: is general effectivity if the old approach is used, is a true effect of the new approach (note that the outcome of our sample analysis is only an estimate of this parameter). It is possible to test whether the is bigger than 10% by t-test. Common statistical packages offer t-tests directly, These need to be adjusted as they test whether there is an effect at all H0 :θ = 0

1. Two-sample t-test

Two sample test has the same null and alternative hypothesis as written in the text. The major disadvantage of this approach to the regression approach is a missing possibility of controlling for other variables.

1. Non-parametric Man-Whitney U test (MW test)

MW test t is a non-parametric counter version to t-test. Although both tests are used in an analysis of expected values, t-test is a test of mean values whereas MW test analyses mode values. MW test can be used in case of unequal variances of groups and when the sample size is smaller than 30. Major disadvantage of the test lower power compared to t-test (chance of finding an existing effect θ is lower than in case of t-test).

**The third step** involves testing residuals and other conditions. Analysis of residuals can reveal patterns which should be captured in the model (problem of model’s misspecification). Following image describe



Image 2 Residuals of the linear regression model. Source: Own

## Power Sample Analysis

**Power sample analysis** is an important branch of statistical analysis which allows researchers to assess requirements of the sample size and other characteristics before the data-collection phase. There are several types of the analysis:

1. A priori
2. Calibration
3. Post-hoc

**A priori approach** is the most widely and useful approach. This analysis is performed before data collection. It requires information about the nature of the studied phenomena. Usually, this analysis is performed on the pilot data summaries or on previously published results. This analysis usually helps to decide a minimal sample size.

**Calibration approach** is also done before seeing the data. Its main purpose is to find a convenient method and potential compromises on the qualitative side of the research. It can also be used in a decision-making process – whether to start the empirical research given some conditions (feasible sample size ~ required sample size).

**Post-hoc approach** is a controversialapproach which is, according to prominent researchers in the field, redundant. (Cohen, 1990). Post-hoc analysis is performed after the data is collected and the power is computed retrospectively. The main argument against this approach is that the test either finds an effect (therefore is powerful) or does not find an effect (was not powerful). The question how powerful the test was is no longer interesting to the researcher as the research is done.

# regression analysis

The main outcome of any regression analysis is a creation of mathematical model which captures relation between variables. Behaviour of the dependent variable is explained by a set of independent variables stored in the data matrix , by a convenient mathematical function parametrised by coefficients stored in the vector .

Regression model provides a conditional expected value of the dependent variable. This condition is given by values of independent variable. Expected value *E* is usually a mean value. We can rewrite this regression settings as:

This specification only highlights the conditional nature of the function. In the next modelling stages function is estimated by some convenient estimation technique, such as Ordinary Least Squares.

## Types of Regression Model

Regression model introduced in the traditional statistical courses is just a special case of the general regression framework. It is traditionally estimated on the cross-sectional data and can be called a simple linear model. This model can be written as:

(1)

It is a univariate model as there is only one independent variable . From the formula above we can identify followings:

* Mathematical form of relation is linear. There is no or variable.
* Function is linear in parameters. This is important condition for OLS estimation. Relation between coefficients must be additive. It is still possible to have or any non-linearly transformed variables in a model.
* Model contains residual component. Properties of the residual part are not explicitly mentioned.
* Model is fitted on the cross-sectional data as the variables have only one index indicating observation. If the model is created on the panel data, variable would have a time-index in addition to the abovementioned (as).

Following aspects must be considered in the modelling process:

1. Type of dependent variable
2. Data design
3. Type of independent variables
4. Interactions between variables

### Type of dependent variable

We can distinguish following three data-types in the modelling framework:

1. Metric / Scale
2. Ordinal
3. Nominal

All three types can be used in regression models on the position of dependent or independent variable. Simple linear regression assumes that the dependent variable is metric. It is necessary to use Generalised Linear Model in other cases.

As a typical example can serve a regression model with dichotomous dependent variable. If the simple regression model is employed on the data (incorrectly coded as 0,1 for NO and YES answer) model’s predictions would go below and above the limits of dependent variable.



Image 3 Incorrectly specified mode with binary dependet variable. Source Own

The proper approach to this type of analysis is to treat a conditional dependent variable as a probability of having YES. This model is called a logistic regression. A typical sigmoid function which does not allow exceeding bounds 0-1 is showed in the next figure.



Image 4 Sigmoid function bounds the estimates to 0-1 values. Source: Own

Type of the dependent variable closely relates to the distribution of the error term. If the dependent variable has a discrete type, normality of residuals is no longer expected as it is in the simple linear regression.

**Illustration:** Simple regression model analysis conditional mean value of variable . Let’s consider following experiment which fixes levels of independent variable.



Image 5 Regression model and conditional expectations (means). Example of homoscedasticity. Source: Own

Regression function estimates mean values in each level . Errors of the model should be equally distributed above and below the line. Density of the points should diminish as the value deviates more from the mean value. This error can be described by normal or student distribution. Also, as required by the simple regression model, the variance of errors in each conditional level is constant (condition of homoscedasticity). If the constant variance is denoted as , then the model can be rewritten as:

Because the computes the conditional mean value. Specification of the model according to this notation is preferable compared to the equation 1. It is customary to specify regression model in two steps:

Where the first line is so-called link function. In case of logistic regression, link function is called *logit.* Many other models with special dependent variables exist. Another example is a count model with residuals following Poisson distribution.

Models with ordinal, nominal or censored (survival analysis) dependent variable require special modelling strategies in addition to specification of residual component.

### Data design

Data design can be classified according following dimension on which the data is recorded.

* Time
* Hierarchy
* Space

Cross sectional data is a data type which is collected at the same time (or within a time frame with non-changing conditions). Panel data is a data type which records information about entities across time. Time series is usually used as a term for univariate collection of data which is ordered chronologically.

Hierarchical, multi-level or nested data is a special data design which is used if the data contains a structure which is important to analyse to get the whole picture.

**Illustration:** Researcher wants to show that a new approach improves the quality of the outcome. She runs an experiment and collects 10 samples produced by a first shift and 10 samples from the second shift. Then she repeats the experiment, but new approach is used this time. Grouping factor “Shift” is an important hierarchical factor.

Last category is a spatial data design. This type of data design requires a special treatment as the standard tools developed for cross-sectional data (such as proximity of units, correlations, etc.) cannot be used directly.

Special methodology for analysing aforementioned data designs was developed. Ignoring the panel data nature and considering data as cross-sectional would lead to overly-optimistic results. Null hypothesis testing would yield to statistically significant result more often than it should. Time dimension need to be accounted in the analysis due to the autocorrelated values.

### Types of Independent Variables

Similarly to the dependent variable we distinguish various data types. It is important to spend more time on transformation and coding of the data as it directly affects the interpretation. In many cases transformation (such as taking logarithms of x) ease the estimation of models. Estimation of the parameters might be even impossible due to big differences in scales if iterative estimation technique is used. This is also a reason why the standardisation of the underlying data is often recommended.

Coding of the ordinal and nominal variables creates dummy variables. There are several approaches how to code levels effectively through the contrasts (e.g., linear, polynomial, sum or helmert).

## Log models

Simple linear model can be modified by transformation of analysed variables. This transformation is usually done to capture non-linear relation between dependent and independent variables. Transformations change an interpretation of the estimated coefficients although the form of the model remains similar to the original one:

It is possible to create 3 new models from the original one as showed in the table.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | Simple linear model | Linear-Log |
|  |  |
|  | Log-Linear | Log-Log |
|  |  |

Table 1 Three types of log models which can be derived from the simple regression model. Source: Own

Interpretation of the **Simple linear** model is obvious.

**Linear-Log** model is created by taking a values of independent variable before estimation of parameters by standard OLS method. The model is no longer interpreted in the real-world values, but on values instead. It means that the unit change is no longer a unit change of as the model considers a unit change of . A unit increase therefore corresponds to increase of , which is approximately . It is a convention not to interpret values as multiple increasement. Instead, percentage growth is preferred. Value translates in 171,27 % growth (100 % growth from 10 is 20). From this we can derive following relation:

Expected change of associated with change of by percent is computed as , where is the estimated regression coefficient. Therefore, increase of usually leads to increase in by .

**Log-linear model** is computed by taking values of the dependent variable and fitting OLS model. A unit change of corresponds to the change of by .

**Log-Log model** is a popular model in the microeconomics or in models analysing elasticity of change (percent change of leads to percent change of ). Interpretation of the coefficients is not intuitive due to the values on both sides of the regression model. An increasement of is associated with change in by . It is a convention to interpret regression outcomes in term of percent change. Proportional change in to a change of can be computed in two steps. In the first, quantity is evaluated:

Which is supplied to the formula .

## Matrix Notation of the Regression model

Standard notation of the linear model introduced in the previous chapters can be written in a matrix notation.



Image 4 Algebraic notation of the simple regression model. Source: Own

Models written in algebraic way are common in the statistical literature due to its explanatory power and intuitiveness. Benefits become clearer when more complex models are being analysed. Such models usually involve intercorrelated structure of residuals, fixed and random effects, interactions of factors or specific contrasts required for interpretation.

Algebraic notation can also simplify general notation of the model. Consider simple model rewritten as (See figure above).