**Research Methodology - Statistics**

**Lubor Homolka**



**2018**

Contents

[Contents 2](#_Toc1030118)

[1 Power sample analysis 3](#_Toc1030119)

[1.1 T-test example – minimal sample size 3](#_Toc1030120)

[1.2 T-test example – Calibration 5](#_Toc1030121)

[2 Regression Analysis 7](#_Toc1030122)

[2.1 Log models 7](#_Toc1030123)

[2.2 Logistic Regression 9](#_Toc1030124)

[2.2.1 Logistic regression – Intercept Model 10](#_Toc1030125)

[2.2.2 Logistic Regression – Model with Categorical Variable 11](#_Toc1030126)

[2.2.3 Logistic Regression – Model with Metric Variable 13](#_Toc1030127)

[2.2.4 Logistic regression in the research paper 15](#_Toc1030128)

[3 Statistical hypothesis 20](#_Toc1030129)

[3.1 Differences in means – T-test 20](#_Toc1030130)

[3.2 Frequentist and Bayesian Regression Analysis 22](#_Toc1030131)

[3.2.1 Frequentist Model 22](#_Toc1030132)

[3.2.2 Bayesian Approach 23](#_Toc1030133)

[4 Time series modelling 25](#_Toc1030134)

[4.1 Correlation of time series 25](#_Toc1030135)

[4.1.1 Problem of non-stationarity 25](#_Toc1030136)

[4.1.2 Correct measure of correlation 30](#_Toc1030137)

[4.2 Stationarity of the time series 31](#_Toc1030138)

[4.2.1 Deterministic (trend) stationarity 32](#_Toc1030139)

[4.2.2 Stochastic (unit root) stationarity 33](#_Toc1030140)

[references 40](#_Toc1030141)

# 1 Power sample analysis

## T-test example – minimal sample size

The team of researchers know from the literature review that the average value of the target social group spends 500 Kč monthly on a certain product of interest. The previous paper also mentions an estimate of standard deviation of size 30 Kč. Control social group spends on average 515 Kč with standard deviation 40 Kč. The team wants to show that there is a difference in consumption behavior -> target group spends more. How many responses is it necessary to collect if the journal’s requirements are set as:

1. Power of test = 0.8
2. Tolerance to I. type error (alpha) is 5%
3. the team can only use *t-test*

**Solution:** The size effect needs to be estimated in the first step. This can be done in statistical software. Following demonstration uses the GPower software which is free to use. Mean values and standard deviation need to be supplied to compute a standardized value of effect size *d*.



Image 1 Effect size estimations for Welch’s *t-test Source: Own*

Given the small difference in the mean values relative to the data variance, we get the SMALL effect size 0.424. Therefore, research can be considered ambitious which will project in the large sample size.



Image 2 Settings of qualitative indicators. Source: Own

We conduct the analysis before collecting data. Hence the A priori analysis (red box).

Green box contains other qualitative requirements, such as requested power or level. This box also contains the estimated effect size from Image 1.As the research aim is to show that one group spends more than the other we set One-tail analysis. Allocation ratio expresses our expectation of future results, whether there will be more respondents from the target or control group. Number 1 indicates that the sample will be more-or-less balanced. It does not mean that the team will perform quota sampling, though. It means that it’s equally likely to have a member from group 1 as from group 2 in the random sample.

The outcome from the power analysis is in the blue box. It tells the minimal value of the t-statistics, but, more importantly, the required minimum number of respondents in each sample.

It is important to stress that 70 is a bare minimum which needs to be reached by random sampling.

## T-test example – Calibration

Let’s continue from the previous example. Assume that the number of 140 respondents is not feasible for the research team. The research team cannot change estimated sample size – this value is supported by previous research and cannot be reduced. What are the options left?

1. Decrease power



Image 3 Sensitivity of sample size on test’s power . Source: Own

Image 3 points to the non-linear nature of the relation. As the requirement on more power increases, total sample size increases exponentially. If the team can have max. 100 respondents, they will have to accept the power of 0.675.

1. Increase alpha

Similarly to the relationship with power, decreasing of alpha requires non-linear increasement of the sample size. Following the previous example with 100 respondents, the research team would need to work with alpha 0.1. The chance that the team will find a non-existing effect therefore increases.



Image 4 Sensitivity of sample size on test’s power . Source: Own

# Regression Analysis

## Log models

The task is to analyse the elasticity of variable on change of variable Data is visualised in the following scatter plot. Convenient model for this task is a log-log model



Image 5 The left panel shows data on the original scale. Right panel depicts transformed values on which the regression model is fitted. Source: Own

Regression model can be estimated by Ordinary Least Squares technique. Following outcome comes from the statistical package R.

Call:

lm(formula = lnY ~ lnX)

Function *lm* estimates regression parameters and provides set of inferential statistics. Following table provides a summary of residuals. Symmetry of errors around the median, which is virtually null, suggests that the model is unbiased.

Residuals:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min | 1Q | Median | 3Q | Max |
| -0.73647 | -0.15589 | -0.04212 | 0.16543 | 0.68556 |

Estimated coefficients:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |  |
| (Intercept) | 2.7521 | 0.0424 | 64.903 | < 0.01 | \*\*\* |
| lnX | 0.4558 | 0.0511 | 8.913 | < 0.01 | \*\*\* |

Both regression coefficients can be considered as statistically significant at the 5% confidence level. T-value shows that estimated value of intercept value is 64.903 times higher than the standard deviation of intercept’s sampling distribution.

The initial task was to interpret coefficient as elasticity estimates. Let’s say we are interested in estimated effect on if the increases by 25 %. Following the appropriate formula (see chapter in lectures), we can compute a change as . This change corresponds to the 10.7 % growth in . It is important to highlight non-linear response of . This nonlinearity is visualised in the following figure.



Image 6 Estimated elasticity is non-linear (notice a deviation from the grey line).

Although the regression formula ( looks like a simple regression () with straightforward interpretation of the slope the coefficient, interpretation get complicated by the log transformation.

The last part of the outcome from R provides summaries of overall fit qualities.

|  |  |
| --- | --- |
| Residual standard error: | 0.2976 on 48 degrees of freedom |
| Multiple R-squared: 0.6233 | Adjusted R-squared: 0.6155 |
| F-statistic: 79.44 on 1 and 48 DF | p-value: < 0.01 |

Regression model was estimated well. Model explains 62.33 of the data variance. F-test confirms what we know from the t-test (because we have only one independent variable, null hypotheses are the same and the F statistics equals to square of the t-statistics.)

## Logistic Regression

Let’s consider an example with a binary dependent variable indicating positive / negative outcome of analysed process. We want to analyse relations between the dependent variable and one dichotomous variable (old – new approach) and one continuous metric variable (standardised values of pressure). We have a balanced dataset with an equal number of positive and negative outcomes.



Image 7 Scatter plot presents relation between dependent variable (1=Positive, 0=negative), values of pressure (x scale). Color indicates value of binary independent variable representing approach. Source: Own

### Logistic regression – Intercept Model

Let’s start with the simple model which predicts the outcome just by the intercept value. As we know that the experiment is balanced, intercept value should convey information about the fraction of positive values.

As the logistic regression is a nonlinear function, OLS cannot be used to estimate coefficients. There is a function called glm in R package which can be used to fit all generalised regression models. Here is the outcome:

|  |
| --- |
| Call: |
| glm(formula = y ~ 1, family = binomial(link = "logit"), data = dataReg) |

Regression coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z-value | Pr(>|z|) |
| (Intercept) | 0.0000 | 0.2828 | 0 | 1 |

Although the estimate is precisely 0, standard error is non-negative. This is caused by the stochastic nature of the data reflecting the number of observations. Estimated value of intercept equals to which does not have a direct interpretation. To compute a value , which is the probability of having a positive outcome, we need to use a logistic function defined as . In our case we see , which is exactly what we have expected to see due to the balanced data set. If the majority of outcome would have been positives, would be higher than 0.5.

### Logistic Regression – Model with Categorical Variable

We will try to answer the question whether application of the new approach leads to positive results of outcome. Looking back to the Image 3 we can assume that new approach leads to the positive outcomes, though not absolutely.

From the 50 observed outcomes 30 were according to the new approach. From the 30 outcomes 23 were positive (76.7%). Therefore, chance of having a positive outcome given the new approach is . In the old approach scenario was achieved 2 positives out of 20 outcomes. This means chance of positive result is . This translates in odds ratio of . It is times likely to have a positive outcome if the new approach is used.

Chances should be written in a form to indicate comparison. It is, however, a convention to write only the first number. If the chance is 1, then neither of groups leads to more positives.

Odds ratio value smaller than one is harder to interpret. It is necessary to revert the odds. Consider a chance 0.8 of old approach being successful. It would mean a chance of 1.25 of new approach being successful compared to old one. Following plot provides the solution by employing function .



Image 8 Conversion of positive to negative odds. Source: Own

Following code runs the model with the dichotomous group variable.

|  |
| --- |
| Call: |
| glm(formula = y ~ group, family = binomial(link = "logit"), data = dataReg) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z-value | Pr(>|z|) |
| (Intercept) | -2.1972 | 0.7453 | -2.948 | <0.01 |
| groupNew | 3.3868 | 0.8613 | 3.932 | <0.01 |

Both regression coefficients are statistically significant (differ from 0). Coefficients are not directly interpretable as in the previous example. Reading from the outcome above, *groupNew* is a contrasting group. Therefore, Intercept refers to the Old group.

To derive an interpretable value, we need to compute following: . Value contains an information about which can be recovered as (compare with the odds ration computation above). We can proceed in the same way to find an interpretation of New approach variable: . This is the value of odds ratio compared to the reference group. If the task is to find a chance of having a positive outcome in a New approach group, we would just plug both coefficients into the computation:

### Logistic Regression – Model with Metric Variable

Following code runs the model with intercept value and numeric value.

|  |
| --- |
| Call: |
| glm(formula = y ~ x, family = binomial(link = "logit"), data = dataReg) |

Intercept value is statistically insignificant. Pressure seems to be an important (at least statistically) variable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z-value | Pr(>|z|) |
| (Intercept) | -0.018 | 0.367 | -0.049 | 0.961 |
| x | 2.162 | 0.609 | 3.549 | <0.01 |

If the value of equals to 0, then probability of having a positive outcome equals to .

As the increases, the probability of having a positive outcome increases as well (positive sign of the estimate). As in the previous models, estimated effects for a unit change of .

Let’s compute a probability of having a positive outcome if:



Image below the paragraph summarises results and points to non-linear form of relation. It should be clear from the demonstration above that estimated coefficient 2.162 cannot be interpreted as a constant unit change effect, as it is in the linear regression model.



Image 9 Marginal effect of the metric variable x on probability of having a positive outcome. Source: Own

### Logistic regression in the research paper

In the paper Moomen at. al (2018) authors analyse factors which influence truck crashes in the United States. We can read from the Abstract:

*“This paper investigated the factors influencing truck crashes on downgrades; an attempt to fill in some of the research gaps. An empirical analysis of factors affecting truck crashes on two-lane downgrade roadways in Wyoming was carried out using a binary logistic regression technique.”*

The first part of the research paper discusses data collection technique. Authors have selected the Wyoming as roadways with a specific characteristic.

*“The prevalence of mountainous terrain and mountain passes in western United States means states like Wyoming have relatively high incidences of downgrade truck crash occurrence.”*

Therefore, it can be questionable whether the results can be generalised.

Authors have employed a binary logistic regression technique because:

*“The binary nature of the dependent variable (truck crashes or otherwise) makes logistic regression suitable for the analysis. Mathematically, the logistic regression model is flexible and intuitive which results in meaningful interpretations (Hosmer and Lemeshow, 2000).”*

In the section Methodology authors emphasise the difference in the estimation technique.

*“Unlike linear regression which uses the method of least squares to derive parameter estimates, logistic regression estimates parameters by the maximum likelihood method. The maximum likelihood estimation method yields parameter estimates which maximize the probability of obtaining the values observed in a set of data.”*

The first part of the paragraph can be considered as redundant as the estimation technique used by authors (maximum likelihood method) is generally accepted first-choice method. It is advisable not to mention estimation technique in detail, unless you use some method which is new or known for technical deficiencies (such as GMM for small samples).

Authors then continue with description of logistic model with formulas. They point out that:

*“Like any other regression model, logistic regression modeling involves assessment of the significance of variables in the model. In logistic regression, significance testing is achieved with the Wald test. The Wald test statistic is defined as beta/SE. The hypothesis of Wald test H0:beta=0 states that the probability of success is independent of x. Though the Wald test is adequate for larger samples, the more powerful likelihood-ratio test is preferred for sample sizes often used in practice (Agresti, 2007).”*

There are several statistical tests for individual parameters. Due to the nature of the dependent variable t-tests are not used as in the simple linear model. The test ratio is identical to the t-test, though. The Wald test uses computed likelihood which was used when the model was identified (recall the maximum likelihood method). This likelihood is then compared to the likelihood computed for the model without the analysed variable. Ratio of likelihoods then decides whether inclusion of the variable improves the model’s explanatory power.

Variable selection was done with respect to both estimation process as well as with expert’s opinion.

*“A subset of twenty-three explanatory variables was selected from the CARE database for this study. The subset of selected variables was done with a view to build the most parsimonious model that explains the occurrence of truck crashes on downgrades. The rationale for minimizing the number of variables in the model was to produce a numerically stable model which can be easily generalized.”*

This approach is controversial as it admits an error resulting from omitted variables. There are several methods how to solve a problem of a large value of explanatory variables. It is advisable to look at the correlation structure of variables. Correlated variables can be dropped from the analysis (while keeping one in the dataset). Another approach is to use a Factor analysis and compute either regression or Bartlett scores for each observation. Such scores summaries information content of several variables in one number.

*“The subset of variables was classified into six categories: driver characteristics, environmental factors, temporal factors, crash characteristics, traffic characteristics, and geometric features. Driver age and gender were noted for all vehicles involved in the crashes. Environmental factors which included weather, lighting and road condition were of important interest (Table 1).*

*Driver age was categorized into three groups to reflect the proportion of truck drivers involved in crashes based on results of the LTCCS (Federal Motor Carrier Safety Administration (FMCSA), 2006). These groupings are young aged drivers (25 years or younger), middle aged drivers (26-66 years), and old aged drivers (66 years and above).”*

All variables were directly observable. No latent variable was used. It was expected that all measurements were error-free. This is a common and accepted practice in the literature. However, when the latent variables are used, it is required to provide information about reliability and validity of measurements.

The authors continued with descriptive statistics of the sample and followed in a section 6 Model Calibration and results. This is not a common way as the first part of chapter (Calibration) is usually in the methodology section. Authors started with following:

*“A binary logistic regression model was fitted to the data. The variable selection method enables the user to specify the way in which the independent variables are entered into the model. In the stepwise selection procedure, statistically insignificant variables are removed from the model before adding a significant variable. The addition or removal of each independent variable is listed as a separate step and at each step a new model is fitted. The procedure is stopped when no more independent variables can be added or removed from the stepwise model. The variables retained in the model form the final model.”*

This text explains the variable selection process, which was partially covered in the variable introduction section. Stepwise selection was used in the model. This (and backwards) selection methods are common ways to find an optimal set of explanatory variables. There are, however, some drawbacks of usage of these methods. First, forward and backward selection can end up with different set of variables. Second, it requires a lot of models to be fit and compared due to the rules of combinatorics. Instead of this type of selection, we can use a lasso, ridge regression (both are so called shrinkage methods) or factor/principal component analysis.

Next paragraph presented indicators of a good fit.

*“The Hosmere Lemeshow test was used as a standard test for goodness-of-fit of the logistic regression. The prediction ability of the model was evaluated using the area under the receiver operating characteristic (ROC) curve (Hosmer and Lemeshow, 2000).”*

The Wald tests were already described in the previous section. Overall tests (very different than known F-test in simple linear regression) were introduced. These tests decided about a validity of the model based on its predictive power rather than by means of statistical inference. Next section dealt with interpretation of results.

*“The negative coefficient of weather (b= -0.6412) indicates that the probability of being involved in a truck crash on downgrades in clear weather is lower compared to adverse weather conditions. The odds ratio suggests the probability of a truck crash occurring on a downgrade in clear conditions is about 47% lower compared to adverse conditions.”*

The interpretation of effect of weather straightforward. Negative value of coefficient means that the probability of event decreases. Compare with the lighting conditions:

*“The results suggest a higher probability of truck crashes is associated with lighted compared to unlighted conditions (b = 0.4755). The odds ratio of 1.609 indicates that the risk of a downgrade truck crash is 61% higher in lighted conditions compared to unlighted conditions.”*

This paper does not use any continuous explanatory variable.

# Statistical hypothesis

This chapter demonstrates three approaches to statistical hypothesis testing. Consider researcher who starts with a new topic with sparse empirical findings. The question is whether there is a difference in attitudes between two types of managers. An attitude of interest can be measured on a scale large enough to be considered as a metric variable. The reference group “Type A” is compared to the target group “Type B”. The initial research conducted on 100 respondents yields to following data:



Image 10 Two types of managers and attitudinal scores. Source: Own

Dashed horizontal line in each boxplot indicates the mean value, full line is a median. Sample statistics indicate that there is a difference, average value of Type A equals to 5.03 with standard deviation 0.926. Average value of Type B is 5.35 with standard deviation 0.905.

## Differences in means – T-test

The standard t-test assumes that variances of both groups are identical. This assumption should be tested before the analysis by statistical test, such as F-test. Another option is to use Welch t-test with pooled variances. As the research is a new topic, and we don’t know which of the two groups is expected to have higher attitudinal score, alternative hypothesis will be set as two-directional.

|  |  |
| --- | --- |
| **Welch Two Sample t-test** | |
| data: data$typeB and data$typeA | |
| t = 2.2496, df = 97.951, p-value = 0.02671 | |
| alternative hypothesis: true difference in means is not equal to 0 | |
| 95 percent confidence interval: | |
| 0.049 0.775 | |
| sample estimates: | |
| mean of x | mean of y |
| 5.446 | 5.034 |

P-value indicates that there is a statistically significant difference between mean values of scores. Researcher knows that even small effects can be found by the statistical test if the sample size is large. Therefore, researcher will be also concerned about the substantial significance of findings. He decides that an interesting effect would be points (Initial statistical test tried to find any difference). If the effect is within this range, researcher would stop persuading the research as the effect is too small.

This problem can be solved by a series of two one-sided test (two-one sided tests, TOST methodology) to find a proof for an equivalence of minimal effect size. This size is called smallest effect size of interest (SESOI) in the literature (Lakens, 2018). In our example SESOI = .



Image 11 Confidence interval for the t-test and for TOST. Source: Own

Image above presents two confidence intervals. The wider one is used in the null hypothesis testing. It does not contain 0, therefore t-test results in a statistically significant outcome. As the narrow TOST confidence interval intersects with the SESOI region, we cannot claim there is a proof for substantially significant difference between two types of managers. This illustration is import as it shows that not every statistically significant result is also a substantially significant.

## Frequentist and Bayesian Regression Analysis

We will consider identical data. Researcher wants to prove now, that the effect is +0.2.

### Frequentist Model

In the classical (frequentist) regression, t-test of the regression coefficient is computed. Confidence interval can be supplied, too.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |  |
| (Intercept) | 5.0344 | 0.1295 | 38.88 | <0.01 | \*\*\* |
| TypeB | 0.4120 | 0.1831 | 2.25 | 0.0267 | \* |

Confidence interval for effect (TypeB) is approximately . Interpretation is not straightforward: this interval is one of many intervals of which 5 % of them contain the true size of the effect. Additional test of positive effect is based on the comparison of restricted model and fully saturated model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Resid. Df | RSS | Sum of SQ | F | p-val |
| Model 1 | 99 | 83.300 |  |  |  |
| Model 2 | 98 | 82.176 | 1.1237 | 1.34 | 0.2498 |

As the p-value is larger than the 0.05, we cannot reject the null. There is no evidence that the effect is different from 0.2.

### Bayesian Approach

Credible interval is an appropriate tool to assess the true effect. It has much simpler interpretation compared to the confidence interval. Another approach is to compute Bayes factor of the hypothesis. There are no test statistics and p-values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Est.Error | l-95% CI | u-95% CI |
| (Intercept) | 5.03 | 0.13 | 4.78 | 5.28 |
| TypeB | 0.42 | 0.18 | 0.06 | 0.77 |

Estimates of the regression coefficients are similar to the frequentist model. Last two columns of the regression output above are lower and upper limits of credible interval. Probability that the true size of the effect is within 0.06 and 0.77 is 0.95.

Whether the size effect is larger than 0.2 can be tested formally.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Hypothesis | Estimate | Est.Error | CI.Lower | Evid.Ratio |
| (TypeB)-(0.2) > 0 | 0.22 | 0.18 | -0.08 | 7.3 |

Evidence ratio is the Bayes factor. It can be concluded that the probability of effect being bigger than 0.2 is 7.3 times more likely to be true that it’s smaller then 0.2.s

# Time series modelling

Time series modelling is a wide topic which involves various sub-disciplines. From time-series visualization, decomposition and identification of various short and long-term patterns, testing structural shocks, prediction or causality modelling. There are several approaches to the time series modelling. Some of them are similar to standard cross-sectional regression analysis. Some of them differ in its nature as they exploit the auto-correlated nature of the data. This chapter is not a systematic introduction to the topic. Instead, it presents selected topics which should be known to an analyst before working with the time series.

## Correlation of time series

Correlation of time series is often an interest of analysis. Once it’s done on the data-stationary data, correlation coefficient is a good measure of association between the series. The problem is, however, with the inferential test. Let’s address the two problems.

### Problem of non-stationarity

Let’s generate two random series using R in the following way:

|  |
| --- |
| set.seed(123) # allows reproducibility |
| randomData1 <- rnorm(100) # first randomly generated series |
| randomData2 <- rnorm(100) # second randomly generated series |

Following plot shows that there is no correlation at all:



Image 12 Two random time series. Source: Own

This can be further tested using the Pearson correlation coefficient (p-value is incorrect as the next chapter explains).

|  |
| --- |
| cor.test(randomData1, randomData2) |
| Pearson's product-moment correlation  data: randomData1 and randomData2  t = -0.49095, df = 98, p-value = 0.6246  alternative hypothesis: true correlation is not equal to 0  95 percent confidence interval:  -0.2435805 0.1483291  sample estimates:  cor  -0.04953215 |

Correlation was estimated at level -0.0495. Now let’s assume that the third variable will influence both series in the same or similar way. This can be modelled just by adding a common trend:

|  |
| --- |
| ts0 <- 1 + 0.1\*seq(100) # blue line, no residuals  ts1 <- 1 + 0.1\*seq(100) + randomData1 # blue + random1data = red line  ts2 <- 1 + 0.1\*seq(100) + randomData2 # blue + random2data = black line |

The previous line of code generated regression line which represents the third variable. The new situation is depicted in the plot below:



Image 13 Two random time series with the common trend. Source: Own

If we run the same statistical test as above, we would get following result:

|  |
| --- |
| cor.test(ts1, ts2) |
| Pearson's product-moment correlation  data: ts1 and ts2  t = 21.085, df = 98, p-value < 2.2e-16  alternative hypothesis: true correlation is not equal to 0  95 percent confidence interval:  0.8620513 0.9353196  sample estimates:  cor  0.9051989 |

Now we have discovered a strong correlation between the two variables which we know that were randomly generated. This is a common problem called a spurious regression. This problem arises especially if the time series are non-stationary. There is one general solution which can be done in the several ways. The idea is to strip an effect of the common patterns. But how to identify such pattern? As a starting point can be considered de-trending the time series by removing deterministic trend by simple regression. More advanced approach is to pre-whitenise the time-series (white comes from the ideal form of the residuals called white noise). Prewhitenising is means that the model (popular option is to use a autoregressive model) which fits the first time series is then used on the second series. Residuals from both models are then being analyzed to discover a true correlation. Let’s continue with the previous example

|  |
| --- |
| library(forecast) # to allow using advanced function  library(tsa) |

The code above identifies the best autoregressive model based on minimal value of the BIC. Identified model has following properties:

|  |
| --- |
| selectARIMA <- auto.arima(ts1, seasonal=FALSE) # fits the autoregressive model  summary(selectARIMA)  Series: ts1  ARIMA(2,1,0) with drift  Coefficients:  ar1 ar2 drift  -0.6834 -0.4862 0.0977  s.e. 0.0891 0.0898 0.0474  sigma^2 estimated as 1.063: log likelihood=-142.38  AIC=292.77 AICc=293.19 BIC=303.15  Training set error measures:  ME RMSE MAE MPE MAPE MASE  Training set 0.008660619 1.010388 0.764617 -6.988041 20.19891 0.7634066  ACF1  Training set -0.06707982 |

ARIMA(2,1,0) with drift means that the time series is of the order 1 (differencing the time series makes it stationary). Current value can be regressed by lagged values up to two historical periods). This model is then used to clean whitenise the second time series. Correlation between the time series can be summarized by cross-correlation plot.



Image 14 Cross correlation of the time series. Plot is an outcome of the whitenisation and reveals estimated value of the correlation between analyzed time series. Source: Own

From the plot above can be read that the true correlation at lag 0, when the values of two time series are being compared at the same time is slightly negative.

### Correct measure of correlation

Although the computation of Pearson correlation coefficient is correct, its inferential conclusions are wrong. Pearson’s statistics assumes that the data come from the independent observations. This assumption is clearly violated when talking about the time series which is nothing but a series of points of the same variable ordered in time. Therefore, inferential test of the correlation significance is overly optimistic as the effective sample size is much smaller.

Correlation of the time series is usually done by cross-correlations (see figure 14). This plot describes a dynamic relation between the two series. In the center the lag 0 corresponds to the Pearson correlation. Lag 1 means that the second time series is moved one step ahead and the Pearson correlation is computed. Consider following example. The first time series depicts financial stimulus, such as and FDI inflow. The second variable is the unemployment in the corresponding month. If there is a negative value of the correlation in the lag 2, then we can claim, that there is a statistical evidence that increase of FDI is usually associated with the decrease of unemployment two months later.

## Stationarity of the time series

Time series can be described by descriptive statistics. Several moments (mean value, standard deviation, skewness, kurtosis, and other higher moments) can be computed but only some of them make sense if the time series is non-stationary. Variance is often used to analyse uncertainty of the studied phenomena. Higher variance traditionally means that the phenomena is more uncertain. Uncertain phenomena are difficult to analyses and predict. Look at the Figure 12 and 13. Time series in the latter figure are identical to the first one but contain the upward deterministic trend. If we would be asked to make a prediction, we should end up with the same degree of certainty (there is no uncertainty in the deterministic trend). Yet, the standard deviation of the time series in Figure 12 is almost three times higher than in Figure 13.

It is important to define a condition under which the modelling results in a sensible outcome (motivating example and interpretation of variance is just one of many other reasons). This condition is called a stationarity of the time series. If the time series has a constant mean value and standard deviation over time we say that the condition of the weak stationarity is met. Strict stationarity is met if all higher moments are constant throughout the development of the series. Two types of stationarity are being distinguished further:

1. Deterministic stationarity
2. Stochastic stationarity (unit root)

It is a general modelling strategy to work with the time series which are transformed from its original form to the stationary one. If removing trend (for example by employing the simple linear regression function) leads to the stationary time series, we call the time series deterministically stationary. In most of the economic time series removing deterministic trend does not help. Only transformation works in making the time series stationary. If taking the first or higher difference of the series helps, time series is called integrated time series of order first or higher order. In this case we talk about the stochastic stationarity, which is related to the term unit root. Unit root is a mathematical term which is difficult to elaborate without mathematical equations. For now it’s only important to know that if the time series has unit root, time series is exploding and does not have a constant mean value. This time series grows and error in the past has a higher importance for the future than the recent one. There are several approaches to tell whether the time series is stationary.

### Deterministic (trend) stationarity

Trend stationarity is the less frequent type of stationarity. KPSS test can be used to determine whether removing a deterministic trend will result in a stationary time series. Let’ apply the test on the data from Figure 12 whether they are stationary around its mean value:

|  |
| --- |
| > summary(ur.kpss(ts1, type=”mu”))  #######################  # KPSS Unit Root Test #  #######################  Test is of type: mu with 4 lags.  Value of test-statistic is: 2.0569  Critical value for a significance level of:  10pct 5pct 2.5pct 1pct  critical values 0.347 0.463 0.574 0.739 |

Because the test statistics 2.0569 is greater than the critical value 0.463 we can reject the null hypothesis that the time series is stationary. We can proceed further to test whether the time series is stationary around its trend value (whether removing the trend, not only the mean value) will results in stationary time series:

|  |
| --- |
| > summary(ur.kpss(ts1, type = c("tau")) )  #######################  # KPSS Unit Root Test #  #######################  Test is of type: tau with 4 lags.  Value of test-statistic is: 0.0591  Critical value for a significance level of:  10pct 5pct 2.5pct 1pct  critical values 0.119 0.146 0.176 0.216 |

Now the test statistics 0.0591 is lower than the critical value 0.146. We cannot reject the null hypothesis that the time series is stationary (detrending will results in stationary time series).

### Stochastic (unit root) stationarity

Unit root time series can become stationary by differencing time series D time. Several statistical tests were developed to test the time series. Among the most used one stand Augmented Dickey-Fuller test. This test has three settings:

1. Unit root only
2. Unit root with drift (constant)
3. Unit root and deterministic trend

Let’s create an artificial time series which contains both the trend and unit root.

|  |
| --- |
| set.seed(123)  err <- rnorm(100) # generates errors distributed according to normal dist.  trend <- seq(100) # index of values from 1 to 100  tsURtrend <- rep(0, 100) # prepares vector of values which will be computed  for(i in seq(2, 100)){ # iteratively computes values  tsURtrend[i] <- 0.5\*trend[i] + tsURtrend[i-1] + err[i]  } |

This time series has following growth pattern:



Image 15 Time series with both unit root and a trend. Time series is clearly non-stationary. Question remains whether taking differences will solve the problem. Source: Own

If we apply the KPSS test with trend:

|  |
| --- |
| > summary(ur.kpss(tsURtrend, type = c("tau")) )  #######################  # KPSS Unit Root Test #  #######################  Test is of type: tau with 4 lags.  Value of test-statistic is: 0.5172  Critical value for a significance level of:  10pct 5pct 2.5pct 1pct  critical values 0.119 0.146 0.176 0.216 |

Test statistics is bigger than the critical value. Therefore, after detrending time series will be non-stationary. At this point we should proceed to unit-root test to see what is causing the non-stationarity.

We will start with the simplest test of unit root only.

|  |
| --- |
| summary(ur.df(y=tsURtrend, type = "none"))  ###############################################  # Augmented Dickey-Fuller Test Unit Root Test #  ###############################################  Test regression none  Call:  lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)  Residuals:  Min 1Q Median 3Q Max  -3.0761 -0.7504 -0.0163 0.8614 3.5547  Coefficients:  Estimate Std. Error t value Pr(>|t|)  **z.lag.1 -0.0003800** 0.0004923 -0.772 0.442  z.diff.lag 1.0276416 0.0191604 53.634 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1.352 on 96 degrees of freedom  Multiple R-squared: 0.9979, Adjusted R-squared: 0.9979  F-statistic: 2.331e+04 on 2 and 96 DF, p-value: < 2.2e-16  Value of test-statistic is: -0.7718  Critical values for test statistics:  1pct 5pct 10pct  tau1 -2.6 -1.95 -1.61 |

Parameter z.lag.1 which equals to -0.0003800 is a center point of the analysis. If this parameter (in the literature called gamma) equals to 0, then the time series contains unit root. Unfortunately, the given the non-stationary nature of the data, p-value of the estimate is biased. Critical values which are displayed in the end of the outcomes shall be used instead. We need to compare value -0.7718 to -1.95. Null hypothesis states that gamma=0 (unit root, non-stationarity). Notice that the critical regions goes towards negative values (as p decreases). We therefore cannot reject the null and we believe that time series is non-stationary. We can test whether the non-stationarity is caused by present drift in the data.

|  |
| --- |
| > summary(ur.df(y=tsURtrend, type = "drift"))  ###############################################  # Augmented Dickey-Fuller Test Unit Root Test #  ###############################################  Test regression drift  Call:  lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)  Residuals:  Min 1Q Median 3Q Max  -2.7225 -0.8184 -0.1263 0.7330 3.5545  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 1.1575954 0.4189646 2.763 0.00688 \*\*  z.lag.1 0.0011866 0.0007404 1.603 0.11233  z.diff.lag 0.9338435 0.0386763 24.145 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1.307 on 95 degrees of freedom  Multiple R-squared: 0.9918, Adjusted R-squared: 0.9917  F-statistic: 5761 on 2 and 95 DF, p-value: < 2.2e-16  Value of test-statistic is: 1.6026 4.1355  Critical values for test statistics:  1pct 5pct 10pct  tau2 -3.51 -2.89 -2.58  phi1 6.70 4.71 3.86 |

Now we have two tests with the ADF test. The first test statistics is called tau2. This is an analogy to the first case. Because 1.6026 does not belong into the critical region, we cannot reject the null hypothesis (time series is non-stationary). Second test statistics phi1 tests jointly whether the gamma = drift = 0. As the value 4.1355 does not exceed 4.71 we cannot reject this hypothesis either. At this time the only suspect is the trend with combination of unit root.

|  |
| --- |
| > summary(ur.df(y=tsURtrend, type = "trend"))  ###############################################  # Augmented Dickey-Fuller Test Unit Root Test #  ###############################################  Test regression trend  Call:  lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)  Residuals:  Min 1Q Median 3Q Max  -2.42464 -0.60909 -0.00321 0.69992 2.13650  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 0.7947716 0.2981214 2.666 0.00904 \*\*  z.lag.1 0.0006863 0.0005253 1.306 0.19459  tt 0.5130963 0.0522186 9.826 4.31e-16 \*\*\*  z.diff.lag -0.0568259 0.1044551 -0.544 0.58771  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.923 on 94 degrees of freedom  Multiple R-squared: 0.996, Adjusted R-squared: 0.9958  F-statistic: 7736 on 3 and 94 DF, p-value: < 2.2e-16  Value of test-statistic is: 1.3064 37.7129 50.8503  Critical values for test statistics:  1pct 5pct 10pct  tau3 -4.04 -3.45 -3.15  phi2 6.50 4.88 4.16  phi3 8.73 6.49 5.47 |

Tau3 has the same meaning as before (result: there is unit root), phi2 tests whether gamma = drift = trend = 0. We can reject this hypothesis as 37.7129 > 4.88. We have identified that the either trend or drift are the cause. Let’s look at the phi3 test whose hypothesis is that gamma = trend = 0. We can reject this hypothesis, too (50.8503 > 5.47). This hypothesis clearly indicates that the cause of problem is the trend in the combination with the unit root. Detrending and first order differencing will stationaries the time series.

# references

Lakens, D., Scheel A. M. & Isager P. M. Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science* [online]. 2018, **1**(2), 259-269 DOI: 10.1177/2515245918770963.

Moomen, M., Rezapour, M., & Ksaibati, K. (2018). An investigation of influential factors of downgrade truck crashes: A logistic regression approach. *Journal of Traffic and Transportation Engineering (English Edition)*. [DOI> 10.1016/j.jtte.2018.03.005](https://doi.org/10.1016/j.jtte.2018.03.005)