Nils Karl Sørensen ©
Advanced Tools Statistics
Edition 2015

UgetStatz +

Advanced Tools Statistics

1. Introduction – Why These Notes?

UgetStatz + is the natural extension of the note set ***UgetStatz.*** The purpose of ***UgetStatz +*** is to dig deeper into especially the topic of regression and time series analysis. In addition, the issue of sampling is discussed, and it is practically shown how to draw a sample. Finally, ***UgetStatz +*** shows some simple methods for analyzing questionnaires ie. non-parametric methods and logistic regression. Most of the analysis can be undertaken by use of tne Analysis Toolpack in Excel or the add-in Megastat.

The material should be appropriate for a 5 ECTS course.

Nils Karl Sørensen

2. Table of Contents

Nils Karl Sørensen ©
Advanced Tools Statistics

Set 1: Model Selection, Autocorrelation and Tests

by Nils Karl Sørensen

Outline *page*

1. Integrated Statistical Modeling and the use of Regression

(BO Section 14.10 & 14.11)

Let us use our knowledge from the previous course in statistics to perform an integrated model sequence. After having outlined for example a macroeconomic theory for example for the money demand, the consumption function or the investment function we now want to perform a statistical investigation. Let us assume that we have found some statistics from the national statistical bureau. Then such an analysis is split into two parts namely the ***descriptive statistical part*** and *a* ***regression part***. These parts could contain:

Descriptive statistical part

Here the following should be considered:

- Set up some nice time series or cross-section plots
- Compute *descriptive statistics* and comment on the evolution of the data
- If we use cross-section data: Draw Box-plot(s) and comment on data
- If we use time series data: Look for special events (like the 2007 recession) and consider the issue of seasonality

Regression part

- *Set up a matrix of correlation.* Identify the variables with the highest correlation to *y* and comment. Discuss signs and relate to the prior from economic theory. Do we find what we expect? As an alternative we could calculate the *variance inflation factor* for variable *j* defined as $VIF_j = \frac{1}{1-R_j^2}$ where R_j^2 is coefficient of determination for the regression model that related x_j to all other independent variables *x*. If multicollinarity is present the VIF_j will be high (R_j^2 near one) and vice versa. This measure is provided by Megastat.
- Could multicollinarity be present (correlation among *x* variables).
- Based on an initial estimation of the full model a model selection is undertaken. During this process we should observe:
    - We attempt to obtain the simplest model with the highest R^2 coefficient.
    - We attempt to minimize the "standard error of regression" shown in the Excel or Megastat output.
    - We attempt to eliminate multicollinarity.
    - All t-statistics should be significant (p-value < 0.05).
- For the final model some selected highlights of the model control can be shown.

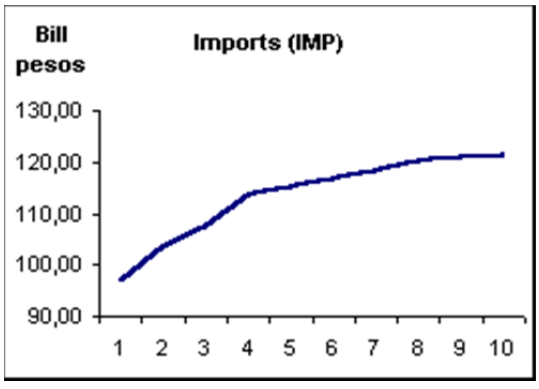
These things can all be undertaken by the use of Excel. Let us perform the regression part of this analysis on a small artificial data set. We want estimate an import *(IMP)* function for the artificial nation "Ruritania" for a 10 year period. We assume that the import depend on money supply *(MS)*, gross domestic product *(GDP)* the exchange rate of US dollar versus the local Peso *(USDvP)*, and finally the interest rate *(RENT)*. Below we find the statistics:

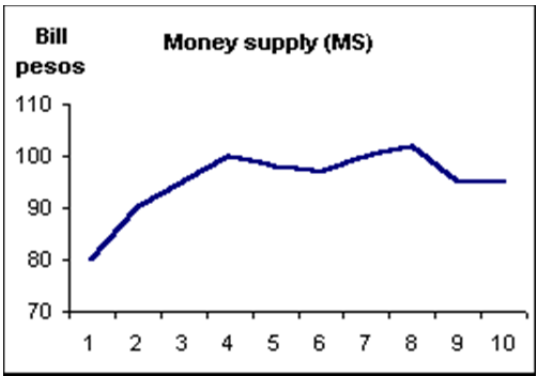
Data

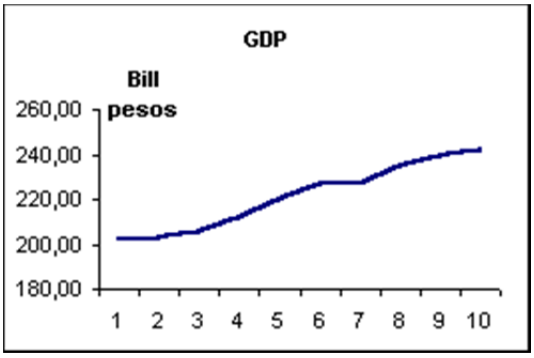
Year	*Imports (IMP) bill. pesos*	*Money supply (MS) bill. pesos*	*GDP* bill. pesos	*Exchange rate USD per Pesos*	*Interest rent (RENT)*
1	97.14	80	202.40	7.45	7.5
2	103.63	90	203.00	7.12	6.7
3	107.65	95	205.50	7.01	6.4
4	113.81	100	212.10	6.85	6.1
5	115.32	98	219.80	7.02	6.3
6	116.96	97	226.80	7.56	6.4
7	118.46	100	227.40	7.62	6.3
8	120.47	102	235.20	7.44	6.2
9	121.21	95	239.80	7.23	6.7
10	121.40	95	242.40	7.02	6.7

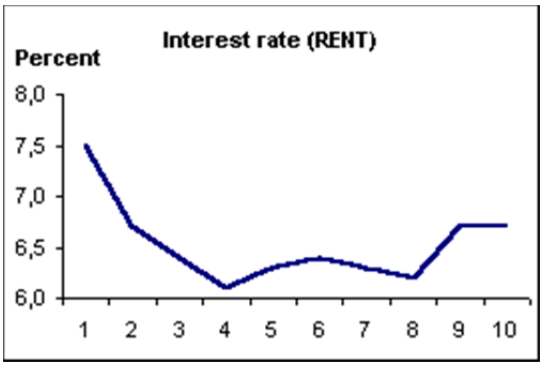
What should we expect from the theory of macroeconomics? When money supply increases, so do demand, so imports should increase. The same holds for GDP. If the interest rate decreases it will be cheaper to lent money. So a low interest rate should stimulate imports. Here we expect negative correlation. Correlation on exchange rate depends on the definition of the exchange rate. Here a low exchange rate should make imports cheaper. So here we expect a negative relation.

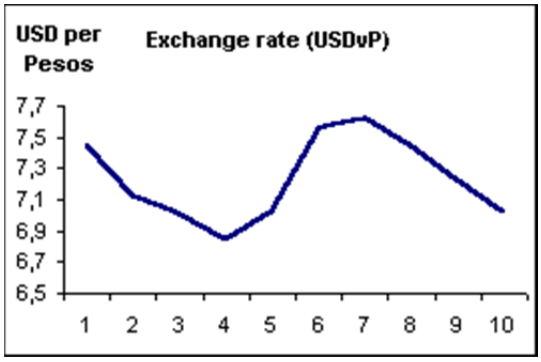
Let's look at some plots











From the plots we can observe that imports and GDP and money supply should be positively correlated. Further imports and the exchange rate as well as the interest rate should be negatively correlated. Notice that the exchange rate and the interest rate have a very similar pattern. If they are correlated with imports as well as with them self we observe a problem of *multicollinarity.* We want to estimate a model of the form:

$$IMP_t = \beta_0 + \beta_1 MS_t + \beta_2 GDP_t + \beta_3 USDrP_t + \beta_4 RENT_t + \varepsilon_t$$

Expected signs: (+) (+) (−) (−)

Let us first look at the matrix of correlation:

	IMP	***MS***	***GDP***	***USDvP***	***RENT***
y: Imports *(IMP)*	1.00				
x1: Money Supply *(MS)*	0.81	1.00			
x2: GDP	0.92	0.53	1.00		
x3: Exchange rate *(USDvP)*	0.06	−0.08	0.21	1.00	
x4: Interest rate *(RENT)*	−0.62	***−0.96***	−0.27	0.20	1.00

Notice, that many of our observations from the plots are confirmed. Besides from the exchange rate all variables are highly correlated with imports. Further, we were wrong with the correlation between the exchange rate and imports. We also observe a severe correlation between the money supply and the interest rate (−0.96). Also among GDP and money supply (0.53) multicollinarity is observed. Let us show the results from the estimation of the model:

Regression Statistics	
Multiple R	1.00
R-squared	0.99
Adjusted R-square	0.99
Standard Error	1.00
Observations	10

This is very high!

ANOVA

	df	*SS*	*MS*	*F-test*	*P-value*
Regression	4	604.62	151.16	152.00	0.00
Residual	5	4.97	0.99		
Sum	9	609.60			

	Coefficient	*Standard deviation*	*t-stat*	*P-value*	*Lower 95 %*	*Upper 95 %*
Constant	78.87	112.72	0.70	0.52	-210.89	368.62
X1: MS	-0.02	0.73	-0.03	0.98	-1.90	1.86
X2: GDP	0.45	0.09	4.96	0.00	0.21	0.68
X3: USDvP	-1.12	1.50	-0.75	0.49	-4.98	2.73
X4: RENT	-8.26	10.26	-0.81	0.46	-34.63	18.11

We have in order to save space omitted the residuals diagrams. The coefficient of determination is very high and from the ANOVA table it is observed that the F-test is significant. Consequently it is meaningful to estimate the model.

However the model is very poor! The only variable that is significant is GDP. All other variables are not significant. The money supply even takes the wrong sign!

In order to proceed we will try to eliminate the most severe problem of multicollinarity namely among the money supply and the interest rate. So we estimate the model without the interest rent. We exclude the interest rate because the money supply is higher correlated with the other variables than the interest rate.

We obtain the following output from Excel, and let us in this case include the residual analysis in the output.

Model without the Interest Rate

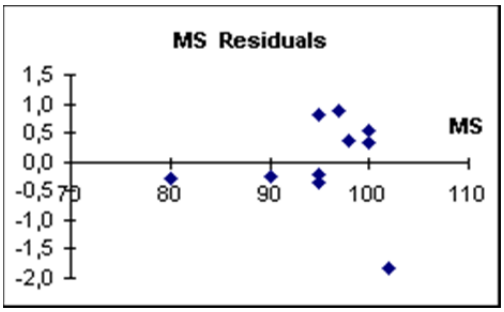
Regression Statistics	
Multiple R	1.00
R-squared	0.99
Adjusted R-square	0.99
Standard Error	0.97
Observations	10

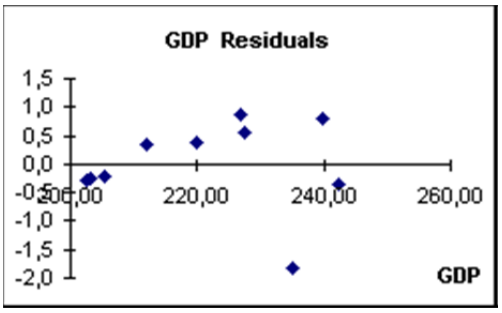
This is smaller than above

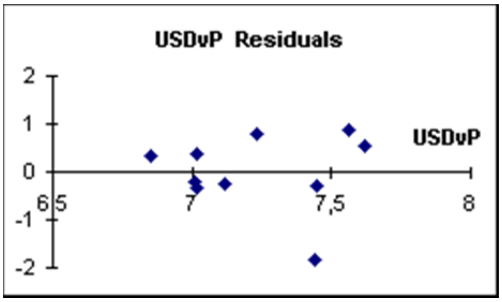
ANAVA

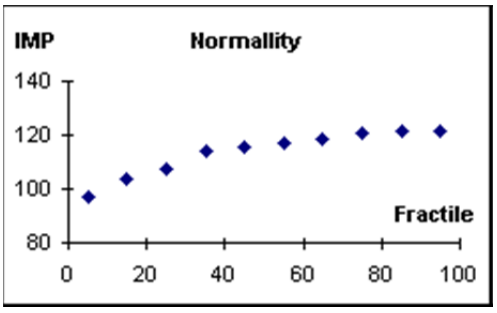
	df	*SS*	*MS*	*F-test*	*P-value*
Regression	3	603.98	201.33	215.04	0.00
Residual	6	5.62	0.94		
Sum	9	609.60			

	Coefficient	*Standard deviation*	*t-stat*	*P-value*	*Lower 95 %*	*Upper 95 %*
Constant	-11.52	10.33	-1.12	0.31	-36.80	13.75
X1: MS	0.57	0.06	9.18	0.00	0.42	0.72
X2: GDP	0.38	0.03	14.29	0.00	0.31	0.44
X3: USDvP	-1.72	1.27	-1.35	0.22	-4.82	1.38









Compared to the initial estimation we now observe the correct sign for the money supply. Notice that the size of the coefficient of the GDP-variable has remained quite constant. Both variables are now significant. The sign on the exchange rate variable is as expected, but it is not significant. This means that the model should be reestimated without this variable. This will due to low correlation with MS and GDP as seen from the matrix of correlation not affect these variables. Finally the analysis of residuals as well as the plot for normality looks quite satisfactory.

2. The Partial F-test and Variable Selection

In Section 1, we performed a selection of the included variables by inspection of the *p*-values only. However, when comparing models several alternatives is optional depending on the nature of the data.

First, the partial F-test is frequently used to test the significance of a set of independent variables in a regression model. We use this F-test to test the significance of a portion of a regression model.

We will present the partial F-test, using a little model example. Suppose that we are considering the following models:

Full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

Reduced model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

By comparing the two models, we are asking the question: Given that variables x_1 and x_2 are already in the regression model, would we be gaining anything by adding x_3 and x_4 to the model? Will the reduced model be improved in terms of its predictive power by the addition of the two variables x_3 and x_4.

Let us consider this issue by setting a test of a hypothesis. The null hypothesis that the two variables x_3 and x_4 additional value once x_1 and x_2 are in the regression model. The alternative hypothesis that the two lope coefficients are not both zero. The hypothesis test is stated as:

H_0: $\beta_3 = \beta_4 = 0$ (given that x_1 and x_2 are in the model)
H_1: β_3 and β_4 are not both zero

The test statistic for this hypothesis is the partial *F*-statistic given as:

$$F_{[r\,;\,n-(k+1)]} = \frac{(SSE_R - SSE_F)/r}{MSE_F}$$

Where SSE_R is the sum of squares for error of the reduced model; SSE_F is the sum of squares for error of the full model; MSE_F is the mean square error of the full model; $MSE_F = SSE_F/[n-(k+1)]$; k is the number of independent variables in the full model (k=4 in the case above); and r is the number of variables dropped from the full model in creating the reduced model (in the present case $r=2$).

The difference $SSE_R - SSE_F$ is called the extra sum of squares associated with the reduced model. Since this additional sum of squares for error is due to r variables, it has r degrees of

freedom. In our model selection case in the previous section we had that $SSE_R = 5.62$, $SSE_F = 4.97$ and $MSE_F = 0.99$. Initial $k=4$ so $r=1$. With $n = 10$ then:

$$F_{[1;10-(4+1)]} = \frac{(5.62 - 4.97)/1}{0.99} = 0.66$$

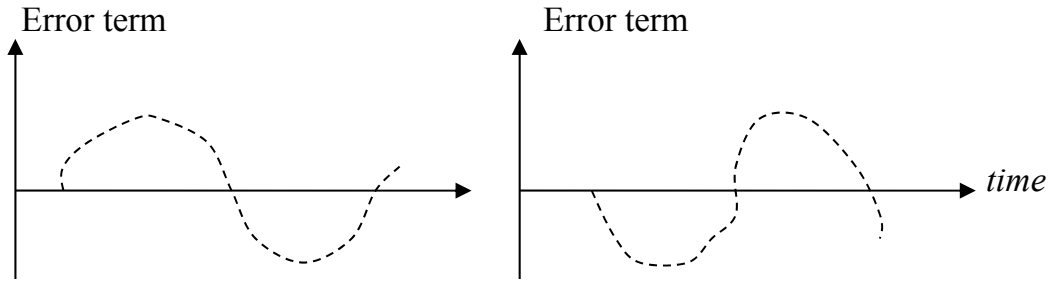
Assuming α=0.05 we find $F_{[1;5]} = 6.61$. So we cannot reject H_0. This is also what we should expect. The interest rate is namely not significant in the initial model, and should consequently be excluded.

The partial F-test is especially good in situations when working with cross-section models. In cases with times series data other problems occur as will be shown in the next Section.

3. Analyzing Autocorrelation and the Durbin-Watson Test

(BO Section 13.8)

Autocorrelation occurs in non-stationary **time series**[1] where the variables are dependent in time. Autocorrelation may be either positive or negative of nature. Examples are given below:



Positive autocorrelation *Negative autocorrelation*

We test for autocorrelation be setting up the *Durbin-Watson* test. We calculate the tester:

$$DW = \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}$$

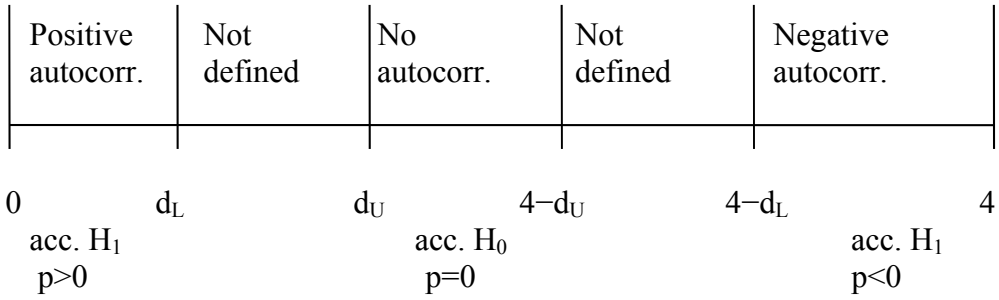
[1] Notice, that the test for autocorrelation only has a meaning, when we work with time series data, and NOT when we work with cross-section data. For example, in the latter case, data may be listed in alphabetic order. If we for example applied the test on regional statistics for Germany, performing the test would imply that Mainz and Munich would be directly related although the distance between the two cities is several hundred kilometers.

This expression is based on the estimation of the regression: $\varepsilon_t = p\varepsilon_{t-1} + v_t$ where the last term is "the error term of the errors". We can state the hypothesis as:

H_0: The error terms are not autocorrelated $(p = 0)$
H_1: The error terms are autocorrelated $(p \neq 0)$

The Durbin-Watson test is a two-sided test, where the alternative hypothesis (H_1) is not defined consistently. This is so because under H_0 the assumption to the error term is by itself not fulfilled. This is exactly what we want to test for!

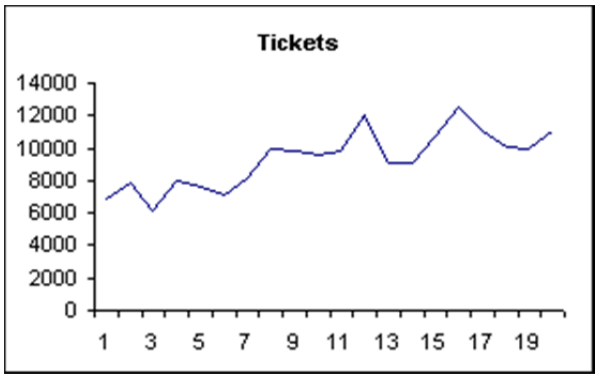
The distribution for the Durbin-Watson test is non-standard and found in Appendix I at the end of these notes or in Bowerman, Appendix A, tables A.11-A.13. k is the number of explanatory variables (the number of X's). There are two critical values to be found named d_L and d_U. The range of the critical value is between 0 and 4. The interpretation can be summarized in the following figure:

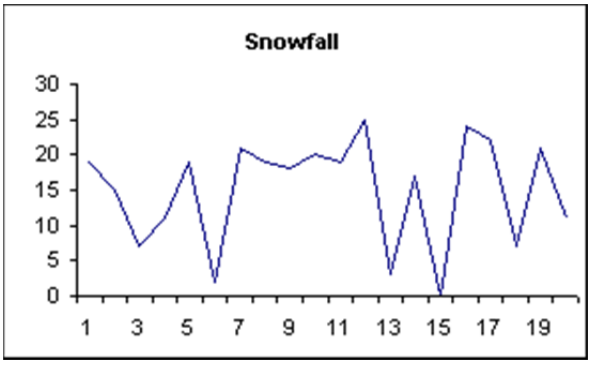


Example

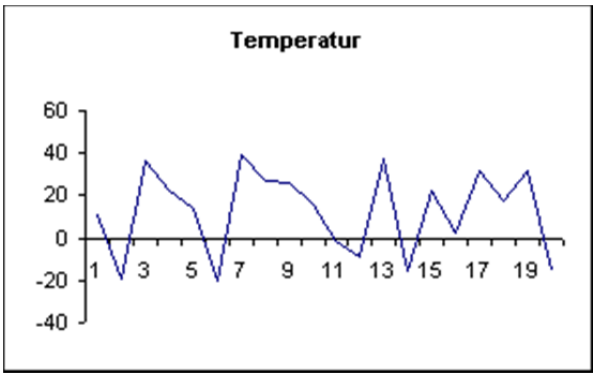
A ski resort wants to determine the effect that the weather have on the sales of lift tickets during the Christmas week. Weekly sales of ski lifts tickets *(y)* are assumed to depend on total snowfall in inches (x_1) and the average temperature in Fahrenheit (x_2). For a data set ranging over 20 years we obtain:

Tickets (y)	Snowfall (x_1)	Temperature (x_2)
6835	19	11
7870	15	-19
6173	7	36
7979	11	22
7639	19	14
7167	2	-20
8094	21	39
9903	19	27
9788	18	26
9557	20	16
9784	19	-1
12075	25	-9
9128	3	37
9047	17	-15





10631	0	22
12563	24	2
11012	22	32
10041	7	18
9929	21	32
11091	11	-15



From the plots it is observed that the relation among the variables not is optimal. So we do not expect the most significant result. This is also confirmed by the matrix of correlation shown below. We obtain from Excel by use of the command: ***Tools/data analysis/correlation***:

	Tickets	*Snowfall*	*Temperature*
Y: Tickets	1.00		
X1: Snowfall	0.33	1.00	
X2: Temperature	-0.11	-0.02	1.00

Let us now estimate a model of the form:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t \qquad \text{where } t = 1,2,\ldots,20$$

From Excel we obtain:

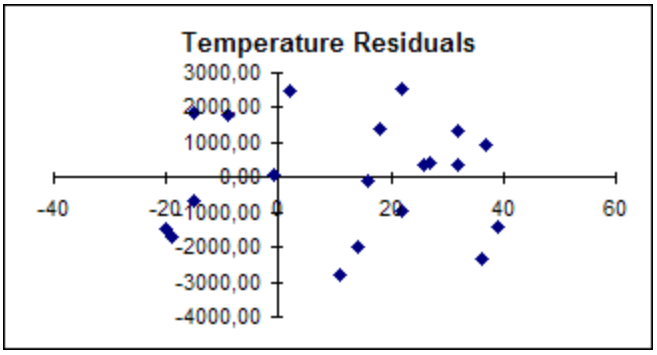
Regression Statistics	
Multiple R	0.35
R square	0.12
Adjusted R square	0.02
Standard Error	1711.68
Observations	20

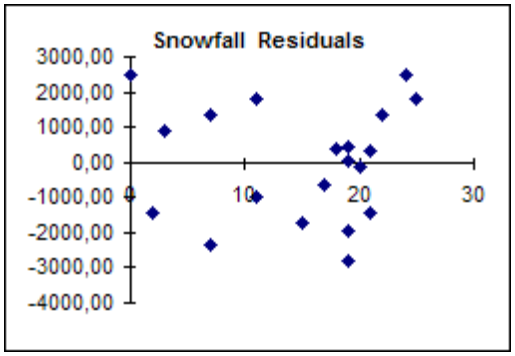
This is a poor correlation

ANOVA

	df	***SS***	***MS***	***F-test***	***P-value***
Regression	2	6793798.5	3396899.1	1.16	0.34
Residual	17	49807214	2929836.1		
Sum	19	56601012.2			

	Coefficient	***Standard Deviation***	***t-stat***	***P-value***	***Lower 95%***	***Upper 95%***
Constant	8308.01	903.73	9.19	0.00	6401.31	10214.71
X1: Snowfall	74.59	51.57	1.45	0.17	-34.22	183.41
X2: Temperature	-8.75	19.70	-0.44	0.66	-50.33	32.82





This is not a very good result, and much worse than our expectations from the plots! The F-test is not significant ($p = 0.34 > 0.10$), so the overall model is not significant. Further, only the constant term is significant. It looks like that neither snowfall nor temperature has an influence on the sales of tickets.

The plots of residuals are also not very nice! Both plots reveal some kind of systematic behavior. Let us perform the Durbin-Watson test first by calculation of the formula given above in Excel. We then obtain:

ε_t	ε_{t-1}	$(\varepsilon_t - \varepsilon_{t-1})^2$	$(\varepsilon_t)^2$
-2793.99			7806391.51
-1723.23	-2793.99	1146528.83	2969525.68
-2342.03	-1723.23	382911.49	5485102.65
-956.95	-2342.03	1918431.85	915762.73
-1963.73	-956.95	1013597.71	3856238.75
-1465.27	-1963.73	248460.53	2147024.00
-1439.07	-1465.27	686.38	2070933.63
414.07	-1439.07	3434133.96	171452.12
364.91	414.07	2416.75	133157.32
-102.82	364.91	218765.62	10571.25
49.96	-102.82	23341.64	2496.31
1823.37	49.96	3144985.16	3324691.68
920.10	1823.37	815908.57	846578.74
-660.40	920.10	2497979.80	436131.76
2515.57	-660.40	10086807.91	6328096.53
2482.26	2515.57	1109.74	6161605.15
1343.06	2482.26	1297779.75	1803801.32
1368.40	1343.06	642.44	1872527.09
334.65	1368.40	1068645.60	111990.59
1831.16	334.65	2239532.65	3353135.13
	Sum	29542666.38	49807213.95

$$DW = \frac{29542666.38}{49807213.95} = 0.59$$

It is assumed that $n=20$ and $k=2$.
With a level of significance equal to 0.05, we have from the critical values in Appendix I that

$D_L = 1.10$ $D_U = 1.54$

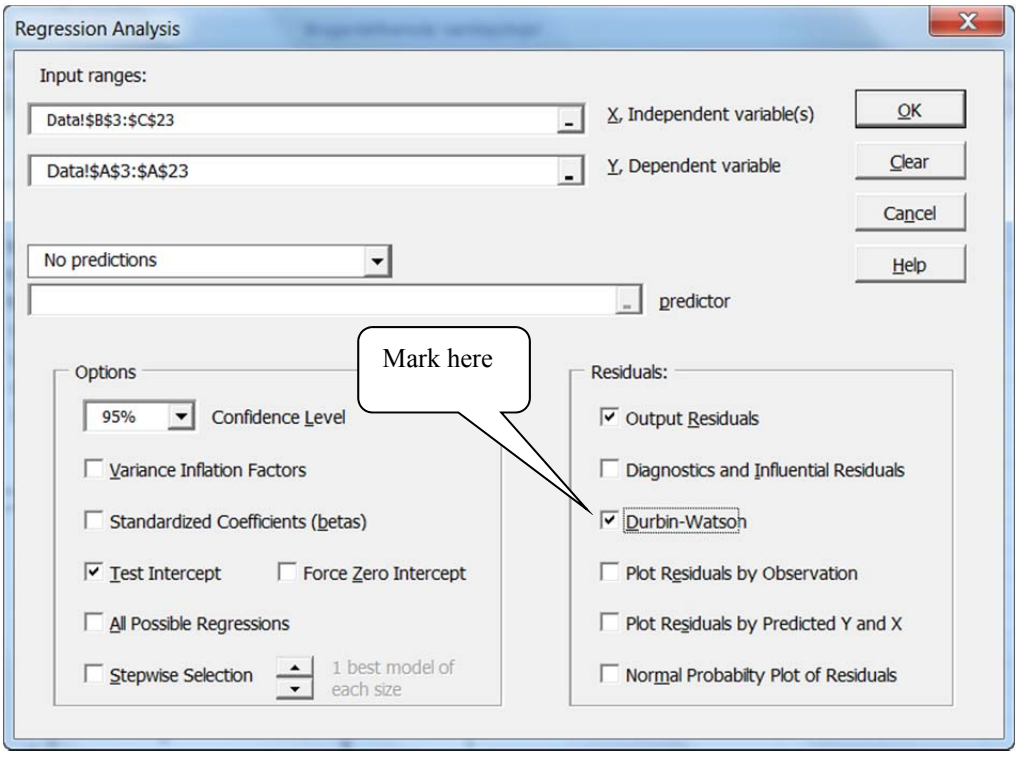
Hypothesis:

H_0: No first order autocorrelation

H_1: Positive first order autocorrelation

As DW < D_L H_1 is accepted.

Alternatively, we can find the DW-value by use of Megastat. Here the test is much easier to perform. In the menu for *regression,* a label with the text *Durbin-Watson* can be found. Just mark the label, and the test will be performed. The menu looks as:



We then find that positive autocorrelation is present. How do we solve the problem? A solution could be to include a ***positive linear trend***. This is a variable taking the values $T = 1,2,..,20$. It is a strongly positive variable. We must from the plots expect this variable to be strongly correlated with the sales of tickets as this has a positive trend.

The result with inclusion of a positive linear trend from Excel is:

Regression Statistics		
Multiple R	0.86	This is has increased a lot!
R square	0.74	
Adjusted R square	0.69	
Standard Error	957.24	This has decreased a lot!
Observations	20	

ANOVA

	df	*SS*	*MS*	*F-test*	*P-value*	
Regression	3	41940217.4	13980072.5	15.26	0.00	Significant!
Residual	16	14660794.8	916299.676			
Sum	19	56601012.2				

	Coefficient	*Standard deviation*	*t-stat*	*P-value*	*Lower 95 %*	*Upper 95 %*
Constant	5965.59	631.25	9.45	0.00	4627.39	7303.78
X1: Snowfall	70.18	28.85	2.43	0.03	9.02	131.35
X2: Temperature	-9.23	11.02	-0.84	0.41	-32.59	14.13
X3: Trend	229.97	37.13	6.19	0.00	151.25	308.69

Compared to the first regression an improvement can be observed. Snowfall is now significant, but temperature has no effect on the model, and should be excluded. Further the coefficient of determination has increased and the standard error has decreased. Consequently this is the model to be preferred.

For this model the Durbin-Watson test can also be undertaken. This will result in a DW-value equal to 1.88. Now $k=3$ because the trend is included. The critical values can again be found be use for the appendix. In this case $d_L=0.998$ and $d_U=1.676$. As $1.676<1.88$ no autocorrelation is observed. The inclusion of the trend variable eliminates the presence of autocorrelation.

4. More General Tests for Model Selection

The Durbin-Watson test for autocorrelation can be criticized in several ways. First, it is only possible to examine for first order autocorrelation. For example, it is not possible to consider for example in influence of ε_{t-4}. That is the residual error 4 periods ago. Second, the Durbin-Watson statistic is not defined for some outcomes of the value of the DW-tester. Third, the alternative hypothesis is not properly tested (this is actually due to the second critic).

In Section 2, we considered the F-test being a little bit similar to the Durbin-Watson test for autocorrelation. Here we compared two models, and used the F-statistic to determine the significance of an improvement of the initial model. The test presented in Section 2 is intuitive more appealing that the Durbin-Watson test, because the F-statistic has a well

defined statistical distribution. However, there is still the problem with the definition of the alternative hypothesis.

The *multiplier tests* is a class of tests that seeks to solve the problems above by use of the methods outlined in Section 2; i.e. the F-test and the Durbin-Watson test outlined in Section 3.

There are three different multiplier tests namely the Wald test, Likelihood Ratio (LR) test and the Lagrange Multiplier (LM) test. In principle, the tests investigate the same hypotheses, but the theoretical point of departure is different for each test.

The tests are most easily explored by an example taken from macroeconomics. Consider an import function estimated on time series data. Let *M* denote imports and let *Y* be GDP. Further *t* is time. The time period considered runs from *1* to *T*. The model can be written as:

$$M_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t \qquad \text{where } t = 1,2,...,T$$

The model states that imports in the present period depend on the level of income in the last period. We now want to examine if the level of investment *I* in the last period also should have been included in the model. The following hypotheses can be stated:

H_0:	$M_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_{1t}$	where $t = 1,2,...,T$	(Model 0)
H_1:	$M_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 I_{t-1} + \varepsilon_{2t}$	where $t = 1,2,...,T$	(Model 1)

Further *m* is the number of additional variables included in the model under the alternative hypothesis. The three tests can be stated as:

Wald Test

Tester $$\xi_W = T \times \left(\frac{SSE_0 - SSE_1}{SSE_1} \right) = \chi^2_m$$

LR Test

Tester $$\xi_{LR} = T \times \log\left(\frac{SSE_0}{SSE_1} \right) = \chi^2_m$$

LM Test

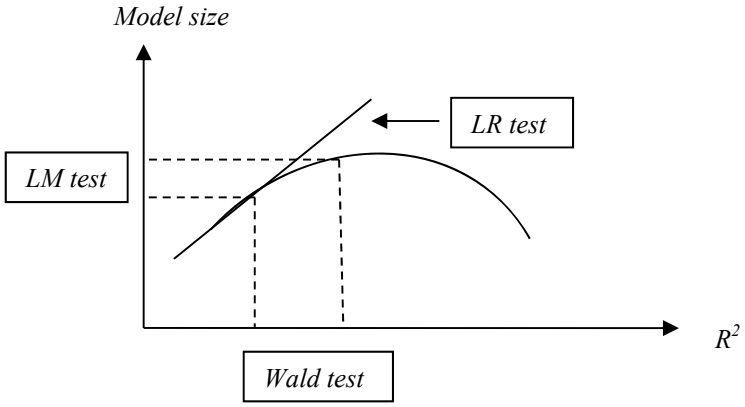
Tester $$\xi_{LM} = T \times \left(\frac{SSE_0 - SSE_1}{SSE_0} \right) = \chi^2_m$$

A nice thing with these tests is fact that they follow the chi-squared distribution. So the tests are in the same class of tests as the distribution free *goodness of fit tests* outlined in Bowerman Chapter 12.

The tests look very similar. What is the difference? Remember that the coefficient of determination is equal to

$$R^2 : 1 - \frac{SSE}{SST}$$

For each of the models the R^2 can be calculated. Then the distribution of the R^2 for the different models can be displayed in a diagram. At some point the model will be over fitted and R^2 will start to decrease. The distribution of R^2 can then be displayed as a quadratic function



The only difference between the Wald and the LM test is the reference. For the Wald test it is the alternative whereas it for the LM test is the null hypothesis. The LR test is a little different taking its point of departure in the model size analysis.

For all tests it is true that they are small sample tests. Further, it is generally true that $\xi_W \geq \xi_{LM} \geq \xi_{LR}$.

The LM-test as a Diagnostic Test

The LM-test can be used to reveal autocorrelation in a more general way. This test is much more flexible than the Durbin-Watson test, and the chi-square distribution is better defined than the Durbin-Watson statistic.

Assume from the example above that we have accepted that the lagged value of the investments should be included in the functional form of the import function. Let this model be valid under H_0.

H_0: $M_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 I_{t-1} + \varepsilon_t$ where $t = 1,2,...,T$ (Model 0)

When we estimate this model the residuals may not be white noise. For example, it could be that the residuals follow a process looking as:

AR(4): $\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \rho_3 \varepsilon_{t-3} + \rho_4 \varepsilon_{t-4} + v_t$

This process is an extension of the process considered in Section 3. Instead of a single lag there are now four lags. The term v_t is the "error of the error" term. This process is called an *autoregressive process of order 4.* The model states that the errors are related 4 periods back in the past. This process can be generalized as:

AR(P): $\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \rho_3 \varepsilon_{t-3} + ... + \rho_p \varepsilon_{t-p} + v_t$

We can incorporate the AR(4) process in our model and test if it is being improved. The import function now looks as:

H_1: $M_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 I_{t-1} + \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \rho_3 \varepsilon_{t-3} + \rho_4 \varepsilon_{t-4} + v_t$ (Model 1)

The test is now undertaken the following way:

- Estimate the model under the null (H_0) and save the residuals → SSE_0
- Use the residuals and estimate the model under the alternative → SSE_1
- In the present case the model is extended with four lags so $m=4$

Perform now:

Tester $\xi_{LM} = T \times \left(\frac{SSE_0 - SSE_1}{SSE_0} \right) = \chi^2_m$

Example

In Appendix II statistics are given for an annual data set on US national accounts ranging over the period from 1929 to 1972. Using these statistics an imports function of the form given above can be estimated. The Excel output for the model estimated under H_0 is given as:

Model H_0

Regression Statistics	
Multiple R	0.949
R-squared	0.901
Adjusted R-square	0.899
Standard Error	4.409
Observations	43

ANOVA

	df	*SS*	*MS*	*F-test*	*P-value*
Regression	1	7282.54	7282.54	374.59	0.00
Residual	41	797.10	19.44		**SSE_0=797.1**
Sum	42	8079.64			

	Coef	*St error*	*t-stat*	*P-value*	*low 95%*	*high 95%*
Constant	-8.77	1.63	-5.38	0.00	-12.06	-5.47
Y-lag	0.07	0.00	19.35	0.00	0.07	0.08

Regression statistics are in order, and the coefficients are significant. The extended model under H_1 is estimated with the following result:

Model H_1

Regression Statistics	
Multiple R	0.950
R-squared	0.902
Adjusted R-square	0.897
Standard Error	4.449
Observations	43

ANOVA

	df	*SS*	*MS*	*F-test*	*P-value*
Regression	2	7287.78	3643.89	184.07	0.00
Residual	40	791.86	19.80		**SSE_1=791.9**
Sum	42	8079.64			

	Coef	*St error*	*t-stat*	*P-value*	*low 95%*	*high 95%*
Constant	-9.07	1.75	-5.19	0.00	-12.61	-5.54
Y-lag	0.08	0.01	7.59	0.00	0.06	0.10
I-lag	-0.03	0.06	-0.51	0.61	-0.14	0.09

Although the overall regression statistics has improved, the performance of the lagged investment coefficient is not satisfactory. First, the coefficient is not significant, and second, the coefficient takes the wrong sign. We expect that an increase in investment will cause imports to rise.

How are the 3 tests performing? First, note that the model is expended by a single parameter, so *m=1*. In addition, the number of observation is equal to *T=43*. Next, the three testers are being set up.

Wald Test

Tester $$\xi_W = T \times \left(\frac{SSE_0 - SSE_1}{SSE_1} \right) = 43 \times \left(\frac{797.1 - 791.9}{791.9} \right) = 0.2823 \approx \chi_1^2$$

LR Test

Tester $$\xi_{LR} = T \times \log\left(\frac{SSE_0}{SSE_1} \right) = 43 \times \log\left(\frac{797.1}{791.9} \right) = 0.1222 \approx \chi_1^2$$

LM Test

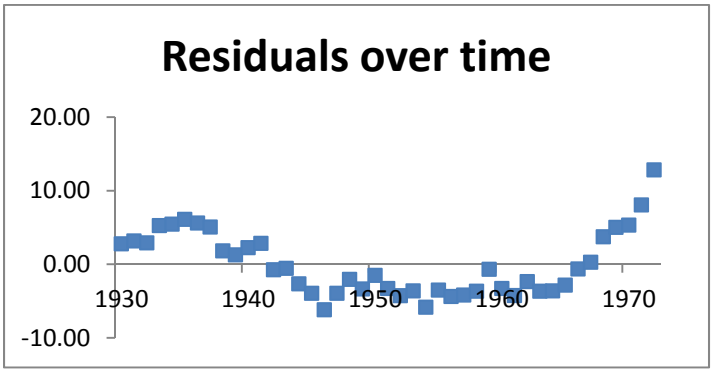
Tester $$\xi_{LM} = T \times \left(\frac{SSE_0 - SSE_1}{SSE_0} \right) = 43 \times \left(\frac{797.1 - 791.9}{797.1} \right) = 0.2805 \approx \chi_1^2$$

Comparison of the size of the testers reveals that $\xi_W \geq \xi_{LM} \geq \xi_{LR} \rightarrow 0.2823 \geq 0.2805 \geq 0.1222$ as expected.

At the 95 % (α=0.05) level we find that $\chi_1^2 = 3.84$. As 3.84 (>) is larger than all the testers *H0 is accepted.*

The conclusion is that the outcome of all the tests is consistent with the finding above. The lagged investments should not be included in the model.

Let us finally consider the case where the LM test is used as a diagnostic test. We have already estimated the model under the null. Now we estimate the model with 4 additional lags of the errors. The model is shown on page 16 mid. The residuals under the null look as:



So autocorrelation of some degree could be present!

The model is a little bit special to set up. The residuals are estimated under H_0 and then the residuals are lagged as shown below. If the residuals not are lagged an additional period the estimation will break down.

Year	M	Y_{t-1}	I_{t-1}	ε_{t-1}	ε_{t-2}	ε_{t-3}	ε_{t-4}
1934	7.1	141.5	5.3	4.97	2.85	3.35	3.25
1935	8.7	154.3	9.4	5.20	4.97	2.85	3.35
1936	9.3	169.3	18.0	5.92	5.20	4.97	2.85
1937	10.5	193.0	24.0	5.59	5.92	5.20	4.97
1938	…	…	…	…	…	…	…
1939	…	…	…	…	…	…	…

The following result is obtained under H_1:

Model H_1

Regression Statistics	
Multiple R	0.99
R-squared	0.99
Adjusted R-square	0.98
Standard Error	1.45
Observations	39

ANOVA

	df	*SS*	*MS*	*F-test*	*P-value*
Regression	6	4872.48	812.08	384.89	0.00
Residual	32	67.52	2.11		**SSE_1 = 67.5**
Sum	38	4940.00			

	Coef	*St error*	*t-stat*	*P-value*	*Low 95%*	*high 95%*
Constant	1928.02	0.70	2747.32	0.00	1926.59	1929.45
Y-lag	0.05	0.00	11.95	0.00	0.04	0.05
I-lag	0.10	0.02	4.05	0.00	0.05	0.15
E-l1	-0.29	0.16	-1.87	0.07	-0.61	0.03
E-l2	0.04	0.20	0.19	0.85	-0.36	0.44
E-l3	0.17	0.20	0.83	0.41	-0.25	0.58
E-l4	-0.27	0.16	-1.72	0.10	-0.59	0.05

The LM-tester is now:

Tester $$\xi_{LM} = T \times \left(\frac{SSE_0 - SSE_1}{SSE_0} \right) = 39 \times \left(\frac{791.7 - 67.5}{791.7} \right) = 35.68 \approx \chi_4^2$$

At the 95 % (α=0.05) level we find that $\chi_4^2 = 9.49$. As 9.49 is smaller than 35.68 *H_1 is accepted.* So the model above is the best, and autocorrelation of order four is present in the initial model.

The interpretation of the outcome of the test should be taken carefully! Inspection of the Excel output reveals that lag 1 is weak significant only, and lag 4 is on the margin to be weak significant. This suggests that the model is reestimated with these two lags only. This is seen already at the initial estimation from the P-values.

This underlines that these test in general has weak performance compared to the initial modeling sequence with inspection of the P-values.

Reference

Engle, R. F., 1982, *Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics.* In: Griliches and Intrilligatior (editors) *Handbook of Econometrics,* North-Holland.

Appendix I: Critical Points for the Durbin-Watson Test Statistic

95 % level ($\alpha = 0.05$)

n	k = 1		k = 2		k = 3		k = 4		k = 5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.316	2.645
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.379	2.506
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.445	2.390
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825
32	1.373	1.502	1.309	1.574	1.244	1.650	1.172	1.732	1.109	1.819
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820

Source:

Durbin, J. and G. S. Watson, 1951, *Testing for Serial Correlation in Least Squares Regression,* Biometrika 30, pp. 158−178.

Appendix II: US National Accounts 1929 to 1972

Year	Total Gross National Product	Disposable Income	Personal Consumption Expenditures	Gross Private Domestic Investment	Exports	Imports	Government Purchases of Goods and Services
1929	203.6	150.6	139.6	40.4	11.8	10.3	22.0
1930	183.5	139.0	130.4	27.4	10.4	9.0	24.3
1931	169.3	133.7	126.1	16.8	8.9	7.9	25.4
1932	144.2	115.1	114.8	4.7	7.1	6.6	24.2
1933	141.5	112.2	112.8	5.3	7.1	7.1	23.3
1934	154.3	120.4	113.1	9.4	7.3	7.1	26.6
1935	169.5	131.8	125.5	18.0	7.7	8.7	27.0
1936	193.0	148.4	138.4	24.0	8.2	9.3	31.6
1937	203.2	153.1	143.1	29.9	9.8	10.5	30.8
1938	192.9	143.6	140.2	17.0	9.9	8.0	33.9
1939	209.4	155.9	148.2	24.7	10.0	8.7	35.2
1940	227.2	166.3	155.7	33.0	11.0	8.9	36.4
1941	263.7	190.3	165.4	41.6	11.2	10.8	56.3
1942	297.8	213.4	161.4	21.4	7.8	9.9	117.1
1943	337.1	222.8	165.8	12.7	6.8	12.6	164.4
1944	361.3	231.6	171.4	14.0	7.6	13.4	181.7
1945	355.2	229.7	183.0	19.6	10.2	13.9	155.4
1946	312.6	227.0	203.5	52.3	19.6	11.2	48.4
1947	309.9	218.0	206.3	51.5	22.6	10.3	39.9
1948	323.7	229.8	210.8	60.4	18.1	12.0	46.3
1949	324.1	230.8	216.5	48.0	18.1	11.7	53.3
1950	355.3	249.6	230.5	69.3	16.3	13.6	52.8
1951	383.4	255.7	232.8	70.0	19.3	14.1	75.4
1952	395.1	263.3	239.4	60.5	18.2	15.2	92.1
1953	412.8	275.4	250.8	61.2	17.8	16.7	99.8
1954	407.0	278.3	255.7	59.4	18.8	15.8	88.9
1955	438.0	296.7	274.2	75.4	20.9	17.7	85.2
1956	446.1	309.3	281.4	74.3	24.2	19.1	85.3
1957	452.5	315.8	289.2	68.8	26.2	19.9	89.3
1958	447.3	318.8	290.1	60.9	23.1	20.9	94.2
1959	475.9	333.0	307.3	73.6	23.8	23.5	94.7
1960	487.7	340.2	316.1	72.4	27.3	23.0	94.9
1961	497.2	350.7	322.5	69.0	28.0	22.9	100.5
1962	529.8	357.3	338.4	79.4	30.0	25.5	107.5
1963	551.0	381.3	353.3	82.5	32.1	26.6	109.6
1964	581.1	407.9	373.7	87.8	36.5	28.2	111.2
1965	617.8	435.0	397.7	99.2	37.4	31.2	114.7
1966	658.1	458.9	418.1	109.3	40.2	36.1	126.5
1967	675.2	477.5	430.1	101.2	42.1	38.5	140.2
1968	706.6	499.0	452.7	105.2	45.7	44.7	147.7
1969	725.6	513.6	469.1	110.5	48.4	48.3	145.9
1970	722.1	533.2	477.0	104.0	52.2	50.0	139.0
1971	741.7	554.7	495.4	108.6	52.6	52.5	137.6
1972	789.7	578.7	524.8	123.8	56.9	58.7	142.9

Source:
Michael Lovell, M, 1975, *Macroeconomics, Measurement, Theory, and Policy,* Wiley.

Nils Karl Sørensen ©
Advanced Tools Statistics

Set 2: Transformation of Linear Models

by Nils Karl Sørensen

Outline *page*

1. Non-linear Models

In some situations it is not possible to model a linear relationship among the dependent variables *y* and one or more of the independent *x* variables. The plot of residuals will continue to exhibit some kind of curvature. In such a case we can set up either a polynomial model or some kind nonlinear or transformed model.

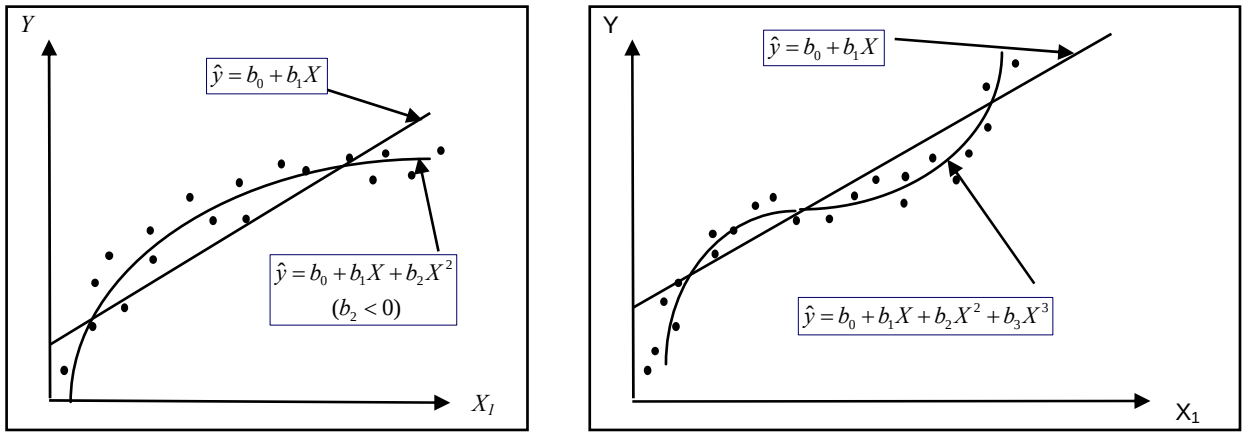
1.1. Polynomial Models

The one-variable polynomial regression or quadratic model is given by:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_m x^m + \varepsilon$$

where *m* is the *degree* of the polynomial - the highest power of X appearing in the equation. The degree of the polynomial is the ***order*** of the model. The model is easily estimated by Excel by transformation of *x*. Some forms of this model are given in the illustration below:

Illustrations of Models of Polynomial Order Two and Three

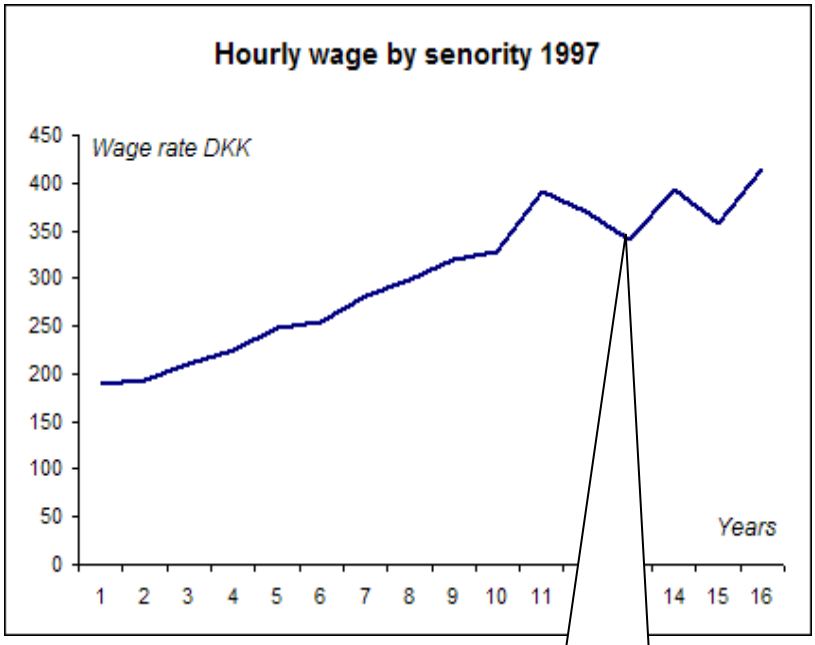


1.2. Simple Types of Non-Linear Models Estimated by Excel

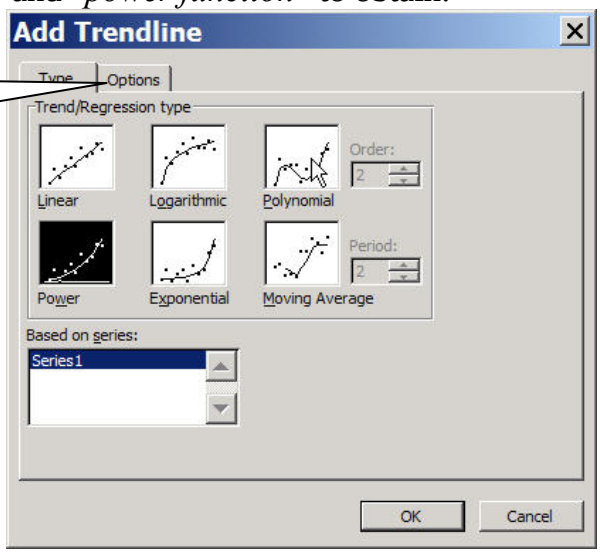
The “add trend line function” in Excel lists several possible transformations for the simple regression model. Let us consider an example.

The table on the next page shows a relation between the average hourly wages obtained by Masters of Economics by seniority, i.e. the number of years since graduation. As visible a non linear relation is present. A decreasing behavior in time is observed. What kind of relation should be used in order to estimate this model properly?

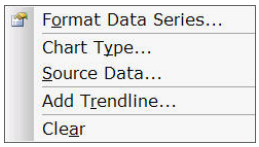
Senority	*Wage DKK*	*Ln(senority)*	*Ln(Wage)*
1	*189*	*0,000*	*5,242*
2	*194*	*0,693*	*5,268*
3	*210*	*1,099*	*5,347*
4	*225*	*1,386*	*5,416*
5	*248*	*1,609*	*5,513*
6	*255*	*1,792*	*5,541*
7	*281*	*1,946*	*5,638*
8	*299*	*2,079*	*5,700*
9	*320*	*2,197*	*5,768*
10	*327*	*2,303*	*5,790*
11	*391*	*2,398*	*5,969*
12	*371*	*2,485*	*5,916*
13	*341*	*2,565*	*5,832*
14	*394*	*2,639*	*5,976*
15	*357*	*2,708*	*5,878*
16	*414*	*2,773*	*6,026*



Use the "**add trendline"** and "*power function*" to obtain:

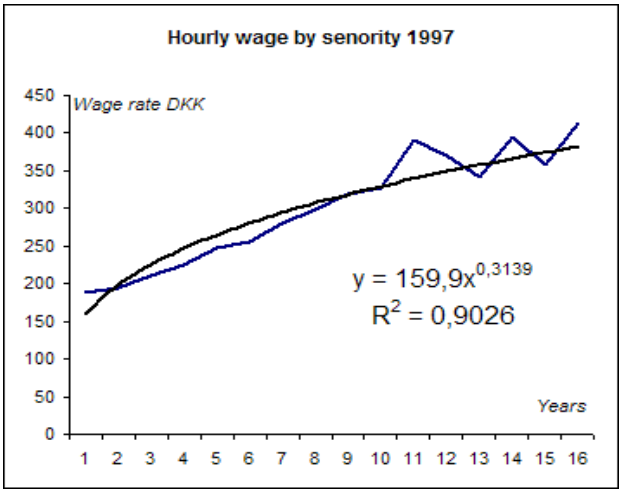


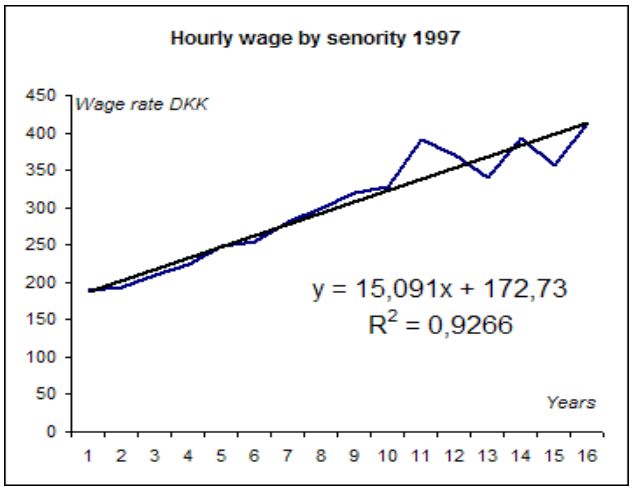
Use "options" to mark for "show equation" and "show R^2".



Place the mouse on an observation and right click. Then the menu above appears.

We then obtain the following:





For comparison purpose the linear trend line is also considered. Although the coefficient of determination is the highest for the linear model this will be insufficient for forecasting purposes. So the power functional form is selected in order to provide the most efficient model to explain the evolution of the hourly wage rate.

1.3. Other Transformations

Let us consider some cases where a non-linear model may be changed to a linear model by use of an appropriate transformation. Most models that can be transformed to linear models are called *intrinsically linear models*.

Consider first the *multiplicative* model:

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3} \varepsilon$$

This is a multiplicative model of three variables x_1, x_2 and x_3 with multiplicative errors. Assuming this behavior for the errors the model we are examining the *Cobb-Douglas* production function for three variables for example labor, capital and human capital. We can transform this model to a linear regression model by use of a *logarithmic transformation.* Taking natural logs (sometimes denoted by ln) of both sides of the equation gives the following linear model

$$\log y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + \beta_3 \log x_3 + \log \varepsilon$$

Notice that the usual assumptions with regard to the errors are valid, so initially the errors are not additive. If this was the case the model would not be an intrinsically linear model. In Excel we can take the natural logarithm by ***insert/function/ln*.** This results in a statement "=ln(cell)", and then copy.

Next consider the *exponential model*. For example, an exponential model in two independent variables can be stated as

$$y = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} \varepsilon$$

Taking the natural logs of both sides gives us the following regression model:

$$\log y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + \log \varepsilon$$

Let us now move to a more straight forward non-linear model. Consider the *logarithmic model:*

$$y = \beta_0 + \beta_1 \log x + \varepsilon$$

We can linearized by substituting the variable $x' = \log x$ into the equation. This gives us the linear model in x':

$$y = \beta_0 + \beta_1 x' + \varepsilon$$

Another nonlinear model that may be linearized by an appropriate transformation is the *reciprocal model*. A reciprocal model in several variables can be stated as:

$$y = \frac{1}{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon}$$

This model becomes a linear model upon taking the reciprocals of both sides of the equation. In practical terms, we run a regression of $1/y$ versus the x_i variables unchanged.

1.4. Variance Stabilizing Transformations

Remember that one of the assumptions of the regression model is that the regression errors ε has equal variance. If the variance of the errors increases or decreases as one or more of the independent variables changes, we have the problem of heteroscedasticity. In this case our regression coefficient estimators are not efficient. This violation of the regression assumptions may sometimes be corrected by use of a transformation. We will consider three major transformations of the dependent variable y to correct for heteroscedasticity.

1. The square root transformation: $y' = \sqrt{y}$
   This is the last "severe" transformation. It is useful when the variance of the regression errors is approximately proportional to the mean of y, conditional on the values on the independent variables x_i.

2. The logarithmic transformation: $y' = \log y$ (by any base)
   This is a transformation of a stronger nature and is useful when the variance of the errors is approximately proportional to the square of the conditional mean of y.

3. The reciprocal transformation: $y' = 1/y$
   This is the most severe of the three transformations and is required when the violation of equal variance is serious. This transformation of the errors is useful when the variance of the errors is approximately proportional to the conditional mean of y to the fourth power.

2. Modeling the US Electricity Supply

The article *Returns to Scale in Electricity Supply* by *Marc Nerlove* uses the model outlined in Section 1.3, and investigates for economics of scale in the US electricity sector vintage 1955. This is a classic article in econometrics. Nerlove uses a cross-section data set covering 145 privately owned electricity plants.

This issue of returns to scale has important bearing on the institutional arrangements necessary to secure an optimal allocation of resources. About 80 % of the electricity supply is supplied by private owned firms. A special problem with the production of electricity is that power cannot be stored.

The model considered for the production that determines supply has the form:

c = total production costs
y = output (measured in kwh)
x_1 = labor input — p_1 = wage rate
x_2 = capital input — p_2 = "price" of capital
x_3 = fuel input — p_3 = price of fuel

ε = a residual explaining neutral variations in efficiency of the firms

The generalized Cobb-Douglas production function can be stated as:

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3} \varepsilon \qquad (1)$$

Minimization of costs implies:

$$c = p_1 x_1 + p_2 x_2 + p_3 x_3 \qquad (2)$$

Solution to the system of (2) minimized subject to (1) implies the marginal productivity conditions[1]:

$$\frac{p_1 x_1}{\beta_1} = \frac{p_2 x_2}{\beta_2} = \frac{p_3 x_3}{\beta_3} \qquad (3)$$

However, if the efficiency of firms varies neutrally, as indicated by the error term in (1), and input prices varies from firm to firm, then the levels of input are not determined independently but are determined jointly by use of the firm's efficiency, level of output, and the factor prices it must pay to labor, capital and fuel.

[1] The solution to this problem can be found in a standard textbook on Microeconomics. For the mathematical description of this function see for example Ian Jacques *Mathematics*, sixth edition, FT Prentice Hall, pages 169 and 394. This is the textbook used in *Tools for Quantitative Analyses I.*

This problem of identification is known as the *confluent relation problem.* However, it is possible to fit *the reduced form* of the system of equations such as (1) and (3) and derive estimates of the structural parameters from estimated of the reduced form parameters. An important reduced form turns out to be the cost function:

$$c = ky^{1/r} p_1^{\beta_1/r} p_2^{\beta_2/r} p_3^{\beta_3/r} \nu \tag{4}$$

where

$$k = r(\beta_0 \beta_1^{\beta_1} \beta_2^{\beta_2} \beta_3^{\beta_3})^{-1/r} \tag{5}$$

$$\nu = \varepsilon^{-1/r} \tag{6}$$

and

$$r = \beta_1 + \beta_2 + \beta_3 \tag{7}$$

In our case k is a parameter measuring the level of technology imboided in the components of the cost function. Further, ν is the monotonic transformation of the residuals ε.

The most important parameter is r measuring the degree of returns to scale. If $r > 1$ there are increasing returns to scale (IRS); if $r = 1$ there are constant returns to scale (CRS), and if $r < 1$ there are decreasing returns to scale (DRS).

A production function that is appropriate for estimation can now be stated as:

$$C = K + \frac{1}{r} Y + \frac{\beta_1}{r} P_1 + \frac{\beta_2}{r} P_2 + \frac{\beta_3}{r} P_3 + E \tag{8}$$

where capital letters denote logarithms *(ln)* of the corresponding lower case letters. Note that under the special case with constant returns to scale $r=1$ and the estimates of the β*'s* give the correct estimates of the model.

The model (8) is called the *unrestricted model.* Here it will be called **MODEL 1**. How do we incorporate the restriction that the coefficients of the prices of the inputs add up to one? This can be done for example by dividing costs and two of the prices by the remaining price[2]. When fuel price is used as the divisor, the result is:

$$C - P_3 = K + \frac{1}{r} Y + \frac{\beta_1}{r}(P_1 - P_3) + \frac{\beta_2}{r}(P_2 - P_3) + E \tag{9}$$

This is called **MODEL II**.

[2] It does not matter either economically or statistically which price is chosen.

The two first models assumes that the relevant "price" of capital is available and that this price varies significantly from firm to firm. In reality most firms will finance its production plant by loans at the market interest rate. In such case the price of capital is the same for all firms. Incorporation of this assumption leave us with **MODEL III:**

$$C = K^* + \frac{1}{r} Y + \frac{\beta_1}{r} P_1 + \frac{\beta_3}{r} P_3 + E \qquad (10)$$

where $K^* = K + \frac{\beta_2}{r} P_2$.

The Excel file in Blackboard called *US Electricity Supply Nerlove.xlsx* brings the data. Estimate by yourself and confirm my findings.

For **MODEL I** we obtain:

MODEL I

Regression Statistics	
Multiple R	0.96
R-squared	0.93
Adjusted R-square	0.92
Standard Error	0.39
Observations	145

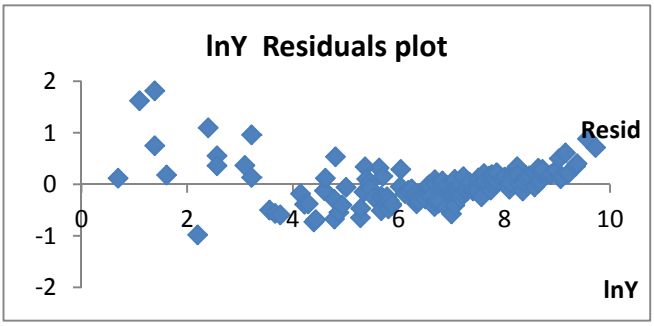
ANOVA

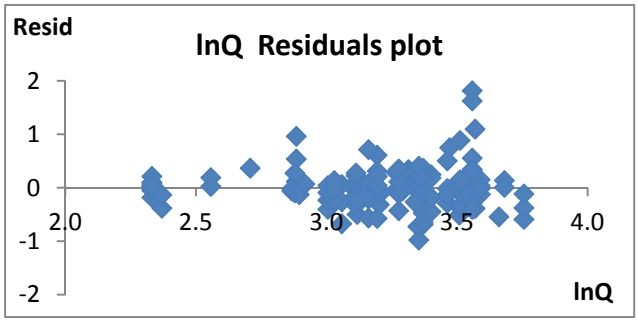
	df	*SS*	*MS*	*F-test*	*P-value*
Regression	4	269.42	67.35	437.79	0.00
Residual	140	21.54	0.15		
Sum	144	290.95			

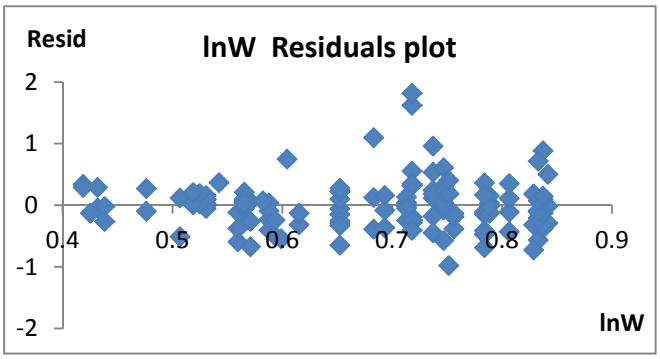
	Coef	*St error*	*t-stat*	*P-value*	*low 95%*	*high 95%*
Constant	-3.51	1.77	-1.98	0.05	-7.01	0.00
lnY (output)	0.72	0.02	41.25	0.00	0.69	0.75
lnW (wages)	0.43	0.29	1.49	0.14	-0.14	1.01
lnQ (fuel)	0.43	0.10	4.25	0.00	0.23	0.62
lnR (capital)	-0.22	0.34	-0.66	0.51	-0.89	0.45

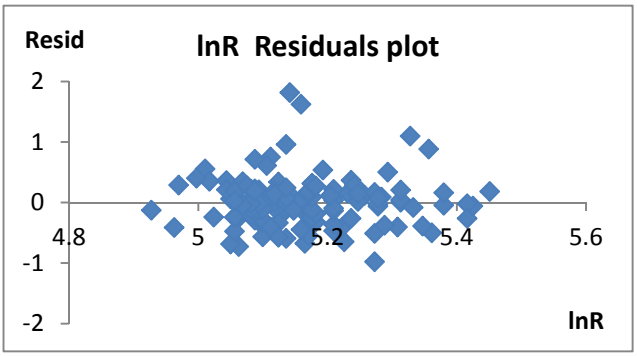
This is pretty good. The estimated coefficients of wages and especially capital are not significant. Looking at the coefficients then the sum of the three coefficients is equal to 0.64 (0.43 + 0.43 – 0.22). This could indicate decreasing returns to scale, but not all variables are significant, so we are not able to judge.

The plots of residuals are fairly good, and can be found on the next page:









The most critical of the plots of residuals is the plot of the logarithm of the output. This issue is also considered in the illustration on page 179 and page 182 in the article. It is clear that the model is not true. It looks as there are great diversity among the suppliers with low output. An explanation on this issue could be that the small firms experience decreasing returns to scale, whereas the large firms experience increasing returns to scale.

Next turn to the result for **MODEL II.** Our data needs here some additional calculations because the logarithm of the price of fuel has to be subtracted from the other variables. The following output is obtained:

MODEL II

Regression Statistics	
Multiple R	0.97
R-squared	0.93
Adjusted R-square	0.93
Standard Error	0.39
Observations	145

ANOVA

	df	*SS*	*MS*	*F-test*	*P-value*
Regression	3	294.58	98.19	640.10	0.00
Residual	141	21.63	0.15		
Sum	144	316.21			

	Coef	*St error*	*t-stat*	*P-value*	*low 95%*	*high 95%*
Constant	-4.69	0.88	-5.30	0.00	-6.44	-2.94
lnY* (output)	0.72	0.02	41.34	0.00	0.69	0.76
lnW* (wages)	0.59	0.20	2.90	0.00	0.19	1.00
lnR* (capital)	-0.01	0.19	-0.04	0.97	-0.38	0.37

Although appealing from a theoretical point of view this model is not performing satisfactory. First, the coefficients do not sum to unity, and second there is still the problem with the coefficient of the capital stock.

A solution could be to leave out the capital stock, but then with only two input variables the restriction does not give a meaning.

This suggests that we estimate **MODEL III** leaving out the capital variable. The result is:

MODEL III

Regression Statistics	
Multiple R	0.96
R-squared	0.93
Adjusted R-square	0.92
Standard Error	0.39
Observations	145

ANOVA

	df	*SS*	*MS*	*F-test*	*P-value*
Regression	3	269.35	89.78	585.93	0.00
Residual	141	21.61	0.15		
Sum	144	290.95			

	Coef	*St error*	*t-stat*	*P-value*	*low 95%*	*high 95%*
Constant	-4.65	0.34	-13.57	0.00	-5.33	-3.97
lnY (output)	0.72	0.02	41.43	0.00	0.69	0.76
lnW (wages)	0.48	0.28	1.70	0.09	-0.08	1.04
lnQ (fuel)	0.41	0.10	4.21	0.00	0.22	0.61

This model is performing better! The only concern is that the coefficient of the wage variable is only weak significant. Taking into consideration the confidence intervals of the coefficients it is not possible to determine the degree of returns to scale, but overall it seems to be close to unity.

Finally, let us take size into consideration. In the data set the firms are sorted according to output size and divided into five groups.

First the overall performance by group is considered.

Comparison Overall Regression

	Group A	*Group B*	*Group C*	*Group D*	*Group E*
Multiple R	0.69	0.82	0.80	0.94	0.97
R-squared	0.47	0.67	0.65	0.88	0.93
Adjusted R-square	0.41	0.63	0.61	0.87	0.93
Standard Error	0.59	0.22	0.18	0.12	0.15
Observations	29	29	29	29	29

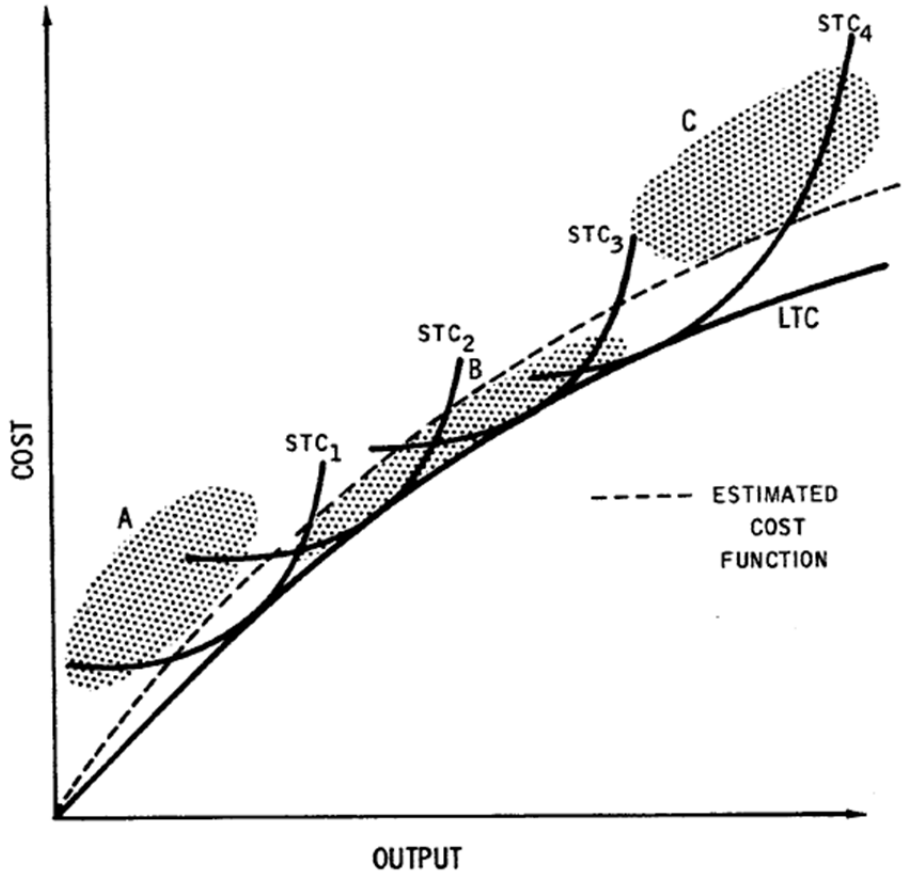
The table confirms the view that the model fits better for the larger plants. Turning to the coefficients the following results are obtained:

Comparison of coefficients

	Group A		*Group B*		*Group C*		*Group D*		*Group E*	
	Coef	*Sig*	*Coef*	*Sig*	*Coef*	*Sig*	*Coef*	*Sig*	*Coef*	*Sig*
Constant	-3.14	**	-4.12	***	-6.03	***	-6.14	***	-8.07	***
lnY (output)	0.39	***	0.66	***	0.99	***	0.93	***	1.04	***
lnW (wages)	-0.02		-0.40		-0.02		0.33		0.70	**
lnQ (fuel)	0.42		0.49	***	0.33	**	0.43	**	0.64	***

Note: *** significant at the 1 % level; ** significant at the 5 % level; * significant at the 10 % level

The model for the small plants only performs good with regard to output. Especially the wage rate is only significant for the larger plants. Notice, also that the sum of the coefficients of the input variables increases as the size of the plants increases. This indicates clearly economics of scale as the plant size increases. See also figure 2 in Nerlove page 180 displayed on the next page.



Reference

Marc Nerlove, 1963, *Returns to Scale in Electricity Supply*. Chapter 7 in C. Christ et.al. (ed), *Measurement in Economics*. Pages 167-198. Stanford University Press.

Nils Karl Sørensen ©
Advanced Tools Statistics

Set 3: Modeling Issues of Tourism

by Nils Karl Sørensen

Outline ***page***

1. Introduction

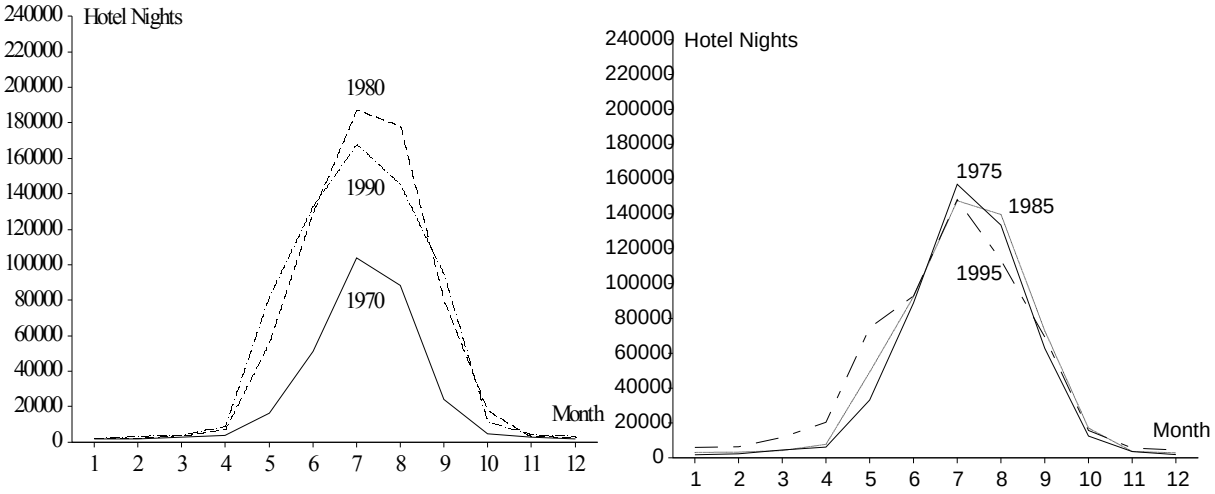
The present set of notes deals with time series modeling related to tourism. There is a wide literature on tourism and statistical estimation of issues related to tourism. The literature can divided into 3 themes: First, modeling of seasonal fluctuations, and secondly modeling tourism demand. In both cases, the estimates very frequently are used to forecast the future demand. Finally, various kinds of marketing analyses are used in order to identify the segments the market.

2. Working with Seasonality and Dummy Variables

An important issue in tourism economics is seasonality. The reason is obvious. The season peaks in different periods over the year, and consequently labour demand and the incomes varies accordingly. For example in Denmark, tourism demand is peaking during the summer, whereas the winter is the off season. In Norway, the seasonal pattern is different with a peak in February as well as in July. This is due to the winter ski season.

Considered over a year monthly tourist arrivals in Denmark looks very much like the curve of the normal distribution. See the examples below.

Bays-Ballot plots of hotel nights for the county of Bornholm, all nationalities



Source: Sørensen (1999)

A Bays-Ballot is a diagram is showing the fluctuations per year at the seasonal frequencies i.e. quarterly, monthly or weekly observations. The plot to the left displays the seasonal pattern of the island of Bornholm located in the Baltic Sea, whereas the panel to the right shows the seasonal pattern for all nationalities of hotel nights for the total of Denmark. Observe the very stable pattern. Nothing is happening outside the season. Also here the

seasonal pattern is very stable. Of course, the tourism authorities do all what they can to change this seasonal pattern and enlarge the season by developing new activities etc.

How can we model the seasonal pattern of tourism, and how can we use statistics to test for the effect of the tourism policy? Let us first turn to the nature of seasonality.

In statistics we deal with two types of seasonality, namely deterministic and stochastic seasonality.

- ***Deterministic seasonality*** is *predetermined* and constant from year to year. The seasonal pattern is constant and will not move.
- ***Stochastic seasonality*** is changing, but based on an underlying trend. If stochastic seasonality is present then we say that a *unit root* is present in the data series.

In both cases we can set up a regression and perform a test for an investigation of the type of seasonality. We consider quarterly data in the following examples[1]. The calculations can be found in the Excel file ***Example Dummy Seasonal.xls***. Try to go though the steps by yourself.

2.1. Deterministic Seasonality

Dummy variables are used to model deterministic seasonality. Let us for a given variable y consider a situation with deterministic seasonality. A dummy variable can take either the value zero or one.

In time series analysis we use *not adjusted seasonal statistics* at the quarterly frequency for example. If seasonality is of *deterministic* nature we can set up the following model:

$$y_t = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon_t$$

where $D_1 = 1$ in quarter 1 and 0 otherwise
$D_2 = 1$ in quarter 2 and 0 otherwise
$D_3 = 1$ in quarter 3 and 0 otherwise

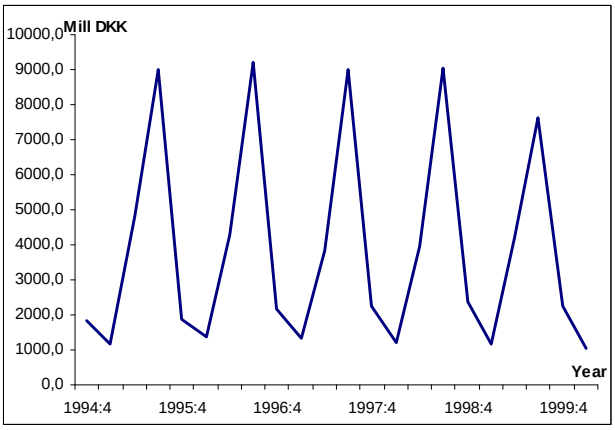
Notice that in this case only three dummy variables are required to represent four seasons. In this formulation β_1 shows the extent to which the expected value of y in the first quarter differs from the expected value in the fourth quarter, the omitted condition. β_3 and β_4 can be interpreted similarly. Alternatively we can use 4 dummies and exclude the constant term.

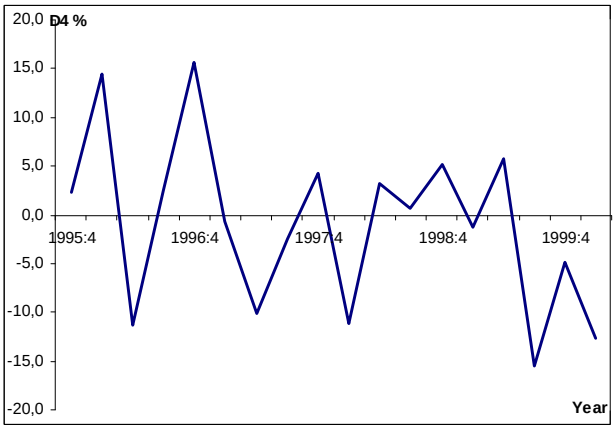
We can then observe the *p-values* (t-tests) in the Excel output and examine the effect of seasonality. This approach was applied by Barsky and Miron (1989).

[1] The set up can be expanded to monthly or bimonthly data as well, but complexity increases.

Example

We consider a quarterly data set of turnovers from renting Danish holiday cottages ranging from 1994.4 to 2000.1. So we only have 22 observations. Consequently, the data set can only serve as an illustration.





Data are shown below with the seasonal dummies. The graph to the left shows the raw data, whereas the graph to the right shows the fourth difference change in percent i.e. $((y_t - y_{t-4})/y_t) \times 100$. It is observed that a seasonal behavior is surely present in the data set with a summer peak in the third quarter followed by a winter slump. The seasonal behavior is preserved when calculation the percentage change, but the series becomes stationary around zero.

Data set

Year	Y: Mill DKK	Q_1	Q_2	Q_3	Q_4	D_4
1994:4	1,837.8	0	0	0	1	
1995:1	1,184.4	1	0	0	0	
1995:2	4,831.6	0	1	0	0	
1995:3	8,992.9	0	0	1	0	
1995:4	1,879.7	0	0	0	1	2.3
1996:1	1,354.7	1	0	0	0	14.4
1996:2	4,284.5	0	1	0	0	-1.3
1996:3	9,207.6	0	0	1	0	2.4
1996:4	2,174.7	0	0	0	1	1.,7
1997:1	1,345.2	1	0	0	0	-0.7
1997:2	3,849.7	0	1	0	0	-10.1
1997:3	8,987.9	0	0	1	0	-2.4
1997:4	2,267.8	0	0	0	1	4.3
1998:1	1,196.0	1	0	0	0	-11.1
1998:2	3,972.4	0	1	0	0	3.2
1998:3	9,040.1	0	0	1	0	0.6
1998:4	2,384.6	0	0	0	1	5.2
1999:1	1,180.0	1	0	0	0	-1.3
1999:2	4,197.8	0	1	0	0	5.7
1999:3	7,642.2	0	0	1	0	-15.5
1999:4	2,270.2	0	0	0	1	-4.8
2000:1	1,029.7	1	0	0	0	-12.7

Let us now perform the regression:

$$y_t = \beta_0 + \beta_1 Q_1 + \beta_2 Q_2 + \beta_3 Q_3 + \varepsilon_t$$

(Notice we have omitted Q_4 because the constant term is included). We obtain:

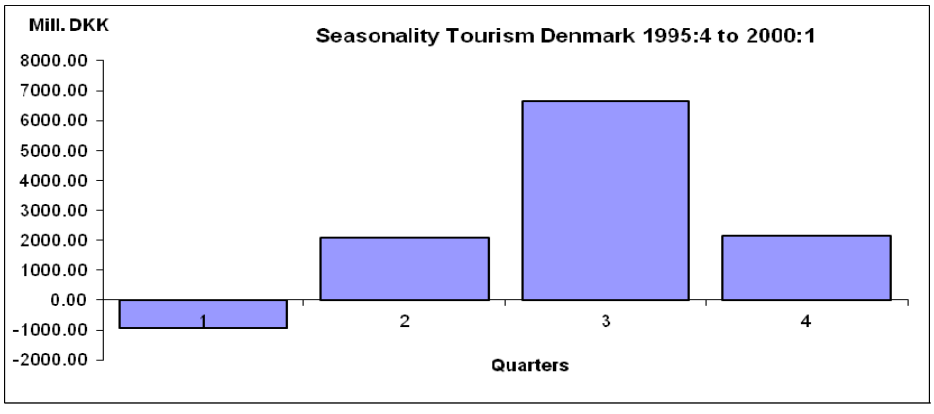
Regression Statistics	
Multiple R	0.99
R Square	0.99
Adjusted R Square	0.98
Standard Error	375.42
Observations	22

ANOVA

	df	*SS*	*MS*	*F-test*	*Signif. F*
Regression	3	181,229,715.82	60,409,905.27	428.62	0.00
Residual	18	2,536,955.83	140,941.99		
Total	21	183,766,671.65			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2,135.80	153.27	13.94	0.00	1,813.80	2,457.80
Q1	-920.80	216.75	-4.25	0.00	-1,376.18	-465.42
Q2	2,091.40	227.33	9.20	0.00	1,613.80	2,569.00
Q3	6,638.34	227.33	29.20	0.00	6,160.74	7,115.94

We observe a peak in the third quarter as expected a negative coefficient in the first quarter as expected. In general, this model performs very well, and here seasonality is deterministic of nature. A graph of the coefficients of the regression shows the seasonal behaviour:



The problem with this model is that data may be non-stationary, and have autocorrelation. If this is the case (I did not check it) then we should use the differenced series instead. The model is then:

$$D_4 = \beta_0 + \beta_1 Q_1 + \beta_2 Q_2 + \beta_3 Q_3 + \varepsilon_t$$

By running this regression we obtain:

Regression Statistics	
Multiple R	0.40
R Square	0.16
Adjusted R Square	-0.02
Standard Error	8.92
Observations	18

ANOVA

	df	*SS*	*MS*	*F*	*Signif. F*
Regression	3	208.85	69.62	0.87	0.48
Residual	14	1,114.31	79.59		
Total	17	1323.16			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.52	3.99	1.13	0.28	-4.04	13.08
Q1	-6.82	5.64	-1.21	0.25	-18.92	5.28
Q2	-7.67	5.98	-1.28	0.22	-20.51	5.16
Q3	-8.24	5.98	-1.38	0.19	-21.08	4.59

This result is not satisfactory because we do not find significant parameters. If this model is the true one then the nature of the seasonality is not deterministic. Instead it could be stochastic. We examine for this in the next Section.

2.2. Stochastic Seasonality

During the 1980'ties a new trend emerged in time series analysis or econometrics introduced by Clive Granger and Robert F. Engle. They argued that in a statistical sense data series may be attracted by each other. If this is the case then the series will cointegrated[2].

Consider for example Keynesian consumption theory arguing that consumption is a function of disposal income. If this is true then the fluctuations in consumption and disposal income should be highly correlated.

Normally, we remove the trend from a data series by taking the first order difference. This is the percentage change. If a data series has this property it is said to be integrated of order one. If consumption and income at the annual level both are integrated of order one then they will be attracted by each other and the difference among the two series will result in only white noise. If this is true cointegration will prevail. This information can be used to model their long run relation. A model along these lines will be considered in Section 3.

This theory can be extended to seasonal data. This case is more complex because we have an increased number of observations, and because seasonal statistics usual has as specific seasonal pattern. As a consequence, we only consider the issue of integration, and we only consider quarterly statistics and leave away the monthly case.

Hylleberg et.al. (1990) hereafter HEGY developed a test for the examination of seasonal integration. They wanted to examine for *stochastic seasonality*.

Stochastic seasonality will be present if we observe that over time there will be stochastic fluctuations around a given pattern. If this is the case then we say that a seasonal unit root is observed.

When a seasonal unit root is observed this information should be taken into account when a statistical model for forecasting purposes is set up. This model should then be better than the model with deterministic seasonal dummies presented in the previous section.

We can test for stochastic seasonality by running the following regression also called the "HEGY-regression":

$$y_{4.t} = \beta_0 + \beta_1 y_{1.t-1} + \beta_2 y_{2.t-1} + \beta_3 y_{3.t-2} + \beta_4 y_{3.t-1} + \varepsilon_t$$

The auxiliary variables are defined as:

[2] In 2003 they received the Nobel Prize in economics for their work on cointegration. Clive Granger passed away in 2009. Both Clive Granger and Robert F. Engle has/had close relations to an econometric group in Denmark at the University of Aarhus.

$$y_1 = y_t + y_{t-1} + y_{t-2} + y_{t-3} \quad \text{annual frequency}$$
$$y_2 = -(y_t - y_{t-1} + y_{t-2} - y_{t-3}) \quad \text{biannual frequency}$$
$$y_3 = -(y_t - y_{t-2}) \quad \text{1 and 3rd frequency}$$
$$y_4 = y_t - y_{t-4}$$

These transformations works like filters and remove all other variation then the one being tested. As the transformations are linked to the β's the investigation for stochastic seasonality is linked to the t-test of the significance of the coefficients.

The "t-test" is then performed on the β's. Here:

β_1 is the annual frequency (annual)
β_2 is the biannual frequency (second quarter)
β_3 is the first seasonal frequency (first quarter)
β_4 is the third seasonal frequency (third quarter)

The hypotheses are:

H_0: If $\beta_i = 0$ stochastic seasonality is present (seasonal unit root)
H_1: If $\beta_i \neq 0$ stochastic seasonality is not present (no seasonal unit root)

Then the tests are as t-tests. The critical values for the test are unfortunately not standard. We cannot apply a normal distribution or a t-distribution. We can set up the following table for some of the critical values assuming a level of significance equal 5 %:

Sample size		β_1	β_2	β_3	β_4
Observations:	*Years:*				
48	12	−2.96	−1.95	−1.90	−1.72
100	25	−2.88	−1.95	−1.90	−1.68
136	34	−2.89	−1.91	−1.88	−1.68
200	50	−2.87	−1.92	−1.90	−1.66

The interpretation of the critical values is as follows: If we for example have a sample with 100 observations (or around) and we estimate the t-statistic of β_1 to equal −1.80 i.e $\frac{\beta_1}{s(\beta_1)}$.
Then −1.80 > −2.88 and we accept H_0 so stochastic seasonality is present.

The critical values are taken from Hylleberg et.al. (1990). The test can be extended to the monthly case. For an analysis of tourism data with HEGY tests for Australia on quarterly data see Kim (1999). For an analysis on monthly tourism data by use of the HEGY test see Sørensen (1999). Critical values for tests at the quarterly, bimonthly and monthly frequency can be found in Fransens and Hobijn (1997). References are found at the end of this section.

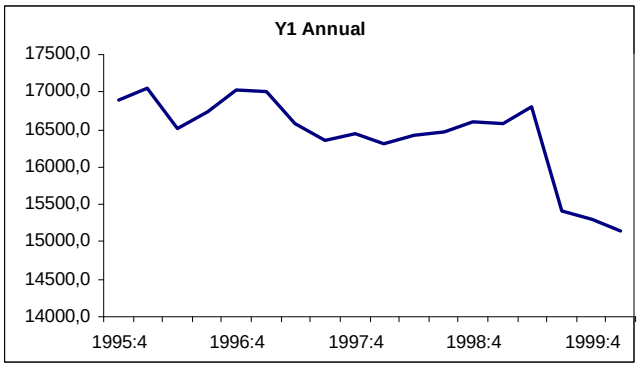
Example

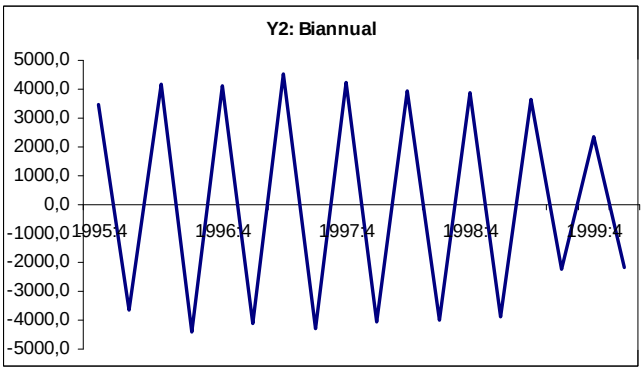
Let us look at the data set from the example above. The result on the differenced data was not convincing. Alternatively, the seasonal movements could be of stochastic nature. Again we use the data series *Y: Mill DKK*. We apply the formulas above on this series, and calculate the auxiliary variables. These are shown below at the left panel. As seen from the regression formula some additional lagging of the variables are needed. This task is undertaken in the data set shown to the right side.

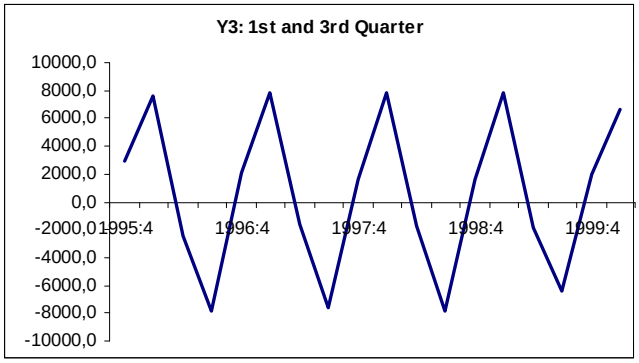
Data set

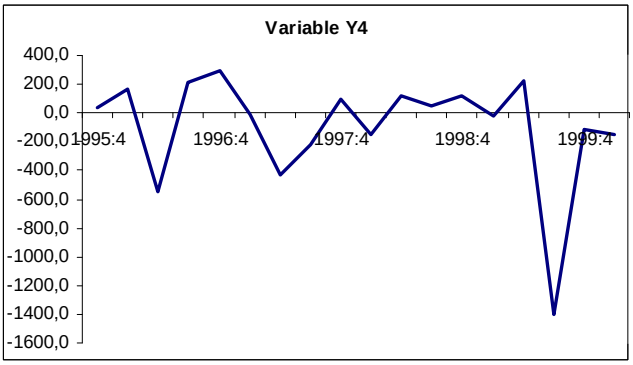
Year	Y_4	Y_1	Y_2	Y_3	Y_4	$Y_{1.t-1}$	$Y_{2.t-1}$	$Y_{3.t-2}$	$Y_{3.t-1}$
1994:4									
1995:1									
1995:2									
1995:3									
1995:4	41.9	16,888.6	3,466.0	2,951.9	41.9				
1996:1	170.3	17,058.9	-3,636.3	7,638.2	170.3	16,888.6	3,466.0		2,951.9
1996:2	-547.1	16,511.8	4,183.4	-2,404.8	-547.1	17,058.9	-3,636.3	2,951.9	7,638.2
1996:3	214.7	16,726.5	-4,398.1	-7,852.9	214.7	16,511.8	4,183.4	7,638.2	-2,404.8
1996:4	295.0	17,021.5	4,103.1	2,109.8	295.0	16,726.5	-4,398.1	-2,404.8	-7,852.9
1997:1	-9.5	17,012.0	-4,093.6	7,862.4	-9.5	17,021.5	4,103.1	-7,852.9	2,109.8
1997:2	-434.8	16,577.2	4,528.4	-1,675.0	-434.8	17,012.0	-4,093.6	2,109.8	7,862.4
1997:3	-219.7	16,357.5	-4,308.7	-7,642.7	-219.7	16,577.2	4,528.4	7,862.4	-1,675.0
1997:4	93.1	16,450.6	4,215.6	1,581.9	93.1	16,357.5	-4,308.7	-1,675.0	-7,642.7
1998:1	-149.2	16,301.4	-4,066.4	7,791.9	-149.2	16,450.6	4,215.6	-7,642.7	1,581.9
1998:2	122.7	16,424.1	3,943.7	-1,704.6	122.7	16,301.4	-4,066.4	1,581.9	7,791.9
1998:3	52.2	16,476.3	-3,995.9	-7,844.1	52.2	16,424.1	3,943.7	7,791.9	-1,704.6
1998:4	116.8	16,593.1	3,879.1	1,587.8	116.8	16,476.3	-3,995.9	-1,704.6	-7,844.1
1999:1	-16.0	16,577.1	-3,863.1	7,860.1	-16.0	16,593.1	3,879.1	-7,844.1	1,587.8
1999:2	225.4	16,802.5	3,637.7	-1,813.2	225.4	16,577.1	-3,863.1	1,587.8	7,860.1
1999:3	-1,397.9	15,404.6	-2,239.8	-6,462.2	-1,397.9	16,802.5	3,637.7	7,860.1	-1,813.2
1999:4	-114.4	15,290.2	2,354.2	1,927.6	-114.4	15,404.6	-2,239.8	-1,813.2	-6,462.2
2000:1	-150.3	15,139.9	-2,203.9	6,612.5	-150.3	15,290.2	2,354.2	-6,462.2	1,927.6

First we graph the variables Y_1 to Y_2 on the left side data set









These plots are instructive (actually, I think that these transformations are the most informative and valuable part of the analysis). The upper left panel on page 9 reveals that the underlying annual trend actually is negative. The upper right panel on page 9 shows the biannual fluctuations. It shows really decreasing amplitude since 1997. The lower left panel above giving us the first and third quarter fluctuations shows also slightly decreasing amplitude. We should then expect varying amplitude at the biannual frequency, but probably not at the other seasonal frequencies. Finally, the transformation Y4 should show no systematic behaviour. This is surely the case here.

Next step is to perform the regression in Excel on the right side data set. The result is:

Result HEGY test

Regression Statistics	
Multiple R	0.43
R Square	0.18
Adjusted R Square	0.11
Standard Error	434.45
Observations	16

ANOVA

	df	*SS*	*MS*	*F*	*Signif. F*
Regression	4	466,687.00	116,671.75	0.62	0.66
Residual	11	2,076,215.45	188,746.86		
Total	15	2,542,902.46			

	Coefficients	*Standard Error*	*t Stat*	Critical HEGY	
Intercept	1480.51	3984.27	0.37		
Y1 (t-1)	-0.10	0.24	-0.40	−2.96	H_0 accepted
Y2 (t-1)	-0.02	0.03	-0.81	−1.95	H_0 accepted
Y3 (t-2)	-0.02	0.02	-0.96	−1.90	H_0 accepted
Y3 (t-1)	-0.01	0.02	-0.52	−1.72	H_0 accepted

We use the critical values for 48 observations. This is not optimal, and the critical HEGY-values will properly be lower.

Notice that we have changed output a little. We have left out the lower and upper 95 % values and the p-value. Instead the critical values from table above are inserted. In general, we accept H_0 so a varying and changing seasonal component is found. However, our result should be written with care because of the very little sample we use.

This explains why the deterministic model above based on differenced data performs so poor. If we use the dummy variable approach we should use the non-transformed data only.

If we use differenced data a more complex model is required. We cannot use the regression for the test. As evident from the regression output the overall performance of the "HEGY-test regression" is poor, so something else should be applied.

References

Nils Karl Sørensen, 1999, *Modelling the Seasonality of Hotel Nights in Denmark by County and Nationality.* Tourism Economics 5, 9-23

Jae H. Kim, 1999, *Forecasting Monthly Tourist Departures from Australia*. Tourism Economics 5, 277-291.

Svend Hylleberg, Robert F. Engle, Clive Granger and Sam Yoo, 1990, *Seasonal Integration and Cointegration.* Journal of Econometrics 44, 215-238.

Svend Hylleberg, Nils Karl Sørensen and Clara Jørgensen, 1993, *Seasonality in Macroeconomic Time Series.* Empirical Economics 18, 321-335.

Phillip Hans Fransens and Bart Hobijn, 1997, *Critical Values for Unit Root Test in Seasonal Time Series.* Journal of Applied Statistics 27, 25-47.

Robert Barsky and Jeff Miron, 1989, *The Seasonal Cycle and the Business Cycle.* Journal of Political Economy 97, 503-534.

3. Introduction to Tourism Demand Analysis

For tourist organisations it is relevant to establish models for demand related to tourism. Apart from attractions the tourist demand for a certain destination can depend on for example the price level on the destination and the income in the home country of the tourist. If the currency exchange rate is low then demand is high. If income in the home land is high then the demand for holiday travels is high. Finally, extreme variations in prices of for example oil can influence transport costs and tourism.

The most commonly used functional form in tourism demand analysis is the power model. This can be expressed as:

$$Q_{ij} = \beta_0 P_i^{\beta_1} P_S^{\beta_2} Y_j^{\beta_3} T_j^{\beta_4} A_{ij}^{\beta_5} u_{ij}$$

Where

- Q_{ij} is the quantity of the tourism product demanded in destination i by tourists from country j
- P_i is the price of tourism for destination i
- P_S is the price of tourism for substitute destinations
- Y_j is the income in origin country j
- T_j is consumer tastes in origin country j
- A_{ij} is advertising expenditure on tourism by destination i in origin country j
- u_{ij} is the error term

The power model has much in common with the Cobb-Douglas production function already considered. For example it may be transformed into a linear relationship using logarithms.

$$\ln Q_{ij} = \beta_0 + \beta_1 \ln P_i + \beta_2 \ln P_S + \beta_3 \ln Y_j + \beta_4 \ln T_j + \beta_5 \ln A_{ij} + \varepsilon_{ij}$$

In order to estimate a demand relation, notice that it frequently not will be possible to find statistics for all the variables included. Further, subscript t has to be added for time series data.

Example: UK visitors to South Korea

To illustrate the model in this section and the two next sections consider a data set on inbound tourism demand for South Korea by UK visitors. The full analysis can be found on the file ***Example ECM-model.xls***. Try to go though the steps by yourself.

The total number of tourist arrivals *UKTA* is used as the dependent variable, and the data are obtained from the Korea National Tourism Corporation (KNTC). Data cover the period ranging from 1962 to 1994 (33 observations). Since the tourist arrivals variable includes both business and leisure travelers, the gross domestic product of UK (*UKGDP*) is used as the income variable, rather than personal disposal income.

In addition, to the model outlined above two variables are added. First, in order to reflect the influence of business activities on tourism demand, a trade volume variable measured by the sum of total imports and exports between South Korea and the UK is included. This variable is labeled *UKTV*. Second, we include a tourism price variable *RCPI* defined as:

$$RCPI_t = \frac{KCPI_t / UKWEX_t}{UKCPI_t}$$

Where KCPI is the South Korean consumer price index; UKWEX is the UK pound versus Korean won exchange rate, and finally, UKCPI is the UK consumer price index. All at time *t.* Defined in this way we should expect a positive relation between *RCPI* and *UKTA*.

Restated in the terms of the power model above the variables are defined as:

$$\ln UKTA_t = \beta_0 + \beta_1 \ln UKGDP_t + \beta_2 \ln UKTV_t + \beta_3 \ln RCPI_t + \varepsilon_t$$

Or by an alternative notation with capital letters:

$$LUKTA_t = \beta_0 + \beta_1 LUKGDP_t + \beta_2 LUKTV_t + \beta_3 LRCPI_t + \varepsilon_t$$

This model also describes the long run behavior of the tourism demand. We shall estimate the model in the next section.

In order to inspect how the variables are related consider the matrix of correlation:

Matrix of correlation

	LUKTA	*LUKGDP*	*LUKTV*	*LRCPI*
LUKTA	1.00			
LUKGDP	0.98	1.00		
LUKTV	0.99	0.95	1.00	
LRCPI	0.95	0.92	0.94	1.00

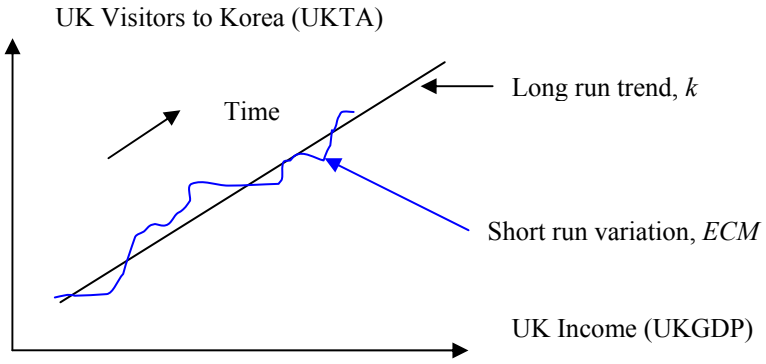
The table confirms all our expectations. But all the explanatory variables are also related to each other so multicollinarity is present. This suggests that the model should be reduced. This issue will be considered in the next Section

4. The Error Correction Model

The Neoclassical growth theory developed among other by Robert Solow is (hopefully) remembered from Macroeconomics[3]. This model claims that a "long run steady state" rule of growth will exist.

Using the example considered above it is claimed that over time the relation among for example hotel nights of UK visitors to for example South Korea will be positively related to the income level in UK. As wealth (income) increases, an increase in the number of visitors to South Korea is expected. This can be illustrated as follows:

Long Run Growth Path



So as time goes by, a positive relation should be observed. The ratio among *UKTA* (UK Tourist Arrivals in South Korea) and *UKGDP* (UK income level) should then be a constant if this long run model is valid, so:

$$k = \frac{UKTA_t}{UKGDP_t} \Rightarrow LUKTA_t = \beta_0 + \beta_1 LUKGDP_t$$

where t is time. If we take the logarithm (L), then our model becomes linear, and can be estimated by OLS, and we can find k.

Engle and Granger (1987) extended this model by pointing out the influence of the short run fluctuations on the long run evolution, and thereby giving name to the notion of the *Error Correction Model (ECM)*.

The Solow model has two problems. First, it does not explain how we come from one equilibrium to another. Therefore, it is static in time and not dynamic. Second, it does not

[3] See for example the textbook for the course in Macroeconomics by Blanchard, Amighini and Giavazzi chapters 11 to 13.

explain how short run fluctuations will influence on the long run growth. The ECM-approach captures this.

Engle and Granger say that the two considered variables (UKTA and UKGDP) *attract* each other. In the short run, we will observe short run fluctuations around the long run trend given by *k*. If the model is stable, then the short fluctuations will approach towards the steady state long run growth rate.

Assume that the short run fluctuations happens immediately one period ahead of the long run fluctuations. The short run variation is the change in the long run, ie. $\Delta LUKTA$, where Δ is the first difference operator ie. the change from year to year. The model is then for our two variables:

$$\Delta LUKTA_t = \beta_0 + \beta_1 \Delta LUKGDPt + \delta(LUKTA_{t-1} - \lambda_0 - \lambda_1 LUKGDP_{t-1}) + \varepsilon_t$$

"Short run" "Long run"

The coefficient δ is called the *error correction term*. This form follows after a series of mathematical manipulations undertaken by Engle and Granger (1987). Finally, ε_t is the residuals.

This form is not so easy to estimate. Engle and Granger suggest a two-step procedure as follows:

1. Estimate the long run relation (notice, that this is a reduced version of the model considered in the last Section):

   $$LUKTA_t = \lambda_0 + \lambda_1 LUKGDP_t + u_t$$

2. Save the residuals from this regression. Now use them and estimate:

   $$\Delta LUKTA_t = \beta_0 + \beta_1 \Delta LUKGDP_t + \delta u_{t-1} + \varepsilon_t$$

Notice, the lag corresponding to one period on the residuals from the first step regression is included[4]. The first relation to be estimated is also called the *cointegrating relation.*

Example: UK Visitors to South Korea

Again we use the data set on inbound tourism demand for South Korea by UK visitors to illustrate this model.

[4] If we used quarterly data, we should lag the residuals by 4 periods and use u_{t-4}. If data were monthly we lag the residuals by u_{t-12} etc.

The long run model is:

$$LUKTA_t = \lambda_0 + \lambda_1 LUKGDP_t + \lambda_2 LUKTV_t + \lambda_3 LRCPI_t + u_t$$

From this regression we save the residuals and estimate the ECM-model:

$$\Delta LUKTA_t = \beta_0 + \beta_1 \Delta LUKGDP_t + \beta_2 \Delta LUKTV_t + \beta_3 \Delta LRCPI_t + \delta u_{t-1} + \varepsilon_t$$

Let us look at some results from Excel. First *the long run relation*:

Regression Statistics	
Multiple R	0.99
R Square	0.99
Adjusted R Square	0.99
Standard Error	0.11
Observations	33

ANOVA

	df	***SS***	***MS***	***F***	***Signif. F***
Regression	3	58.77	19.59	1,735.02	0.00
Residual	29	0.33	0.01		
Total	32	59.10			

	Coefficients	***Std. Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***
Intercept	-3.32	1.10	-3.01	0.01	-5.57	-1.06
LUKGDP	2.28	0.28	8.09	0.00	1.70	2.86
LUKTV	0.43	0.04	9.99	0.00	0.34	0.52
LRCPI	0.29	0.12	2.51	0.02	0.05	0.53

All signs are positive as expected. The R^2 is very high because we are working with logarithmic transformed variables.

From this regression we save the residuals, and use them in the next step. Here we obtain *the ECM-model* as:

Regression Statistics	
Multiple R	0.72
R Square	0.52
Adjusted R Square	0.45
Standard Error	0.08
Observations	32

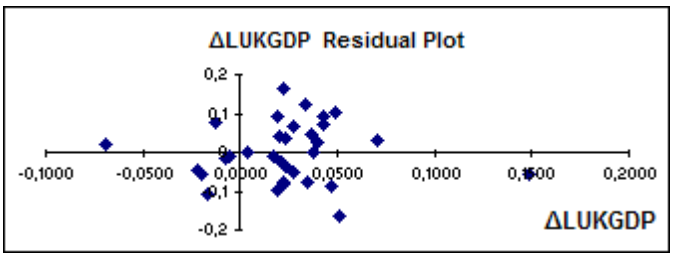
ANOVA

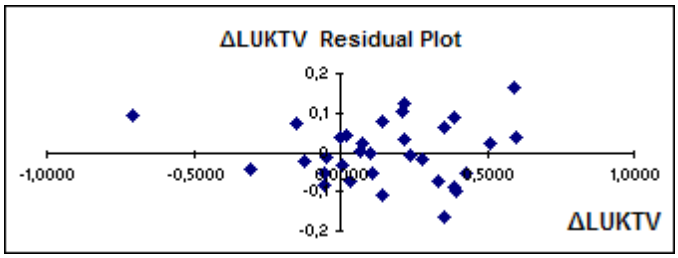
	df	*SS*	*MS*	*F*	*Signif. F*
Regression	4	0.198	0,049	7,42	0.00
Residual	27	0.180	0,007		
Total	31	0.377			

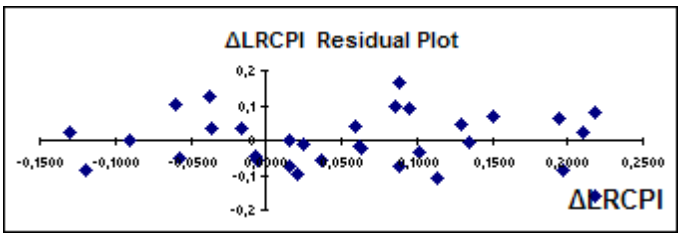
	Coefficients	*Std. Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.056	0.022	2.60	0.02	0.012	0.100
ΔLUKGDP	0.921	0.415	2.22	0.04	0.069	1.772
ΔLUKTV	0.272	0.058	4.72	0.00	0.154	0.390
ΔLRCPI	0.258	0.160	1.61	0.12	-0.071	0.586
Resid(t-1) (**ECM-term**)	-0.508	0.161	-3.16	0.00	-0.839	-0.178

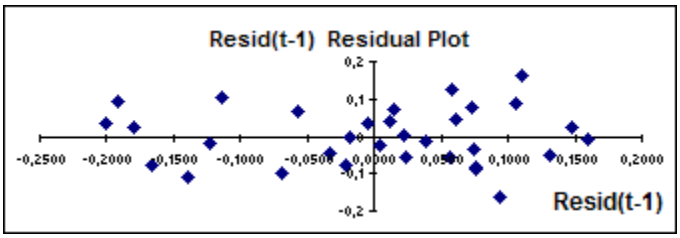
All variables are again significant. Importantly, the $RESID_{t-1}$ (that is the ECM-term) is significant. Therefore, the adjustment process is actually operating.

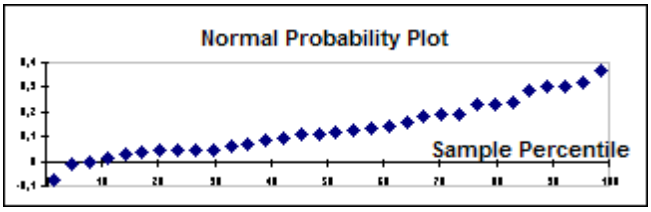
The plots of residuals are actually also very good.











References

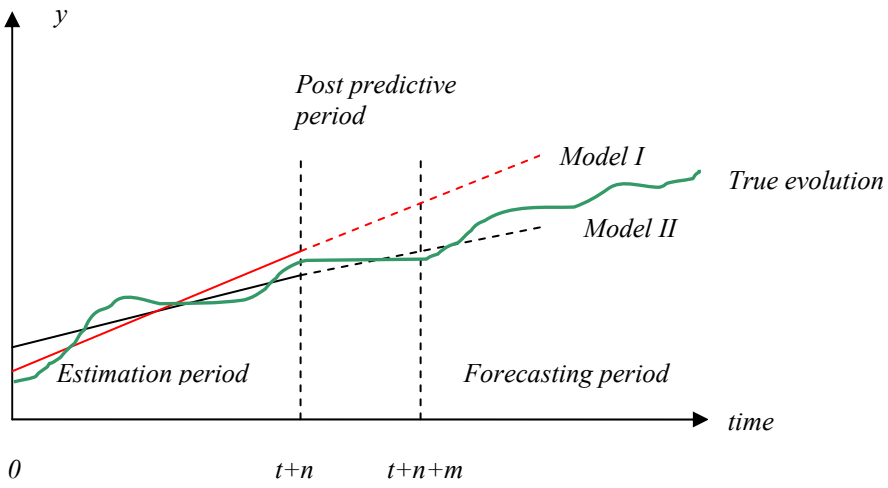
Robert F. Engle and Clive Granger, 1987, Cointegration and Error Correction: Representation, Estimation and Testing. *Econometrica 55,* page 251–276.

Tim Bollerslev and Svend Hylleberg, 1985, A Note on the Relation between Consumers Expenditure and Income in the United Kingdom. *Oxford Bulletin of Economics and Statistics* 47, No. 3 page 153-170.

5. Time Series Forecasting and Evaluation

This Section addresses the question of time series forecasting. Many different types of time series modeling can be stated as for example moving average models (ARIMA), exponential smoothing methods, decomposing models, and models dealing specially with issues of seasonality. We will leave these models to a course in forecasting or business econometrics.

Let us for the present purpose just assume that we have estimated a regression model based on time series statistics (at the annual, quarterly or monthly frequency), and now we want to use this model for forecasting outside the time period used for the estimation. In such a situation it is good to have saved some additional observations for *post predictive testing.* This is illustrated in the figure below:



We consider two models: I and II. We have to find the model that performs most efficient. As earlier we can define the forecasting error in the post predictive period for a give point in time denoted by t as $\varepsilon_t = y_t - \hat{y}_t$. Here y_t is the observed value and $\hat{y}_t$ is the forecasted value by one of the two different models considered above. For the m observations in the post predictive period we can calculate the mean absolute deviation or mean absolute error (MAE) defined as:

$$MAE = \frac{1}{m}\sum_{t=1}^{m}|\varepsilon_t|$$

This gives some kind of an average in absolute terms. For comparison purposes define the mean absolute percentage error (MAPE) as:

$$MAPE = \frac{\sum_{t=1}^{m} \frac{|\varepsilon_t|}{y_t}}{m}$$

This is just MAE written in relative terms. Obviously, the model with minimum MAPE should be chosen. From the figure above this looks to be present for model II. A more handy measure than the MAPE, is the root mean squared percentage error, RMSPE. It can be defined as:

$$RMSPE = \sqrt{\frac{1}{m}\sum_{t=1}^{m}\left(\frac{\varepsilon_t}{y_t}\right)^2}$$

The evaluation of this measure is as with the previously measures. Use the model with the minimum RMSPE.

Example: Forecasting UK Visitors to South Korea

As an illustration consider the data set on UK visitors to South Korea used earlier to set up the ECM-model. The data period is ranging from 1962 to 1994. How can a model for forecasting the number of tourist arrivals *UKTA* best be set up?

First data is divided into sub periods. For example consider the period 1962 to 1990 as the period of estimation. The period from 1991 to 1994 i.e. 4 years is the post predictive period. Assume that we want to forecast the non-transformed number of tourist arrivals'. In order to develop the forecast model we initially present several alternatives. Next the model is estimated for the estimation period. Then the forecasts of the post predictive period are calculated. Finally, the measures of evaluation is calculated and compared. The example can also be found in the file ***Example Forecasting.xls.***

Consider the following 5 models[5]:

Model I: Theory based model

Here a demand model is considered of the form:

$$UKTA_t = \beta_0 + \beta_1 UKGDP_t + \beta_2 UKTV_t + \beta_3 RCPI_t + \varepsilon_t$$

Here t is time, *UKTA* is tourist arrivals from UK to South Korea, *UKGDP* is UK Gross Domestic Product and *UKTV* is the trade variable defined as the sum of imports and exports

[5] This number is arbitrary. The models listed are just what I could come up with! There may be better models or ideas than mine.

between UK and South Korea. Finally, *RCPI* is a price variable defined on page 12 in these notes.

Model II: The Trend Model:
This model is stated as:

$$UKTA_t = \beta_0 + \beta_1 YEAR_t + \varepsilon_t$$

Here *YEAR* is a linear trend defined as: $YEAR_t = 1962, 1963, 1964, \ldots, 1994$. The model is completely without any theoretically foundations, and states that the number of tourists from UK to South Korea wills evolutes linearly over time.

Model III: A Polynomial Approach
The model states that the number of tourist arrivals depends on *UKGDP* in a non-linear way taking *UKGDP* squared into consideration. The model looks as:

$$UKTA_t = \beta_0 + \beta_1 UKGDP_t + \beta_2 UKGDP_t^2 + \varepsilon_t$$

Model IV: Dynamic Model
This model also considers the *UKGDP*, but instead of the squared GDP, the lagged GDP is included. This model states that past income for the last two years has an impact on the current flow of tourists. This is a consequence of fact that the decision of taking a holiday trip may take long time. Theoretically this model has its foundations from the theory of consumption. The model can be written as:

$$UKTA_t = \beta_0 + \beta_1 UKGDP_t + \beta_2 UKGDP_{t-1} + \beta_3 UKGDP_{t-2} + \varepsilon_t$$

Model V: Moving Average model:
This model takes into consideration the lagged values in the tourist flows from UK to South Korea for the past two periods. Such a process is also called an MA (Moving Average) process of order 2. The model claims that the tourist flow can be described by an inertia process without any theoretical foundations. The model can be written as:

$$UKTA_t = \beta_0 + \beta_1 UKTA_{t-1} + \beta_2 UKTA_{t-2} + \varepsilon_t$$

Common for models IV and V are that the period of estimation is two periods shorter than for the first 3 models. This is due to the presence of lagged values.

The next step is to estimate the five models from the start until 1990. The results are summarized in the table next page. The table is build specifically to reduce space and provide an easy comparison of a large Excel output. Notice, that the overall regression statistics are located in the bottom of the table.

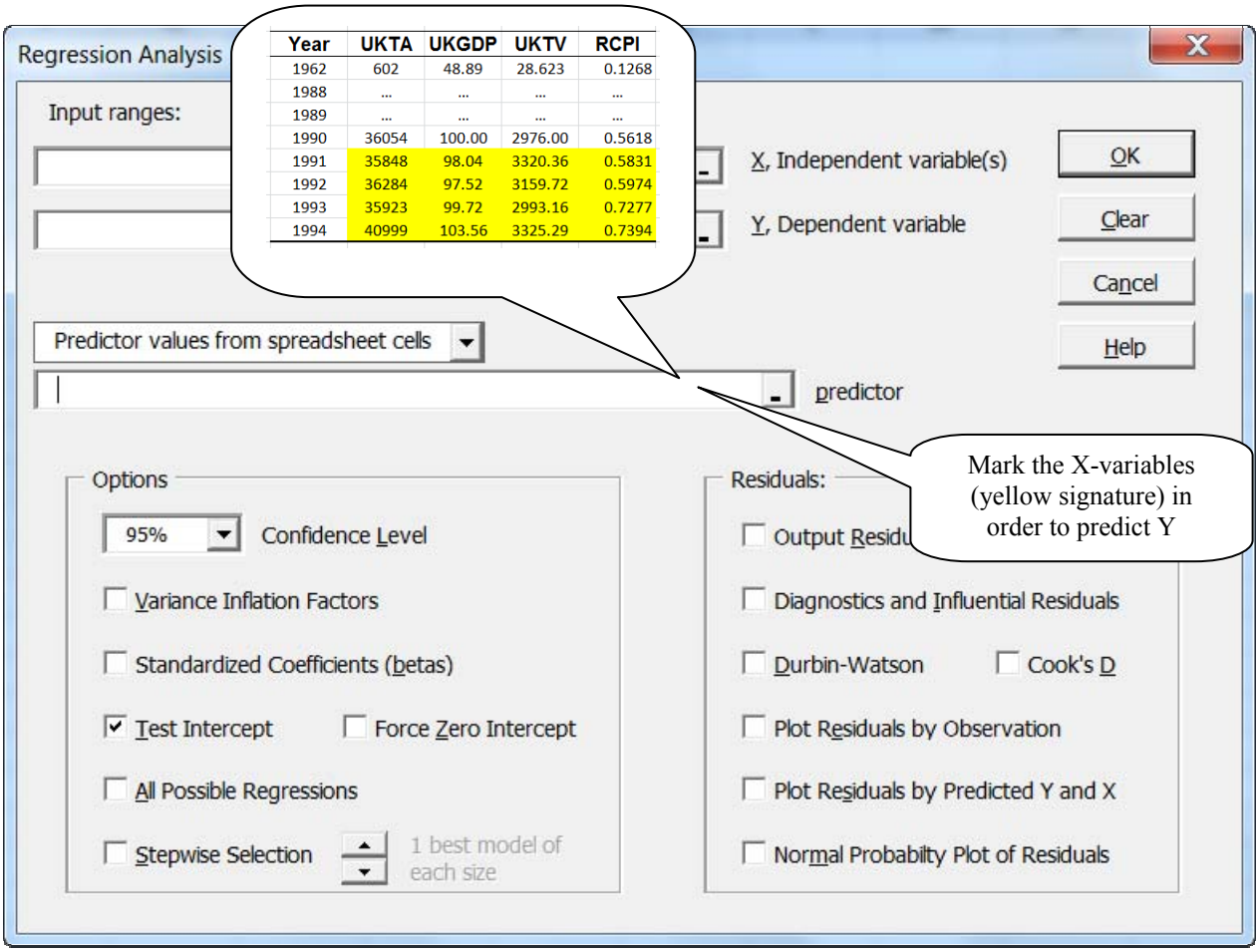
Model Performance and Comparisons

Variable	MODEL I		MODEL II		MODEL III		MODEL IV		MODEL V	
	Coef	*Sig*	*Coef*	*Sig*	*Coef*	*Sig*	*Coef*	*Sig*	*Coef*	*Sig*
Constant	-10,604	**	-2381754	***	12,861		-45,550	***	501.11	
UKGDP	198.76	**			-761.37	**	493.79	**		
UKTV	8.49	***								
RCPI	-1,849.22									
Year / Trend			1,211.19	***						
$UKGDP^2$					9.78	***				
UKGDP (t-1)							277.85			
UKGDP (t-2)							8.91			
UKTA (t-1)									0.74	***
UKTA (t-2)									0.37	*
Multiple R	0.99		0.94		0.98		0.97		0.99	
R-squared	0.98		0.89		0.95		0.94		0.98	
Adjusted R-square	0.98		0.89		0.95		0.93		0.98	
Standard Error	1,659.56		3,639.18		2,428.14		2,836.78		1,524.84	
Observations	29		29		29		27		27	

*Note: *** significant at the 1 % level; ** significant at the 5 % level; * significant at the 10 % level*

The overall statistics gives the smallest values of the standard errors for the models I and IV. In model I significant values is observed of the income and trade variable. The price variable *RCPI* should be excluded. In models II and III all explanatory variables are significant. This suggests that the tourists arrivals can be describes by a polynomial functional form as well as by a simple trend. According to the standard error model III outperforms model II. Overall model II is the worst of the models. Out of the dynamic models model V performs best, and this completely non-theoretical model has the lowest standard error! The model claims that the present tourism flow is a function of past flows. As the size of the coefficient decreases in time last year tourist flows has the largest impact on the present flow. In model IV only the present income has an impact on the flow of tourists. So there is no time decision element involved in the process.

The next step is to calculate the predicted values of the variable *UKTA* for the period 1991 to 1994. This is done by substitution of the values of the explanatory variables into the estimated models. This task is either done manually in Excel or by use of Megastat. In Megastat under *regression* there is a special option for this, see the illustration on the next page.



Undertaking this task for our five models results in the predictions by model presented in the upper part of the table below.

Predictions and Models

Year	Obs	Model I	Model II	Model III	Model IV	Model V
1991	35848	36002	29734	32246	31534	40095
1992	36284	34508	30945	31647	30736	40545
1993	35923	33290	32156	34217	31660	40794
1994	40999	36852	33367	38930	34163	40686

Compared to the observed values the picture looks diversified. In order to obtain a more precise picture of the forecast performance the evaluation indicators presented first in the present section have to be calculated. First the residuals are computed defined as $\varepsilon_t = y_t - \hat{y}_t$

The result is given on the next page where the computed overall indicators also can be found. Looking first at the residuals it can be observed that for most cases the forecasts "under estimates" the observed number of tourist arrivals. In other words; the models are performing too "conservative" relative to reality. This is also the truth in most forecasts in real life. Out of the 20 predictions made the model only "overshoots" in 4 cases. Only model I and V has this kind of behavior. In general, model V is too optimistic.

The lower part of the table reports the calculation of the forecast indicators. All gives a similar ranking of the models. Model I is the overall best model. As observed from the residuals this model is especially good for the first two years. In forecast year 3 it is the second best, and in forecast year 4 it is ranked as number 3.

Model I is followed by model III. – the polynomial model. Model V is ranked as number 3 due to the remarkable performance of the model in forecast year 4.

The forecast performance is a victory for the theory founded model I. This should also be the case, but it may easily be different in real life!

Residuals

Year	Model I	Model II	Model III	Model IV	Model V
1991	-154	6114	3602	4314	-4247
1992	1776	5339	4637	5548	-4261
1993	2633	3767	1706	4263	-4871
1994	4147	7632	2069	6836	313

	Model I	Model II	Model III	Model IV	Model V
MAE	2101	5713	3003	5240	3267
MAPE	0.057	0.152	0.082	0.140	0.095
RMSPE	0.067	0.155	0.088	0.141	0.108

Nils Karl Sørensen ©
Advanced Tools Statistics

Set 4: Methods in Sampling

by Nils Karl Sørensen

Outline ***page***

1. Issues on Sampling

If you are working with for example marketing, one of the central topics is to find out the preferences of the consumer. How do we sell our product? Who are in our primary consumer segment? Do we supply what the consumer need etc.? Such issues can only be investigated by a market survey. For such a purpose, we need to sample. This is the motivation for this note. We need a list of the relevant consumers to contact. Sometimes we have a list of all the relevant consumers to talk with. In this case it is easy. We just have to sample. In most of the cases, the task is not so easy. What do we do in such a situation?

Interviews based on sampled data are only one out of several other methods to gather information concerning the market behavior of the consumer. Other methods are *focus groups* and *consumer test panels*. The present notes are restricted to sampling only.

In Bowerman in the basic statistics course ("Tools for Quantitative analyses, part II"), we dealt a little with *random numbers.* We used Excel or Megastat to draw random numbers, and we showed that a large sample was better than a small sample.

We also dealt with sampling when we proved the *central limit theorem.* In particular, we listed a number of conditions for a sample estimator to be *representative* for the total population. The conditions to be met are:

- The sample estimator should be *unbiased*
- The sample estimator should be *efficient*
- The sample estimator should be *consistent*
- The sample estimator should be *sufficient*

The sample is *unbiased,* if the population mean and the sample mean are equal, and the shape of the distributions are similar. The sample is *efficient*, if the population variance is properly transformed into the sample variance. The sample is *consistent*, if it relative to its size contains all the information in the total population. Finally, the estimator is *sufficient*, if the calculations based on the sample is close to the results based on the total population. We shall especially deal with the two first issues.

When conducting statistical research there are two ways of obtaining data.

1. ***Second data sources*** such as UN, EUROSTAT or National Statistical Offices
2. ***Primary data sources.*** In such a case we collect our sample personally

When we have second hand sources we know the total population. This also means that we are in a position to calculate the mean and the standard deviation of the total. We can also draw a histogram or similar of the total population. This is important, when we want to compare the total population with the sample.

Why do we not use the total population as the point of departure for our research? In many cases, this is not handy. Consider for example, a situation, where we have a dataset of 50,000 individuals, and we are interested in their preferences for buying a specific product. If we want to conduct a telephone interview or an internet based survey; an investigation of a sample of this size is intractable and also too costly.

Instead, it is possible to do interviews with say 500 persons. In such a case, we are interested in drawing a sample equal to 500/50,000 = 0.01 or 1 percent of the total population. How do we draw this list of candidates for the interview?

How large should a sample be? Many times it is the budget of the investigation that determines this question! A better alternative is to use the formulas given in Bowerman in the chapter on *hypotheses testing.* Here we gave tolerances of the mean and standard deviation, and with a given level of significance, we were able to estimate the size of the sample required in order to meet the conditions[1].

Another reason for working with the sampled data is that we only want a list of persons, firms or other units of items that we can use for the questionnaire. Then in the questionnaire we shall ask the units about issues of relevance for our problem in consideration. This could for example be a market survey for the preferences of a given product.

When doing research that involves sampling a procedure must include the following steps:

1. Choice of total population and characteristics (variables)
2. Setting up units of measurement and sampling frame
3. Taking the sample
4. Examining the validity of the sample
5. Use supplementary analysis in order to obtain a sample of the desired size
6. Conduct an investigation of the non-respondents, response rate etc.
7. Using the sample for the purpose

For the present, we focus on the process of sampling. A completely other issue, is the *design of the questionnaire.* We will not use much space on this issue, but the interested reader should consult the literature on marketing research. A few remarks will be made in the note on *Nonparametric Methods.*

Finally, the *response rate* is important. If it is too low, it is important to have additional elements to use in or to supplement. A response rate should exceed minimum 60 percent. However, by use of special techniques it is possible to obtain valid materials with a response rate as low as 10 percent. In such a case the so-called Heckman procedure is used. This is beyond the outline in the present course.

[1] See Bowerman Section 8.3 and 8.4.

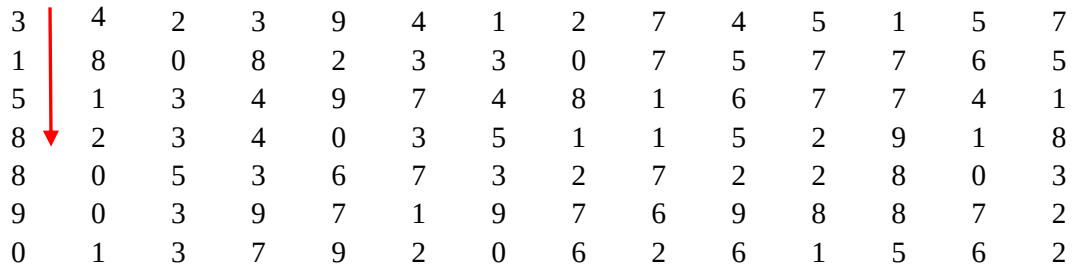
2. How to Sample

Methods for sampling are called *sampling designs*, and the sample we take is called a *sample survey*. The most common used methods in sampling are:

- Random sampling
- Stratified random sampling
- Systematic sampling
- Cluster sampling

In order to do random sampling, we need a list of random numbers. A list could look as:

Random Numbers



3	4	2	3	9	4	1	2	7	4	5	1	5	7
1	8	0	8	2	3	3	0	7	5	7	7	6	5
5	1	3	4	9	7	4	8	1	6	7	7	4	1
8	2	3	4	0	3	5	1	1	5	2	9	1	8
8	0	5	3	6	7	3	2	7	2	2	8	0	3
9	0	3	9	7	1	9	7	6	9	8	8	7	2
0	1	3	7	9	2	0	6	2	6	1	5	6	2

This is just a list of any numbers ranging between 0 and 9, drawn by Excel.

Random Sampling

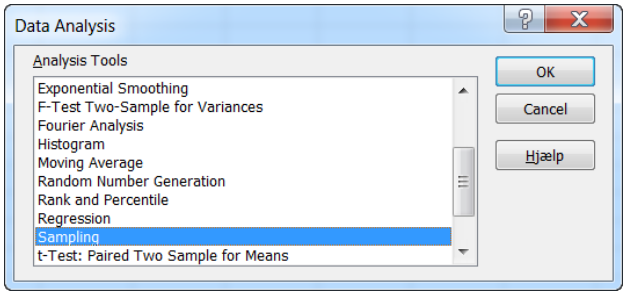
This is straight forward. We apply for example the table above. A *sample plan* is needed for the task. For example we could take all items on the vertical list.

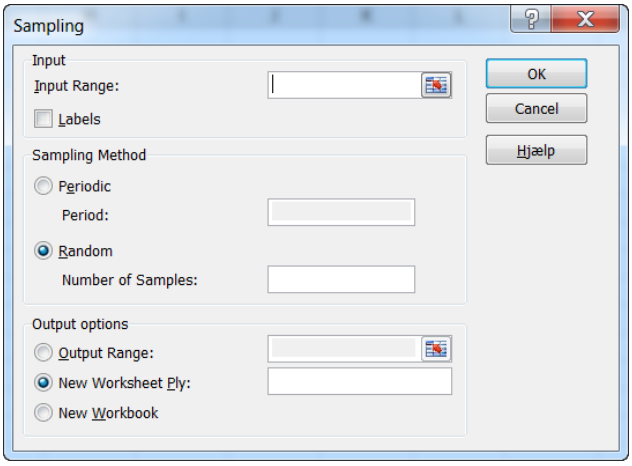
The first item to be selected is item 3, then item 4 (3+1), then item 9 (3+1+5), etc. We proceed until the size of the sample to be used is reached. If the end of the list is reached we just start from the top again.

We can test the validity of the sample by conducting a descriptive statistical analysis of the total population as well as the sample and compare. In this way, we compare the total population with *N* elements with the sample with *n* elements. If the sample is taken correctly the total mean and the sample mean should not be significantly different from each other, and the standard deviation of the sample should be larger because the number of observations in the sample is smaller. As a result the divisor in the standard deviation is smaller, and the numeric value of the standard deviation will be larger.

Random Sampling in Excel and Megastat

How can the table of random numbers above be generated by use of Excel or Megastat? In Excel there are two different options. Select ***data/data analysis/sampling*** as shown in the left panel and obtain the box to the right.





Here the following can be identified:

- Input range: Mark the total population
- Sampling method:
  - If random: Denote the size of the sample. For example if the total is 500 the sample could be equal to 50
  - If period: If for example 7 is marked then every 7^{th} observation is picked. This is also called *systematic sampling*, see below

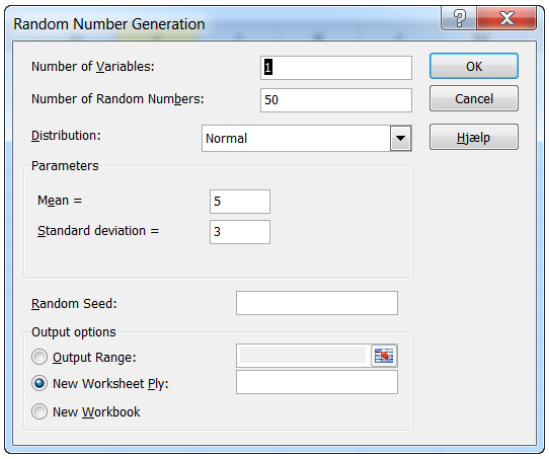
This procedure cannot be found in Megastat. Here use the table of random numbers given above, and use the procedure of counting illustrated above.

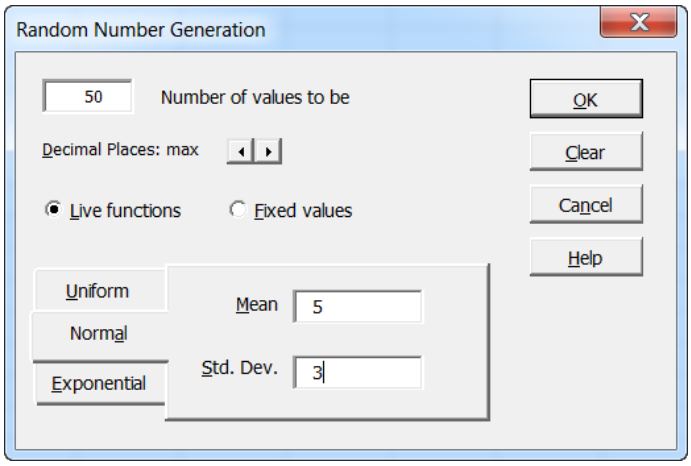
An alternative to this is to use the function *random number generation.* This facility is found in *Data Analysis*. This facility can be used to generate data of a given distribution. Consider an example.

In Excel type ***data/data analysis/Random Number Generation*** and obtain the dialog bow to the left on the next page. The menu box on the right panel is obtained by application of Megastat. In this case type ***Megastat/Random Number Generation.*** The two functions give similar (random) results.

Both functions allow drawing data set following specific distributions like for example data that are uniform or normal distributed etc.

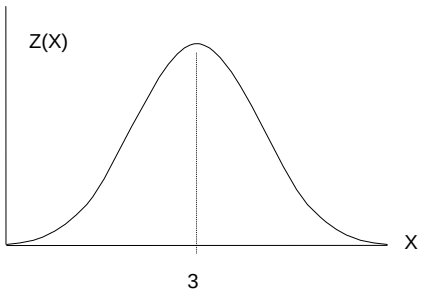
In both cases, we have drawn a sample of 50 observations from a total population assumed to be normally distributed with mean equal to 5 and standard deviation (uncertainty) equal to 3 for a single variable.





Excel ***Megastat***

Theoretically our normal distribution of 50 items can be illustrated as:



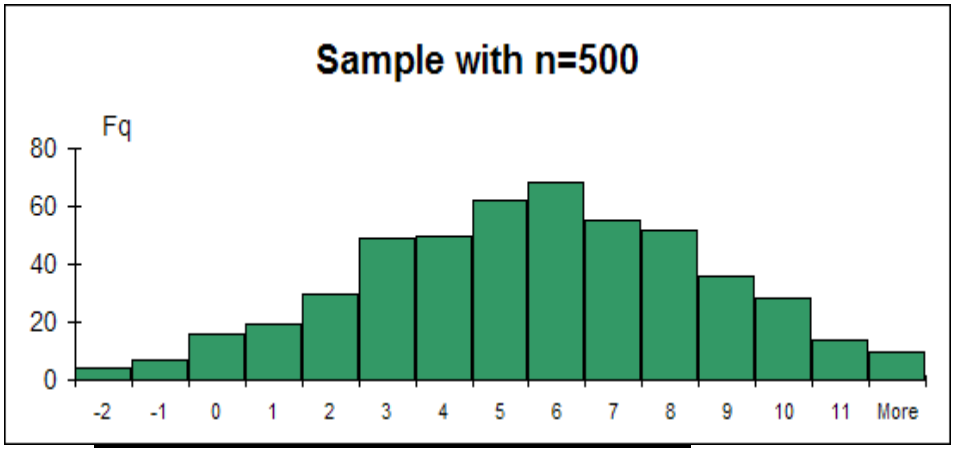
Let us now use one of the two programs to conduct a little experiment. What will happen to the distribution when the number of observations increases from 50 to 500?

Next page gives the results by use of Excel. In order to keep the data sets handy, I have sorted the two data sets into 14 groups or categories ranging from –2 to 11. Next, I have displayed two histograms, so I can compare my results.

I have also used the *descriptive statistics* function to calculate the mean and standard deviation to see how good my samples fit my priors.

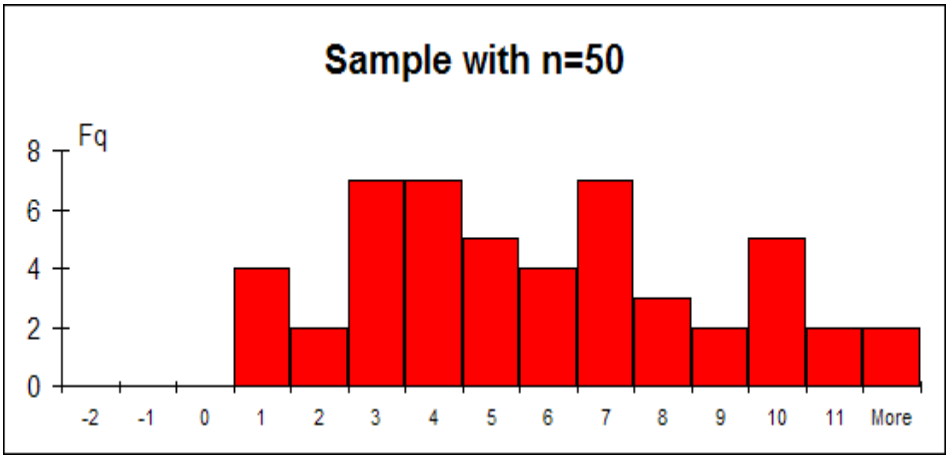
Normal Distribution with Mean = 5 and Standard Deviation = 3

N=500	*Frequency*
-2	4
-1	7
0	16
1	19
2	30
3	49
4	50
5	62
6	68
7	55
8	52
9	36
10	28
11	14
More	10
Sum	500



Descriptive statistics	*Norm 1*	*Norm 2*
Mean	5.13	5.42
Standard deviation	3.05	3.09
Observations	500	50

n=50	*Frequency*
-2	0
-1	0
0	0
1	4
2	2
3	7
4	7
5	5
6	4
7	7
8	3
9	2
10	5
11	2
More	2
Sum	50



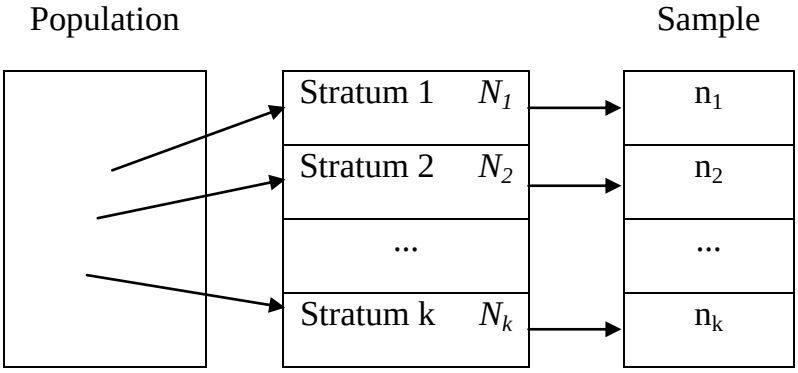
From the graphs, it is evident that a larger sample is closer to the theoretical illustration shown on the previous page. This is also what to be expected. A larger and *correctly* drawn sample is more representative. Consequently, the results will be better.

Stratified Random Sampling

A problem with the method of random sampling is that we easily run into bias if the order of the lists of items that we sample from not is completely random itself. Instead a stratified random sample can be set up.

In order to select a *stratified random sample,* we divide the population into non-overlapping groups of similar elements. So within each group our data should be distributed as homogenously as possible. These groups are called **strata**. Then a random or systematic selected sample is selected from each stratum, and these samples are combined in order to form the full sample.

We sample with regard to the variable of interest. Schematically, the set up can be illustrated as:



To determine the size of the strata, we can apply a methodology that frequently is used when setting up a histogram. To determine the width of the intervals or groups used a formula where:

$$2^k = N$$

Here *k* is the number of groups. Notice that the strata do not need to be of equal size. Doing it in this way is preferable, because no weighting of data is need when we calculate the sample mean and variance. A rougher, but absolutely efficient method is to let Excel determine the bin range of the histogram, and use that for the division into strata.

Inside each stratum, we can select elements by use of random numbers just as before.

Consider a situation with a total population equal to 500 observations where we want to take a sample of 10 percent or 50 items by use of stratified sampling.

In practice a “cookbook” for sampling with stratified material could look as:

- Find the relevant variable to be used for the analysis
- Set up a histogram in order to find the distribution of the data set
- Find the number of strata for the sampling process. For example if our data set has 500 observations then 9 strata could be appropriate as $2^9 = 512 \sim 500$
- Sort data for example in ascending order and organize the material
- There are 500 observations with 9 strata. However the intervals may *not* be of equal size. It could look as:

Strata	1	2	3	4	5	6	7	8	9	Total
Frequency	20	40	55	100	80	70	60	40	35	500
Share	0.1*20	0.1*40	0.1*55	0.1*100	0.1*80	0.1*70	0.1*60	0.1*40	0.1*35	0.1*500
Numbers	2	4	6	10	8	7	6	4	3	50

Note: I have rounded up in strata 3, and down in strata 9. This is arbitrary!

- Use *random sampling* or *systematic sampling* to find the number of items to be selected within each strata.
- Finally, provide a descriptive statistical analysis of the total and well as the sample and compare

Systematic Sampling

This is an alternative to random sampling. For example, within a stratum 5 elements should be selected out of 10 elements. We select every second element.

Sometimes we do not have a complete list of elements to sample from. For example, if we want to do an interview in front of a supermarket. Systematic sampling will then be to select for example every 100^{th} shopper that passes in or out of the supermarket depending on the investigation to be undertaken.

A variation of systematic sampling is today's television and radio stations use of ***voluntary response samples.*** In such samples, participants self-select – that is, whoever wishes to participate does so (usually expression some opinion). These samples over represent people with strong opinions. Such samples are then biased, and should be used with care.

Cluster Sampling

Sometimes it is advantageous to sample in stages. In cluster sampling, we first select groups or clusters and then sample. The method is also labeled *Deming method*. It has been frequently used in the United States for example when having a pool of voters in a system where these have to register. Such a procedure can be undertaken in four steps:

Stage 1: Randomly select a sample of county's from all states in the United States

Stage 2: Randomly select a sample of townships from the sample under stage 1

Stage 3: Randomly select a samples of voting precincts from each township from stage 2

Stage 4: Randomly select a sample of voters from each voting precincts from stage 3

This method is also applied for selecting families to participate in for example consumer surveys.

3. Cases

Case I: Drawing a Sample to use for Interviews of Heads of Housing Associations in Denmark 2010

For the Danish Ministry of Interior a sample has been conducted for an internet based questionnaire. In 2008, Danish housing associations could apply for grants if they wanted to renew for example kitchens, bathrooms, pluming installations etc. The ministry wanted to ask 18 different questions in order to see how the preferences were among the housing associations.

For the purpose, the total population was drawn from Statistics Denmark. The total population was delivered as an Excel file with addresses. Initially there was 7,646 Housing Associations or sub sections.

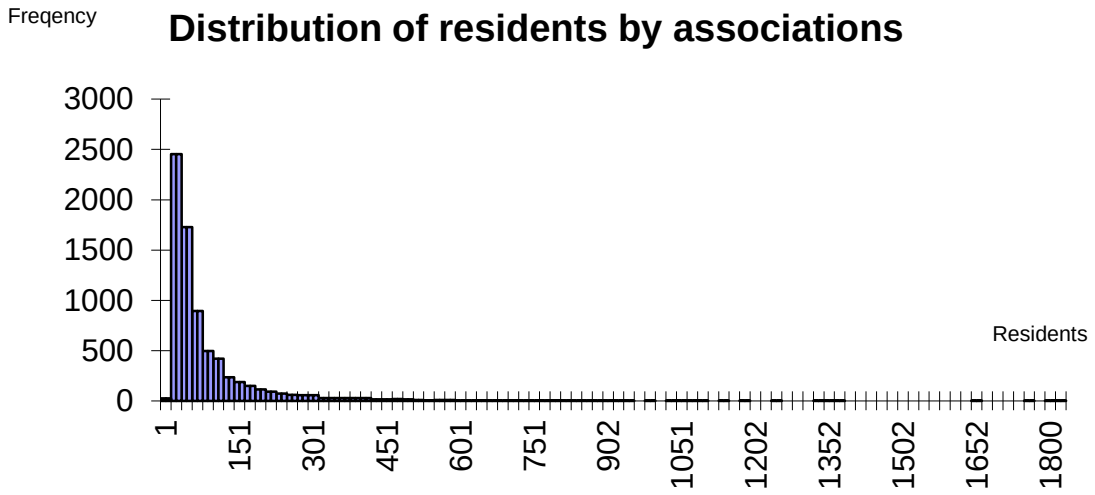
An investigation of the list revealed that for 309 associations no full information was available. For example, the year of establishment (building) could be missing. This is an important parameter when asking for activities related to renewal, because older associations are more inclined to renew than the new associations. This was the case for 195 associations. Further, 144 associations had other kinds of missing information. These associations were sorted out. A total of 309 associations were sorted out.

This process resulted in a total population equal to 7,337 housing associations. In sum these associations amounted for 547,005 residencies (houses, apartments, studios etc.). The 309 associations sorted out amounted for 12,478 residencies or 2.23 percent of the total population. It was evaluated that this number of residencies did not have an impact on the outcome of the investigation.

It was decided to sample 800 elements or 10.90 percent of the total population. The sample was then divided into two sub samples each with 400 associations labeled 400A and 400B respectively. The plan was then to use sample 400A for the survey, and use sample 400B as a as backup.

Initially a histogram was set up for the total population. The distribution of residents by the size of the associations is shown on the next page. A very skewed to the right population is observed. The majority of associations are very small, but a few are very large like for example Toveshøj or Gellerup in Aarhus or Albertslund in Copenhagen West. Further, it was known that these large associations frequently have a low rate of respond. Consequently, the Ministry wanted these associations to be a little over represented in the sample. The Ministry was not interested in associations below 5 residents. Therefore, these were not selected for the sample.

On this basis a system with 85 strata was selected by size of the association. This large number was needed in order to capture the large housing associations. In each stratum 11 percent of the associations were selected by use of a systematic selection procedure. Finally, the sample of 800 items was divided into two equal large samples of 400 items.



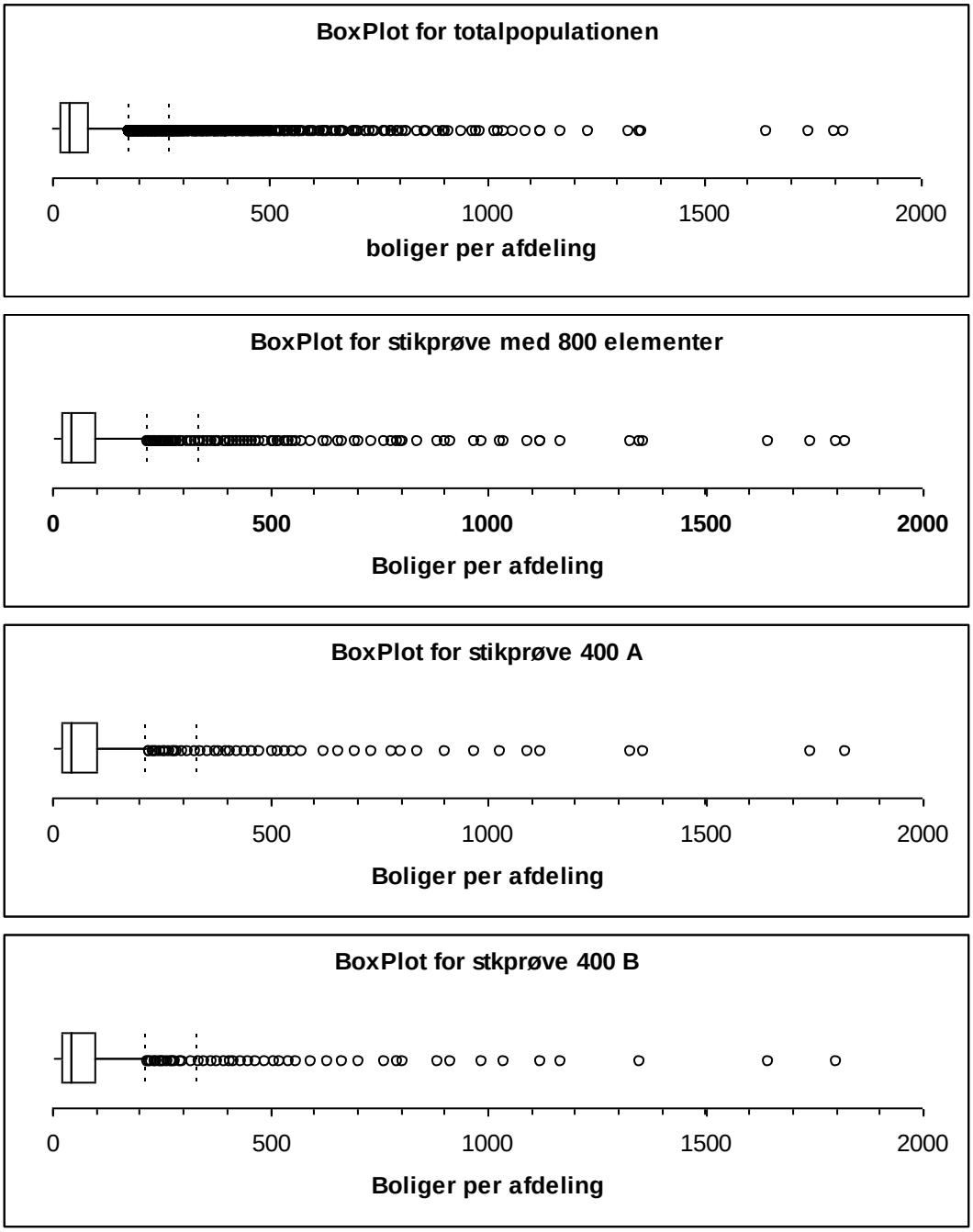
The division of data into strata is shown in the Appendix. We can now perform a descriptive statistical analysis of our relevant parameter namely the number of residents by association:

Descriptive Statistics

	Total	800	400 A	400 B
Mean	74.55	113,91	115.05	112.77
Median	36	42	42	42
Mode	12	12	12	12
Standard deviation	117.76	218,31	222.31	214.50
Variance	13,867.96	47,658,61	49,422.08	46,012.00
Kurtosis	43.68	22,88	23,04	22.93
Skewness	5.18	4,36	4,38	4.35
Range	1,823	1,818	1,818	1,795
Minimum	1	6	6	6
Maximum	1,824	1,824	1,824	1,801
Sum	547,005	91,126	46,018	45,108
Observations	7337	800	400	400

Notice, that the mean is large in the samples. This is expected, because the large associations are given higher priority. The variances for the samples are also larger. This is expected because the number of observations is smaller. Observe that the median and the mode are equal for all samples. Very nice!

If we conduct a test for equal mean assuming unequal variance then H_0 is accepted. So the samples are correct. Finally, consider the very nice box-plots below (sorry in Danish headlines):



Notice, that the distributions look very similar.

Analysis on Background Parameters

We can consider the year of establishment of the associations in the total population and in the samples. A descriptive analysis can be found below. Here similarity is also observed. In the original project a distribution by municipality was also considered. Also in this case, consistence was observed. Due to space limitations it has been left out.

Statistics by year of establishment

	Total	***800***
Mean	1979	1978
Median	1983	1982
Mode	1990	1991
Standard deviation	19.42	19.20
Variance	377.25	368.68
Kurtosis	-0.69	-0.74
Skewness	-0.52	-0.39
Range	98	94
Minimum	1911	1915
Maximum	2009	2009
Observations	7,337	800

Case II: Sampling from the Data set "Euroregions"

At the regional level the members of the European Union can be divided into a total of 356 regions for 2004. This is undertaken in the Excel file *Euroregions.xls*. This file can be found in Blanckboard. For all regions statistics of the income per capita is available from EUROSTAT. The result of the mean and the standard deviation is shown in the bottom line in the table below.

Alternatively the mean and the standard deviation could be calculated from the national data not divided by regions. The result of this calculation is shown in the upper line in the table below. In the Excel file *Euroregions.xls* the folder *Euronations* gives the data used for the calculations.

	Mean	**SD**	**Obs**
Data set Euronations (based on national average)	19,984.90 €	13,624.02 €	26
Data set Euroregions (based on regions)	20,609.28 €	10,439.27 €	356

We can consider the data set calculated on the national data as a sample of the regionalized data set. How good is the mean income calculated at the national level as an indicator for the mean income calculated at the regional level? In other words: Is the national mean income a representative indicator the mean income calculated by use of the regional statistics?

We can examine the validity of the national calculated income by setting up confidence intervals or by examination for equal mean. If overlap of the confidence intervals is observed the sample is a good predictor for the total population.

Set up a 95 % confidence interval for mean for each data set. In general we apply the formula:

$$\overline{X} \pm t_{\alpha/2(n-1)} \frac{s}{\sqrt{n}}.$$

with degrees of freedom *df =n–1*. For the *Euronations* data set we obtain:

$$\overline{X} \pm t_{\alpha/2(n-1)} \frac{s}{\sqrt{n}} \Rightarrow 19{,}984.90 \pm 2.060 \frac{13{,}624.02}{\sqrt{26}} \Rightarrow$$
$$19{,}984.90 \pm 2.060(2{,}724.81) \Rightarrow 19{,}984.90 \pm 5{,}613.11 \Rightarrow$$
$$[14{,}371.79; 25{,}598.01]$$

With *df =n–1 = 26 – 1 = 25*. Assuming *(α/2=0.025)* we find by use of the **Statistics Tables** that *t = 2.060.* For the *Euroregions* data set we obtain:

$$X \pm t_{\alpha/2(n-1)} \frac{s}{\sqrt{n}} \Rightarrow 20{,}609.28 \pm 1.96 \frac{10{,}439.27}{\sqrt{356}}$$
$$20{,}609.28 = \pm 1.96(553.28) \Rightarrow 20{,}609.28 \pm 1{,}084.43 \Rightarrow$$
$$[19{,}524.85\,;\, 21{,}693.71]$$

with degrees of freedom *df =n–1 = 356 – 1 = 355* $\approx \infty$. Assuming *(α/2=0.025)* we find by use of the **Statistics Tables** that *t = 1.96* (Notice that the t-distribution is approximate to the normal distribution).

Can the data set Euronations be said to be a good description for the data set Euroregions?

- This must be the case. The mean of *Euronations* falls inside the confidence interval of *Euroregions.*

- The confidence interval of *Euronations* is much larger than for *Euroregions*. This is so because the former is a sample and much smaller than the total population 25/356 = 7 %
- However, the mean of *Euronations* is lower than for *Euroregions.* This is so because there are many small Eastern European nations with a low income level per capita.
- This gives a negative bias. For example has Estonia the same weight in the sample as Germany

Appendix:

Strata for the sample in the case on residents by associations in Case I

Interval	*Frequency*	Share 11 %	In the sample
1	23	2.53	0
22	2453	269.83	240
43	1728	190.08	185
65	894	98.34	100
86	495	54.45	55
108	419	46.09	45
129	234	25.74	25
151	188	20.68	20
172	149	16.39	20
194	115	12.65	15
215	91	10.01	10
236	72	7.92	8
258	60	6.6	7
279	57	6.27	6
301	58	6.38	6
322	29	3.19	3
344	29	3.19	3
365	28	3.08	3
387	29	3.19	3
408	29	3.19	3
429	14	1.54	2
451	16	1.76	2
472	17	1.87	2
494	14	1.54	2
515	9	0.99	1
537	6	0.66	1
558	8	0.88	1

580	10	1.1	1
601	3	0.33	1
622	2	0.22	1
644	5	0.55	1
665	3	0.33	1
687	3	0.33	1
708	4	0.44	1
730	2	0.22	1
751	3	0.33	1
773	3	0.33	1
794	4	0.44	1
815	2	0.22	1
837	2	0.22	1
859	1	0.11	1
880	2	0.22	1
902	2	0.22	1
923	2	0.22	1
944	1	0.11	1
966	0	0	0
988	3	0.33	1
1009	0	0	0
1030	2	0.22	1
1051	1	0.11	1
1073	1	0.11	1
1094	1	0.11	1
1116	0	0	0
1138	2	0.22	1
1159	0	0	0
1180	1	0.11	1
1202	0	0	0
1223	0	0	0
1244	1	0.11	1
1266	0	0	0
1288	0	0	0
1309	0	0	0
1331	1	0.11	1
1352	1	0.11	1
1373	1	0.11	1
1395	0	0	0
1416	0	0	0
1437	0	0	0
1459	0	0	0
1481	0	0	0
1502	0	0	0
1523	0	0	0
1545	0	0	0

1566	0	0	0
1588	0	0	0
1609	0	0	0
1631	0	0	0
1652	1	0.11	1
1673	0	0	0
1695	0	0	0
1716	0	0	0
1738	0	0	0
1759	1	0.11	1
1781	0	0	0
1800	1	0.11	0
Over 1800	1	0.11	1
Sum	7337	800	800
		807.07	

Nils Karl Sørensen ©
Advanced Tools Statistics

Set 5: Nonparametric Methods

by Nils Karl Sørensen

Outline

page

Literature

Bruce L. Bowerman and Richard T. O'Connell, 2007, *Business Statistics in Practice.* 4th edition. McGraw-Hill. Chapter 15, *Nonparametric Methods.*

This Chapter can be found in Blackboard under *course materials* folder *literature.*

1. Questionnaires and Nonparametric Methods

Questionnaire Design

Many questionnaires claims that the respondent answers by ranking his/hers preferences on a given issue. There may for example be five options ranging from "1" equal to "extremely not satisfied" over "3" "neutral" to "5" equal to "very satisfied". Such an *ordinal* scale is called a *Lickert scale.* An example could be:

Question example:
How did you like the party in the student bar last Friday?

Answer example:

☐	☐	☐	☐	☐
Hated the party	Did not like the party	Neutral	Liked the party	Liked the party very much

All questions should "turn" the same way each time (worse/best or reverse). In the questionnaire, the question above may be accompanied with several other questions needed in order to obtain the relevant information. Here the following things are of importance:

- What kind of information do we need relative to the statement of the problem?
- Don't ask unless it has a purpose
- Think – before asking!

An overlong and unclear questionnaire is NOT wanted. Respondents do not like to use too much time on answering the questionnaire. As a rule of thumb it should not take more than 15 minutes to answer the questionnaire.

Notice, that once the questionnaire is launched it cannot be changed! The design of the questions is therefore extremely important. The content of the questionnaire should reflect the statement of problems for the investigation very strictly.

What kind of information should be achieved from the questionnaire? There are four kinds of information:

- Knowledge - what people know (true or factual)
- Attitudes - preferences (past/future/present)
- Behavior - what people do
- Attributes - what people are

The question will frequently be a mix. First, there will be background information such as gender, age etc. Notice that for example a question on income level of the respondent may be difficult to answer due to the many types of income i.e. household income, pretax income, disposal income etc. Try to be as precise as possible.

Appendix I give an example of a questionnaire. The example attempts to reveal the preferences for Online Clothing Shopping among university students.

Coding and handling of Data

The answers can be coded for example in an Excel spreadsheet, a SPSS data file or a specific program designed for questionnaires like Survey Exact or similar. For a given question like the one in the example above we may code with the following numbers:

1	2	3	4	5

A *Lickert* scale will normally have an *unequal* number of outcomes. Why? This is due to the theory of distribution. If the answers are assumed to be normally distributed the mean outcome will be "neutral". (code = 3). The Lickert scale may have 3, 5, 7 or 9 categories of answers. My experience is that 5 is a sufficient number in nearly all cases.

If the survey is large it may be a good idea to build up a code book with instruction on "how to do". Appendix II gives an example of how a system for coding can be build up in an Excel or SPSS spreadsheet.

If the respondent not is answering the question it should simply be associated with an "empty space". The number zero is not appropriate. Why? If zero is included it will be counted as a number, and impose a positive or negative bias on the final result. Blank answers should be taken out of the analysis and treated separately.

We can use our data with answers and calculate descriptive statistics and present the material in histograms or similar. We can sort the main questions with regard to the background variables gender, age, income etc. and obtain more detailed information. We can also perform various tests, investigations for independence, regression etc.

BUT wait a minute! The purpose of performing this ranking is to reveal the preferences of the respondent. In a microeconomic sense we are trying to find the utility maximizing allocation of the respondent or consumer. Originally, marketing took its point of departure from the consumer theory in microeconomics. So the statistics that we obtain from the answer of the question on the party last Friday put forward in example above is a ranking of the satisfaction rate. As remembered (hopefully) from consumer theory, statistics of utility are *rankings* of preferences. The distance from one level of utility to another does depend on

an *individual* interpretation of the distance from utility level 3 (neutral) to 2 (satisfied). This feature of our data has implications for the conducted statistical analysis. Therefore, a new class of tests have to be developed taking into account that data are rankings of preferences.

Nonparametric Methods

The ranking of the preferences introduces namely a measurement problem. For example if a consumer gives the product rank "2", and another consumer gives the same product rank "4" we cannot say that the last consumer prefer the product twice as much as the first one. The ranking is an *ordinal variable.* Frequently ranked or *ordinal* statistics may be very skewed and non-normal in behavior.

As shown in the notes to Chapter 2 and 3 in Bowerman on descriptive statistics, the median may in such cases be a more stable measure than the mean. In the example, it was found that in the case of outliers, the mean and the variance increased very much in the presence of an outlier, whereas the median remained constant. For example, wage distributions may be very non-symmetric with a few persons with very high incomes and many on a lower level. When undertaking wage negotiations, the median wage is normally the wage to be negotiated. This is so, because this is the wage that is important to most of the employees. Official wage statistics also uses the median wage. Another example could be student's evaluations of the lecturers' ability to teach. Suppose that the lecturer receives a good evaluation from a very large number of the students. A little minority has the opinion that the lecturer is very inefficient and gives the lecturer a very bad evaluation. In this case, the distribution of the evaluations will also be very non-symmetric and a test is required taking this into consideration.

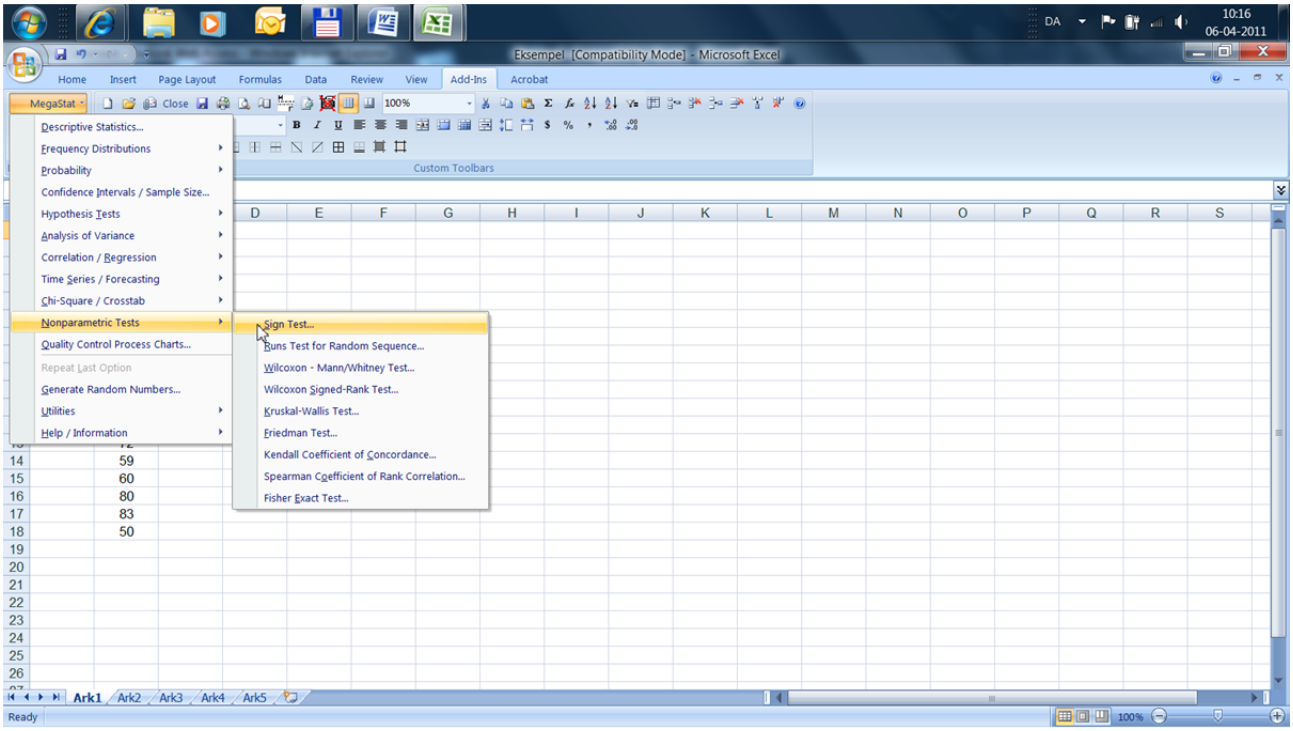
Nonparametric tests are a class of methods used when the underlying assumption of Normal distribution (or t-distribution if the sample is small) not is fulfilled. Here we consider 3 tests namely:

- Sign test — For a single data set — - equivalent to the t-test
- Mann-Whitney U test — For two data sets — - inferences on two samples
- Kruskal-Wallis test — *R* data sets — - equivalent to ANOVA

The nonparametric tests are also referred to as *distribution free tests.* Together with the *chi-squared test* presented in Bowerman Chapter 12, these tests are very frequently used when working with questionnaires.

Nonparametric Methods with Megastat

Nonparametric tests can most easily be undertaken in Megastat. Open Excel with Megastat loaded and click on *add-ins*. Here select *nonparametric tests*. A menu with a range of tests will occur. The maximum number of observations allowed in Megastat is restriction to 180. Alternatively, SPSS can be used[1]. This issue will be covered in another set of notes.



2. Testing for Normality in Data

Very frequently we assume that the underlying data generating process can be described by the Normal distribution for example in the notes on ANOVA to Chapter 11 in Bowerman. Remember from the notes to Chapters 2 and 3 in Bowerman that we defined skewness and kurtosis. These two measures provide information with regard to the asymmetry and concentration of a given data set. Then the question arises: How skewed or concentrated can a distribution be before it not is considered as Normal distributed?

Bowman and Shenton (1975) have provided a test for this issue using the measures of skewness and kurtosis. Mathematically the measures may be defined as:

[1] In SPSS use the *data view* mode, select *Analyze* and then *nonparametric* methods.

Skewness: $$SK = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 / n}{s^3} \qquad i = 1,2,...,n$$

Kurtosis: $$KU = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 / n}{s^4} \qquad i = 1,2,...,n$$

Where x_i is observation *i*, *n* is the number of observations, $\bar{x}$ is the mean *(the first moment)*, and *s* is the standard deviation *(the second moment)*. In Excel the number 3 is subtracted from kurtosis. Skewness is also called the *third moment* and kurtosis is the *fourth moment.*

Remember from the notes to Chapters 2 and 3 in Bowerman that

- Skewness: An expression for how much the distribution is away from the ”normal”. If SK>0 data are skewed to the right, if SK=0 data are symmetric, and if SK<0 data are skewed to the left.

- Kurtosis: A measure of the ”concentration” of the distribution”. If KU is large then we have a concentrated data set, and if KU is small we have a “flat” distribution.

Let us now define the hypotheses:

H_0: The data set can be described by a Normal distribution.
H_1: The data set can *not* be described by a Normal distribution.

Bowman and Shenton now set up a tester that we will label by *B*. This is:

$$B = n \times \left[\frac{(SK)^2}{6} + \frac{(KU)^2}{24} \right]$$

For $n \to \infty$ the tester[2] is chi-squared distributed with degrees of freedom equal to 2 ($\chi^2_{(2)}$). However, if *n >100* there is a bias towards rejecting the null hypothesis although it may be true. To correct for this Bera and Jarque (1981) has simulated the critical values shown in the table next page.

[2] The numbers ”6” and ”24” comes from the deviation of the tester. The method involves a statistical way of thinking beyond the scope of this course. However, the deviation is given in D’Agostino, R. and E.S. Pearson (1973).

Critical Values for Bowman-Shenton Test

Observations, n	***10 % significance***	***5 % significance***	***Observations, n***	***10 % significance***	***5 % significance***
20	2.13	3.26	200	3.48	4.43
30	2.49	3.71	250	3.54	4.51
40	2.70	3.99	300	3.68	4.60
50	2.90	4.26	400	3.76	4.74
75	3.09	4.27	500	3.91	4.82
100	3.14	4.29	800	4.32	5.46
150	3.43	4.39	∞	4.61	5.99

Example

Consider the data set analyzed in the notes to Chapters 2 and 3 in Bowerman. By use of the descriptive statistics function in Excel we can calculate skewness and kurtosis. We find that $SK = -0.35$ and $KU = 0.12$. The data set considered has 20 observations. The tester equals

$$B = n\left[\frac{(SK)^2}{6} + \frac{(KU)^2}{24}\right] = 20\left[\frac{(-0.35)^2}{6} + \frac{(0.12)^2}{24}\right] = 0.42$$

With 20 observations the critical value at 5 % level of significance is equal to 3.26 as observed from the table above. Because $0.42 < 3.26$ we accept H_0. So the data set can be said to be Normal distributed.

We also considered an example where an extreme observation was added to the dataset. For example if the value of the maximum was increased from 24 to 34 then $SK = 1.19$ and $KU = 3.88$. The now tester equals

$$B = n\left[\frac{(SK)^2}{6} + \frac{(KU)^2}{24}\right] = 20\left[\frac{(1.19)^2}{6} + \frac{(3.88)^2}{24}\right] = 17.27$$

As $17.27 > 3.26$ we accept H_1. So the distribution easily becomes non-normal!

References

Bera, A.K. and C.M. Jarque (1981): "An Efficient Large-sample Test for Normality of Observations and Regression Residuals" *Working Papers in Economics and Econometrics* no. 40, Australian National University.

Bowman, K.O. and L.R. Shenton (1975): "Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2". *Biometrika* ***62***. Pp. 243-250.

D'Agostino, R. and E.S. Pearson (1973): "Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$ ". *Biometrika* ***60***. Pp. 613-622.

2. Sign Test

(BO Section 15.1)

If a population is highly skewed, then the median might be a better measure of the central tendency than the mean. Further, if the sample size is very small, then the *t-test* outlined in Bowerman Chapter 9 may not be valid. In such a case, it may be better to set up a hypothesis with regard to the *median* than with regard to the mean.

Exact, we want to test the hypotheses:

H_0: $M_d = M_0$ (The median and the stated median are equal)
H_1: $M_d \neq M_0$ (The median and the stated median are *not* equal)

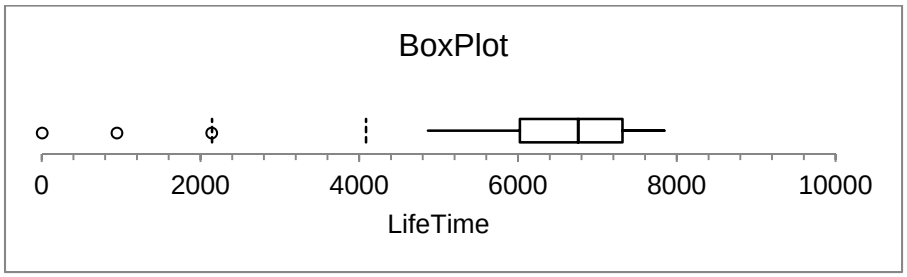
How can a test for the median be developed? Bowerman considers a case of a DVD or compact disc player. The developer of the product wishes to show that the lifetime of the player exceeds 6,000 hours of continuous play[3]. To examine the issue the developer randomly selects 20 new players.

Consider initially a descriptive statistical analysis of the data set:

LifeTime
5
947
2142
4867
5840
6085
6238
6411
6507
6687
6827
6985
7082
7176
7285
7410
7563
7668
7724
7846

	Life Time Data
count	20
mean	5,964.75
Variance	5,160,816.20
Standard deviation	2,271.74
minimum	5
maximum	7846
range	7841
skewness	-1.81
kurtosis	2.32
1st quartile	6,023.75
median	6,757.00
3rd quartile	7,316.25
interquartile range	1,292.50
low extremes	3
low outliers	0
high outliers	0
high extremes	0

[3] 6,000 hours is a lot of time – actually 250 days. In real life, the process will be speeded up, and a simulation process will be used in order to generate the sample data. This process is normally used in many industries for example drilling equipment to off-shore operation (it is not that easy to replace an item for a drill on deep waters).



It is observed that the distribution is highly skewed to the left with 3 low extremes. Especially, the range is very high. A t-test may then be misleading. The Bowman-Shenton test give:

$$B = n\left[\frac{(SK)^2}{6} + \frac{(KU)^2}{24}\right] = 20\left[\frac{(-1.81)^2}{6} + \frac{(2.32)^2}{24}\right] = 9.12 > 3.26 \text{ at the 5 \% level}$$

This is very significant, so the distribution is not normal.

What can be done? By inspection of data it is evident that 5 observations only are below 6,000 hours. This is indicated by the solid line in the table above.

Remember that the median divides data into two parts of equal size. We can describe the distribution as a Binominal distribution with $p = 0.5$ and $n = 20$. By inspection of data we can observe that 15 observations are above 6,000.

The p-value for testing H_0: $p = 0.5$ versus H_1: $p > 0.5$ is the probability computed assuming H_0 is true of observing a sample result that is as least as contradictory to H_0 as the sample result we have actually observed. Since any number of lifetimes out of 20 lifetimes that is greater or equal to 15 is at least that contradictory we have:

$$\textit{p-value: } P(X \geq 15) = \sum_{x=15}^{20} \frac{20!}{x!(20-x)!}(0{,}5)^x(0{,}5)^{20-x}$$

and calculate for $x = 15,16, ..., 20$

By use of the Appendix in Bowerman or **Statistics Tables** we find that $P(X \geq 15) = 0.0207$

If H_0 is true there is only 2.07 percent probability that the distribution will be as the sample above. This implies that it is reasonable to conclude that the median lifetime of the player exceeds the advertised median life time equal to minimum 6,000 hours playing time.

This is only good in small samples. If our sample is larger than 20 observations the table of the Binominal distribution is inefficient, and it is not handy to use a calculator. What can then be done?

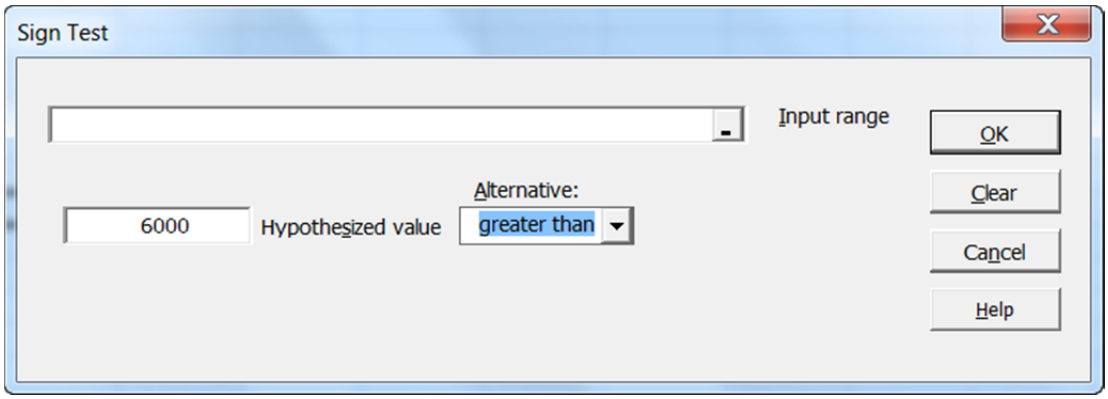
Remember from Chapter 6 in Bowerman that the Binominal distribution can be approximated to the Normal distribution. Define *S* as the number of observations over the hypothetical median . As the median divides the distribution into two parts of equal size $p = 0.5$. For the Binominal distribution mean and standard deviation is given as $\mu = np = 0.5n$ and $\sigma = \sqrt{np(1-p)} = 0.5\sqrt{n}$

Then the *Z* test can by use of the transformation $Z = \left(\frac{X-\mu}{\sigma}\right)$ be written as:

$$Z = \frac{(S-0.5)-0.5n}{0.5\sqrt{n}} = \frac{(15-0.5)-0.5(20)}{0.5\sqrt{20}} = \frac{14{,}5-10}{2.236} = 2.01$$

As 2.01 > 1.96 we reject the H_0 hypothesis. What is the implication? The median in the present dataset is equal to 6,757. This value is significantly higher than 6,000. As this was the claim put forward by the producer, the statement of minimum 6,000 hours life time is valid.

We can perform the sign test by use of Megastat. Use ***add-in / Megastat / Nonparametric Tests / Sign test*** and obtain



We have to mark the hypothesized value (here 6,000) and specify the alternative hypothesis. Loading in the 20 observations will result in the out given below:

Sign Test

6,000	hypothesized value
6,757	median LifeTime
5	below
0	equal

15	above
20	n
	binomial
.0207	p-value (one-tailed, upper)
	normal approximation
2.01	z
.0221	p-value (one-tailed, upper)

The output confirms the results found above.

3. Mann-Whitney Test

(BO Section 15.2)

This test is also called the *Wilcoxon Rank Sum Test* or the *U−test.* We consider two data sets, so the procedure is very similar to the procedure developed in Bowerman Chapter 10. In that test we examined, if the *central tendency* or *locations* were equal among samples. The nonparametric test for comparing the locations of the two samples is not necessarily a test about the difference between the population means. It is a more general test to detect whether the probability distribution of population *1* has shifted to the right or to the left of population *2*.

We assume independence as earlier. As the case with the sign test, the Mann-Whitney test is valid for any shapes that might describe the sampled populations. For each data set, a distribution is given as D_1 and D_2 respectively. The samples have observations n_1 and n_2.

First we combine the data of the two samples. For this we use *ranked* data. This is done in order to bring the data into similar levels. The ranking is done as follows: Rank the $n_1 + n_2$ observations from the smallest (rank 1) to the largest (rank $n_1 + n_2$). If two or more observations are equal, we assign to each "tied" observation a rank equal to the average of the consecutive ranks that would otherwise be assigned to the tied observations.

Next, we for each data set calculate the sum of the rank, and denote them by T_1 and T_2. The outcome of the test can then be examined by the **test statistic T** to be T_1 if $n_1 \leq n_2$ and to be T_2 if $n_1 > n_2$.

The null hypothesis can be stated as:

H_0: D_1 and D_2 are identical probability distributions

The alternative hypothesis is a little bit more complicated:

	Reject H_0 if	
H_1: D_1 is shifted to the *right* of D_2	$T \geq T_U$	*if* $n_1 \leq n_2$
	$T \leq T_L$	*if* $n_1 > n_2$
H_1: D_1 is shifted to the *left* of D_2	$T \leq T_L$	*if* $n_1 \leq n_2$
	$T \geq T_U$	*if* $n_1 > n_2$
H_1: D_1 is shifted to the *right or left* of D_2	$T \leq T_L$	*or* $T \geq T_U$

The critical values of T_L and T_U can for small samples be found in Bowerman Table 15.2 page 741.

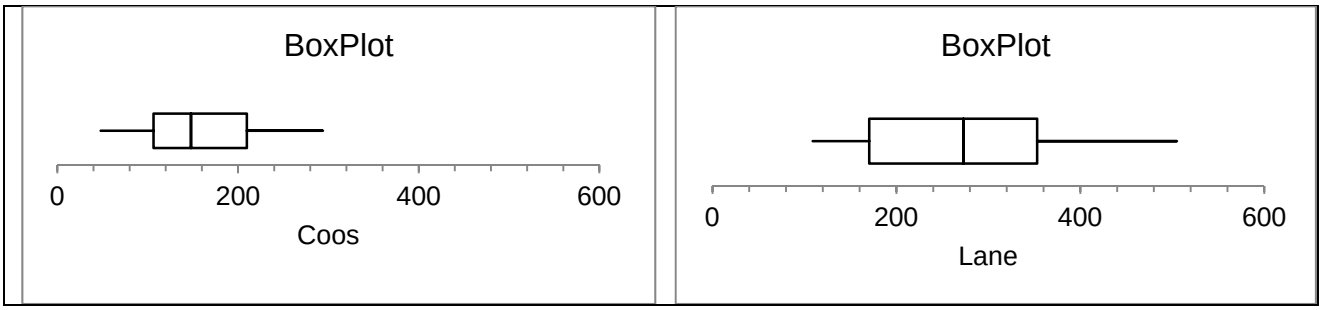
In Bowerman an example is given of processing times for two different courts on similar cases. The hypothesis to be examined is if the processing time is equal among the two courts for a very little data set with $n_1 = 10$ and $n_2 = 7$, see also the descriptive statistics below.

In the example $T_1 = 72.5$ and $T_2 = 80.5$. Using Table 15.2 we for our sample sizes find that $T_L = 46$ and $T_U = 80$.

As $n_1 > n_2$ and $T_2 = 80.5 \geq T_U = 80$ we reject the H_0 and accept the alternative. So D_1 has shifted to the *left* of D_2. Stated differently, the processing time of *Coos* is lower than the processing time of *Lane.* This is confirmed by the Box-plots of the two data sets below.

Coos	**Lane**
48	109
97	145
103	196
117	273
145	289
151	417
179	505
220	
257	
294	

	Coos (D_1)	*Lane (D_2)*
count	10	7
mean	161.10	276.29
Variance	5,914.54	20,658.24
Standard deviation	76.91	143.73
minimum	48	109
maximum	294	505
range	246	396



For a larger data set, the use of the critical values is not handy. Instead, as the case with the sign test, we approximate our data by use of the Normal distribution. The calculation of the mean and the variance of the two pooled data set is given by the following formulas, see also Bowerman page 744:

$$\mu_T = \frac{n_1(n_1+n_2+1)}{2} = \frac{10(10+7+1)}{2} = 90$$

$$\sigma_T = \sqrt{\frac{n_1 n_2(n_1+n_1+1)}{12}} = \sqrt{\frac{10\times 7(10+7+1)}{12}} = 10.24$$

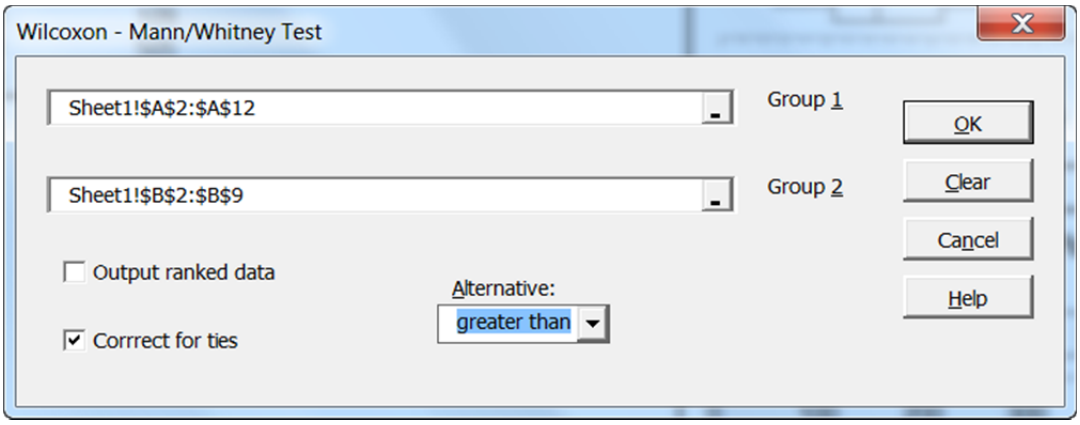
The Z tester is:

$$Z = \frac{(T-\mu_T)}{\sigma_T} = \frac{(72.5-90)}{10.24} = -1.71$$

Where T is the relevant sum of ranks to be tested here T_1. If we had used T_2 the outcome would have been reverse. What is the outcome? If we consider a one-sided test and assume that $\alpha = 0.05$ then $Z = -1.645$. As $-1.71 > -1.645$ we reject the null and accept H_1.

This is consistent with the finding above. Notice that the two values of Z are very close, so the p-value is just below 0.05 (actually it is 0.0436), see below. If we had undertaken a two-sided test the p-value would have been so high that the null hypothesis was accepted. If we inspect our findings above the result is not surprising. The value of $T_U = 80$ is very close to the sum of the ranks for the second sample $T_2 = 80.5$.

We can perform the test by use of Megastat. Use ***add-in / Megastat / Nonparametric Tests / Wilcoxon – Mann-Whitney*** and obtain



Loading in the data the following output will appear

Wilcoxon - Mann/Whitney Test

n	sum of ranks	
10	72.5	Coos
7	80.5	Lane
17	153	total

90.00	expected value
10.24	standard deviation
-1.66	z, corrected for ties
.0485	p-value (one-tailed, upper)

The *p-value* is a little different from mine properly due rounding off.

5. The Kruskal Wallis-Test

(BO Section 15.4)

The Kruskal-Wallis H test is a nonparametric technique for the location of the median for 3 or more data sets. Contrary to the ANOVA procedure it does not require any assumptions about the distribution of data.

Intuitively, the test is identical the ANOVA single factor test with data replaced by their ranks. Hypotheses are also as under the ANOVA procedure.

We have our data divided into p groups. We first rank all of the observations in the p samples from smallest to largest. If n_i denotes the number observations in the ith sample, we are ranking a total of $n = n_1 + n_2 + \ldots + n_p$. What if several observations have similar rank? Then we assign the tied observations the average of the consecutive ranks that would

otherwise be assigned to the tied observations. Megastat has a procedure to undertake this task. Next, we calculate the sum of the ranks of the observations in each sample called T_i. From this the ranked average $\overline{R}_i$ by group can be calculated.

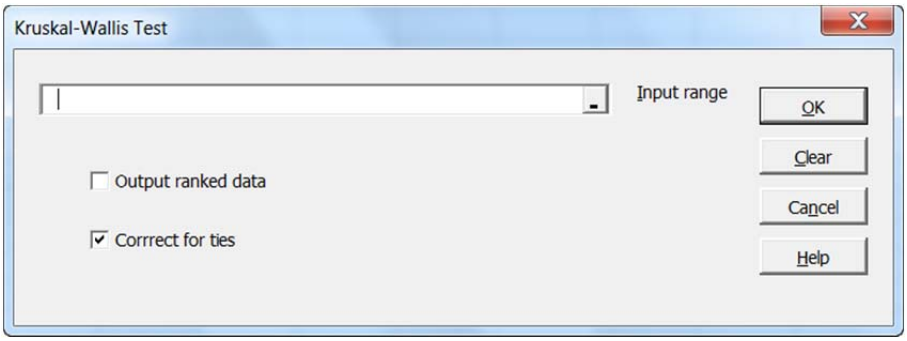
The *Kruskal-Wallis* tester is now found as:

$$H = \frac{12}{n(n+1)}\sum_{i=1}^{p} n_i\left(\overline{R}_i - \frac{n+1}{2}\right)^2 \approx \frac{12}{n(n+1)}\sum_{i=1}^{p}\frac{T_i^2}{n_i} - 3(n+1)$$

This tester is χ^2 distributed with degrees of freedom equal to $(p-1)$ similar to Bartlett's test that was considered in the notes to Chapter 11 in Bowerman.

The Kruskal-Wallis Test in Megastat

The Kruskal-Wallis test can be performed in Megastat by selecting ***Add-in / Megastat / Nonparametric Tests / Kruskal-Wallis Test***. The following screenshot will appear:



Example

In the notes to Chapter 11 in Bowerman, we investigated a case for seasonality in the distribution of advertisements over a year. We used weekly observations of the weight of advertises. Using a one-way ANOVA analysis, we found evidence that the amount of advertisements was higher during the fourth quarter of a year. We also found evidence that the data not was normally distributed. This motivated the use of the Kruskal-Wallis test.

The result from Megastat is given below:

Kruskal-Wallis Test

Median	*n*	*Avg. Rank*	
854.00	13	16.23	1 Q
1,075.00	13	31.81	2 Q
923.00	13	17.62	3 Q
1,436.00	13	40.35	4 Q
1,028.50	52		Total

22.885	H (corrected for ties)
3	d.f.
0.00	p-value

multiple comparison values for avg. ranks

15.68 (.05) 18.69 (.01)

The tester is calculated as:

$$H = \frac{12}{n(n+1)}\sum_{i=1}^{p}\frac{T_i^2}{n_i} - 3(n+1) = \frac{12}{52(52+1)}\left[\frac{(16.23\times 13)^2}{13} + \frac{(31.81\times 13)^2}{13} + \frac{(17.62\times 13)^2}{13} + \frac{(40.35\times 13)^2}{13}\right]$$
$$- 3(52+1)$$
$$= 0.004525(3{,}424.37 + 13{,}154.39 + 4{,}036.04 + 21{,}134.13) - 204 = 181.88 - 159 = 22.88$$

The tester is χ^2 distributed with degrees of freedom equal to *(p−1) = (4−1) = 3*. Assuming a level of significance equal to 5 percent we find that $\chi^2_3 = 7.81$. As $7.81 < 22.88$ the H_0 is rejected and the H_1 is accepted. The result is then that the median amount of advertises is different from quarter to quarter.

In the bottom of the output from Megastat a multiple comparison of the ranks is provided by use of the Mann-Whitney test. It is found that quarter 4 is different from quarter 1 (at the 1 percent level) and from quarter 3 (at the 5 percent level), but not different from quarter 2.

References

Kruskal, W.H., and Wallis, W.A., 1952. Use of Ranks in One-Criterion Variance Analysis. *Journal of American Statistical Association* 47, p. 583-621 (December 1952)

Siegel, S., and Castellan, Jr., N.J., 1988. *Nonparametric Statistics for the Behavioral Sciences*. 2th edition. New York: McGraw-Hill.

Appendix I: Example of a Questionnaire

The example below shows how a questionnaire on online shopping can be undertaken mostly by use of answers based on a Lickert scale.

Online Clothing Shopping Survey

This is a survey being conducted for a marketing research class at a university to determine shopping behavior. Please answer the following questions honestly and thoroughly to help us with our research and final project. Thank you in advance for your assistance in our research project.

1. Are you... ___Female ___Male
2. What is your status at this university? ___Bachelor ___Master

3. Within the past month, how many times have you visited a web site for online shopping purposes?
   ___ 0 ___ 1-2 ___ 3-4 __ 5 or More
4. Within the past month, how many times have you visited a web site in which clothing is offered for sale?
   ___ 0 ___ 1-2 ___ 3-4 __ 5 or More

For questions 5 through 7, please circle the number that reflects your opinion

5. Shopping online for clothing will become increasingly popular.

Strongly Agree Agree Neutral Disagree Strongly Disagree
|_____|_____|_____|_____|
5 4 3 2 1

6. Shopping online for clothing is a wise action for today's consumers.

Strongly Agree Agree Neutral Disagree Strongly Disagree
|_____|_____|_____|_____|
5 4 3 2 1

7. Within the next 30 days, how likely are you to purchase clothing online?

Very likely Somewhat likely Not sure Somewhat unlikely Very unlikely
|_____|_____|_____|_____|
5 4 3 2 1

8. What do you believe to be the main reason why people shop online for clothes?

___ Convenience ___ Low Prices ___ More Selection ___Others (Please Specify): _____

9. What do you believe to be the main reason why people AVOID shopping online for clothes?

___ Credit Card Security ___ Items May Not Fit ___ Difficult to Return ___Others (Please Specify): _____

Appendix II: Example of a Questionnaire with a Codebook

In this questionnaire the goal is to find consumers preferences for pizzas in a pizzeria. The numbers behind the answers gives to options for the coding.

Variables	Questions
1	1 Have you purchased a Pizzazza pizza in the last month? _X_ Yes (1) ___ No (2) ___ Unsure (3)
2	2 The last time you bought a Pizzazza pizza did you (cross only one): ___ Have it delivered to your house? (1) ___ Have it delivered to your place of work? (2) ___ Pick it up yourself? (3) _X_ Eat it at the Pizzazza pizza restaurant (4) ___ Purchase it some other way? (5)
3	3 In your opinion, the taste of a Pizzazza pizza is (tick only one): ___ Poor (1) ___ Fair (2) _X_ Good (3) ___ Excellent (4)
	4 Which of the following toppings do you typically have on your pizza? (Cross all that apply.)
4	___ Green pepper (0;1)
5	___ Onion (0;1)
6	_X_ Mushroom (0;1)
7	___ Sausage (0;1)
8	_X_ Pepperoni (0;1)
9	___ Hot peppers (0;1)
10	_X_ Black olives (0;1)
11	___ Anchovies (0;1)
12	___ Pineapple (0;1)
13	___ Shrimps (0;1)
14	5 How do you rate the speediness of Pizzazzas restaurant servi ce once you have ordered? (Circle the appropriate number if a 1 means very slow and a 7 means very fast.) Very slow 1 2 3 4 5 (6) 7 very fast
15	6 Please indicate your age: ___ 0–15 years (1) ___ 15–25 years (2) _X_ 26–40 years (3) ___ 41–60 years (4) ___ Over 60 years (5)
16	7 Please indicate your gender: _X_ Male (1) ___ Female (2)
17	8 Please indicate which country you are from (cross only one): ___ US (1) ___ Canada (2) _X_ UK (3) ___ Germany (4) ___ France (5) ___ Italy (6) ___ Russia (7) ___ China (8) ___ Other country (9)

Note: the 0;1 indicates the coding system that will be used. Each response category must be defined as a separate question = variable, 0=No, 1=Yes

The accompanying code sheet could look as:

Variables

Respondent number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1																	
2																	
3																	
.																	
.																	
.																	
10	1	4	3	0	0	1	0	1	0	1	0	0	0	6	3	1	3
.																	
.																	
.																	
n																	

Nils Karl Sørensen ©
Advanced Tools Statistics

Set 6: Two-way ANOVA

by Nils Karl Sørensen

Outline

page

Literature

Bruce L. Bowerman, Richard T. O'Connell, J.B. Orris and Emily S. Murphree, 2010, *Essentials of Business Statistic.* (**3rd edition**). McGraw-Hill. Section 11.3 and Appendix D. The latter is available from the homepage of the book or in Blackboard (course documents and then folder *literature*).

Alternative:

Bruce L. Bowerman, Richard T. O'Connell, J.B. Orris and Emily S. Murphree, 2012, *Essentials of Business Statistic.* (**4th edition**). McGraw-Hill, Chapter 11, Sections 11.3 and 11.4.

Look in Blackboard under *course materials* folder *literature* to find the material.

1. Introduction

Back in Bowerman Chapter 11, Sections 11.1 and 11.2, we worked with the Analysis of Variance (ANOVA). This method was a generalization of the test for comparing two independent samples for mean to *p* groups. We examined a hypothesis stating that the mean between the groups were equal. We considered the mean of the variable in the columns only. In some cases also the mean of the *rows* may be of importance.

In such cases, we label the measurement variable of the column *factor 1* and the measurement variable of the row *factor 2* respectively. The factor of the row may be a single row or a group of rows. The first case is called a *Randomized Block Design*, whereas the latter is called *Two-way ANOVA*.

Related to questionnaires, the more advanced approach to the ANOVA analysis will be of interest if we consider cases where several questions are related to each other.

2. Randomized Block Design

(BO Section 11.3)

A randomized block design compares *p* treatments (columns) with *b* blocks (rows). Each block is used exactly one time to measure the effect of each and every treatment. The assumptions are similar to the ones used for the one-way ANOVA. Note, that for each treatment, the number of blocks has to be similar.

Hypotheses

Can be stated as:

H_0: The means of *p* and *b* are equal
H_1: Minimum one is different

This can be elaborated further, but now things become more complicated:

H_0: The means are equal	i) For the *p* treatments
	ii) For the *b* blocks
H_1: Minimum one is different	i) For the *p* treatments
	ii) For the *b* blocks

Method

Compared to the one-way ANOVA, the variation has to be decomposed into treatments as well as blocks.

Total variation	= Treatment	+ Block	+ Errors
Sum Square Total	= Sum Square Treatment	+ Block Sum Square	+ Sum Square Error

Or:

$$SSTO = SST + SSB + SSE$$

The calculations can by summarized into the extended *ANOVA-table:*

Variation	*Squared sum (SS)*	*Degrees of freedom (df)*	*Mean square (MS)*	*F-value*
Treatment	$SST = b\sum_{i=1}^{p}(\bar{x}_{i\bullet} - \bar{x})^2$	p – 1	MST = SST/(p−1)	F = MST / MSE
Blocks	$SSB = p\sum_{i=1}^{b}(\bar{x}_{\bullet j} - \bar{x})^2$	b – 1	MSB = SSB/(b−1)	F = MSB / MSE
Error	$SSE = SSTO - SST - SSB$	(p−1)(b−1)	$MSE = \frac{SSE}{(p-1)(b-1)}$	
Total	$SSTO = \sum_{i=1}^{p}\sum_{j=1}^{b}(x_{ij} - \bar{x})^2$	p(b – 1)		

There are now two F-tests to be considered. One for the *p* treatments and one for the *b* blocks.

Degrees of freedom for the testers are:

Treatments: $df_1 = (p-1)$ $df_2 = (p-1)(b-1)$

Blocks: $df_1 = (b-1)$ $df_2 = (p-1)(b-1)$

Example in Excel and Megastat

Let us consider a little case, where the price of a basket of daily commodities is purchased in four different cities. In each city, the basket is purchased in five different supermarkets. The dataset measured in DKK look as:

	Sønderborg	Aabenraa	Kolding	Ribe
Netto	750	800	810	680
Fakta	780	790	790	740
Føtex	820	830	840	750
Lidl	790	770	730	730
Aldi	740	770	770	750

In order to substantiate our analysis, let us first perform an ANOVA-analysis. The hypotheses to be inspected are:

H_0: The means of the cities are equal ie. $\mu_{Sønderborg} = \mu_{Aabenraa} = \mu_{Kolding} = \mu_{Ribe}$
H_1: Minimum one is different

We perform along the lines described in the notes from the statistics course. In Excel perform ***data/data analysis/One-way ANOVA.*** The following output will appear:

ANOVA: Single Factor

Summary

Groups	*Obs*	*Sum*	*Average*	*Variance*
Sønderborg	5	3,880	776	1030
Aabenraa	5	3,960	792	620
Kolding	5	3,940	788	1720
Ribe	5	3,650	730	850

ANOVA

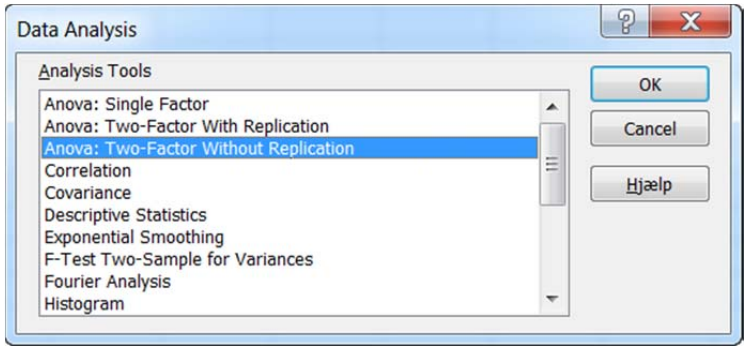
Source	*SS*	*Df*	*MS*	*F-value*	*P-value*	*F-crit*
Between Groups	12,175	3	4,058.33	3.85	0.03	3.24
Within Groups	16,880	16	1,055.00			
Total	29,055	19				

The test is significant at the 5 percent level. A supplementary analysis will reveal that Ribe is cheaper than the other three cities.

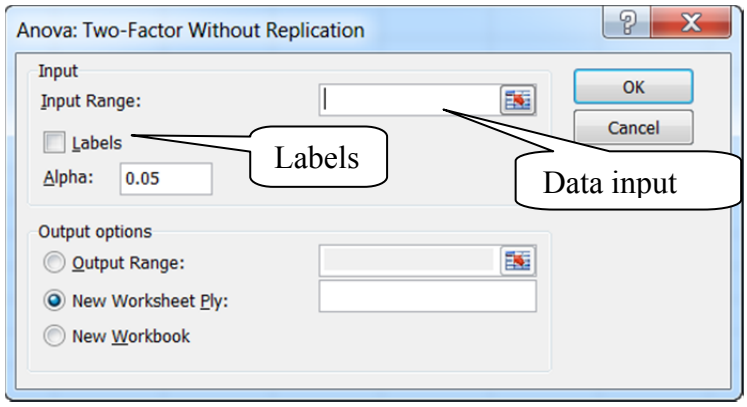
Now expand the problem, and consider the supermarket chain as well. In this case the hypotheses are:

H_0: The mean of cities as well as supermarket chains are equal
H_1: The means are different

In Excel select ***data / data analyses / ANOVA: Two-Factor without Replication***



A box is obtained looking as:



Performing this sequence will result is the following output:

ANOVA: Two-factor without Replication

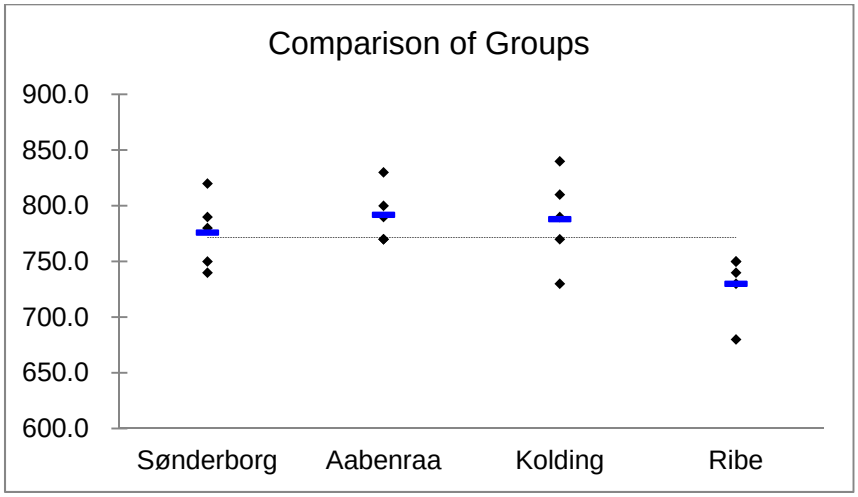
Summary	*Obs*	*Sum*	*Average*	*Variance*
Netto	4	3,040	760	3533
Fakta	4	3,100	775	567
Føtex	4	3,240	810	1667
Lidl	4	3,020	755	900
Aldi	4	3,030	758	225
Sønderborg	5	3,880	776	1030
Aabenraa	5	3,960	792	620
Kolding	5	3,940	788	1720
Ribe	5	3,650	730	850

ANOVA

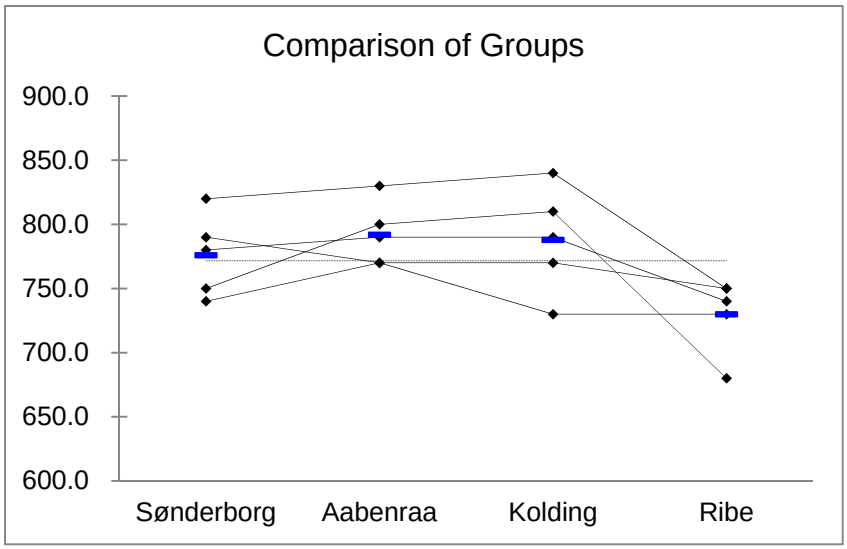
Source	*SS*	*Df*	*MS*	*F-value*	*P-value*	*F-crit*
Rows (supermrk.)	8,380	4	2,095.00	2.96	0.06	3.26
Columns (cities)	12,175	3	4,058.33	5.73	0.01	3.49
Error	8,500	12	708.33			
Total	29,055	19				

It is observed that the p-value for the cities has decreased from 0.03 to 0.01. This is for cities. It is also lower, than for rows (supermarkets). The p-value for the supermarkets is not significant at the 5 percent level, but at the 10 percent level only. Although, Føtex is the most expensive, it is not significantly more expensive than the cheapest supermarket namely Lidl. So the price span among the supermarkets remains constant, but level differs among cities.

In ***Megastat,*** a similar analysis can be conducted. Her post-hoc or supplementary analysis is performed as well. In Megastat use ***add-ins / Megastat / Analysis of Variance /Randomized Block Design***. The dialog box looks as under one-way ANOVA. However, in Megastat some interesting plots can be produced see below.



and



3. Two−Way ANOVA

(Appendix D or BO Section 11.4, 4th edition)

The two-way ANOVA is a further extension of the randomized block design. The treatments and blocks are now assumed to *interact* with each other. In our example this implies that the price setting of the Supermarkets not only follow an overall company policy, but also varies from city to city.

Two factors are said to interact if the difference between levels (treatments) of one factor depends on the level of the other factor. Factors that do not interact are called additive.

The three questions answerable by two-way ANOVA are:

- Is there any factor A main effect (treatments)?
- Is there any factor B main effect (blocks)?
- Are there any interaction effects of factors A and B?

Moving back to the example of the Supermarkets we add also information with regard to the supermarket chain "Coop". The table looks now as:

		Sønderborg	Aabenraa	Kolding	Ribe
Seg 1	Netto	750	800	810	680
	Fakta	780	790	790	740
Seg 2	Føtex	820	830	840	750
	Coop	840	850	860	820
Seg 3	Lidl	790	770	730	730
	Aldi	740	770	770	750

There are now a levels of factor A ($a = 3$), i.e. the segments, and there are b levels of factor B ($b = 4$) i.e. the cities. Thus, there are $a \times b$ ($3 \times 4 = 12$) combinations of segments and cities. Finally, there are $n = 2$ elements/supermarkets in each segment.

In the table, the elements of factor B rows have been further decomposed into 3 segments. So the grouping of factor B is curial for the outcome of the investigation. The 3 segments represents: 1) Danish owned discount supermarkets; 2) Normal Danish supermarkets, and 3) German owned discount supermarkets. Besides from an investigation of the price setting across the cities (factor A), we can analyze the price setting across supermarket chains (factor B), and finally the price setting of each segment across the cities (interaction among factor A and factor B).

In the previous Section it was assumed that the numbers of the blocks are of equal size among all treatments. In addition, it is in the present case assumed that the sizes of the segments are equal. Therefore, we had to add "Coop" otherwise segment 2 would only consist of "Føtex".

Hypotheses
Can be stated as:

Factor A:	H_0: All *a* factors are equal
(treatments)	H_1: Minimum one is different
Factor B:	H_0: All *b* factors are equal
(blocks)	H_1: Minimum one is different
Factor AB:	H_0: All *ab* factors are equal
(interaction)	H_1: Minimum one is different

Method
The decomposition of the total variation is further extended relative to the randomized block design. The mathematical formulas are not handy, so for the present purpose we only state:

Total variation = Variation A + Variation B + Variation AB +Variation Error

SST = SSA + SSB + SS(AB) + SSE

The calculations can by summarized into the extended *Two-way ANOVA-table:*

Variation	*Squared Sum*	*Degrees of freedom (df)*	*Mean square (MS)*	*F-value*
Factor A	SSA	a – 1	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSE}$
Factor B	SSB	b – 1	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSE}$
Interaction	SS(AB)	(a–1)(b−1)	$MSAB = \frac{SS(AB)}{(p-1)(b-1)}$	$F_{AB} = \frac{MSAB}{MSE}$
Error	SSE	ab(n–1)	$MSE = \frac{SSE}{ab(n-1)}$	
Total	SST	abn – 1		

There are now three F-tests to be considered. One for each factor and the F-test for the interaction among the two factors A and B. Degrees of freedom for the testers are:

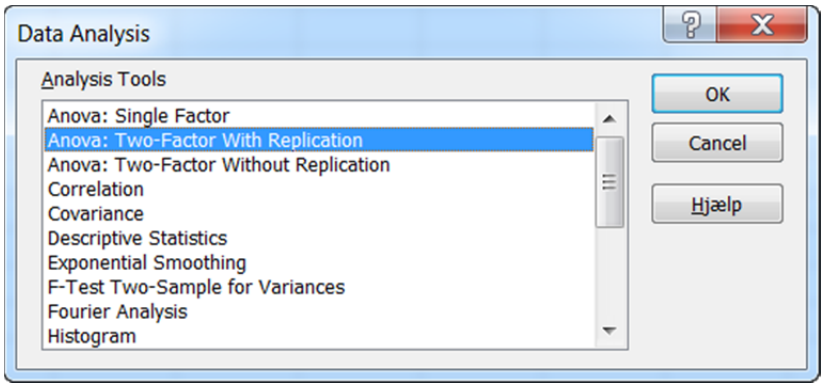
Factor A:	$df_1 = (a-1)$	$df_2 = ab(n-1)$
Factor B:	$df_1 = (b-1)$	$df_2 = ab(n-1)$
Interaction	$df_1 = (a-1)(b-1)$	$df_2 = ab(n-1)$

Example in Excel and Megastat

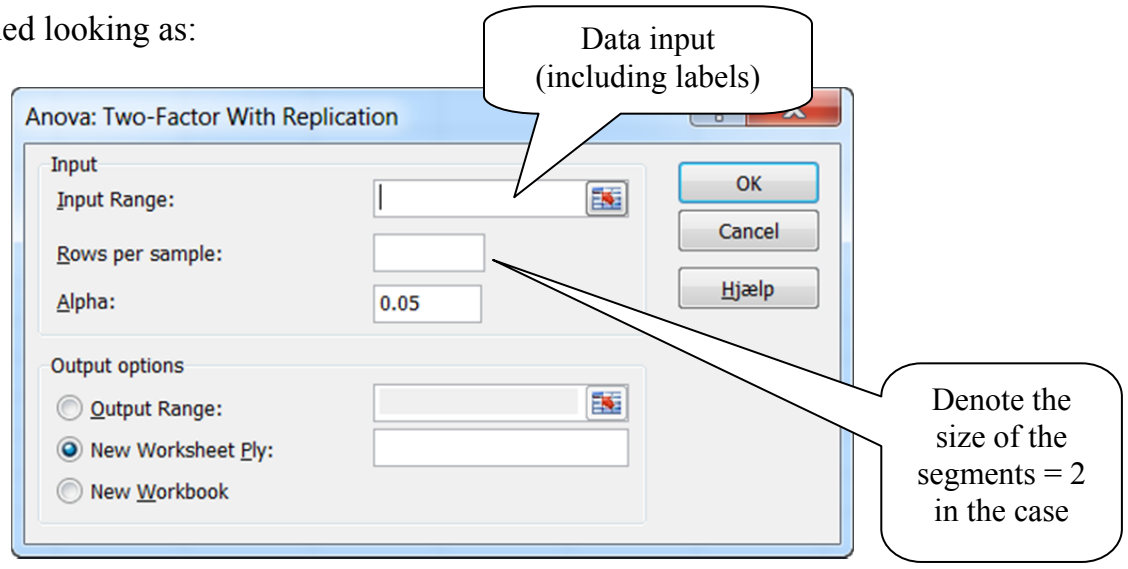
First the hypotheses have to be stated. This is a little more complicated than earlier.

Factor A:	H_0: The price basket among the segments is similar
(treatments)	H_1: Minimum one is different
Factor B:	H_0: The price basket among the cities is similar
(blocks)	H_1: Minimum one is different
Factor AB:	H_0: The segments of the Supermarkets and the cities are similar
(interaction)	H_1: Minimum one is different

In Excel select ***data / data analyses / ANOVA: Two-Factor with Replication***



A box is obtained looking as:



Performing this sequence will result is the following output:

ANOVA: Two-way with Replication

Summary	Sønderborg	Aabenraa	Kolding	Ribe	Total
Netto/Fakta					
Obs	2	2	2	2	8
Sum	1,530	1,590	1,600	1,420	6,140.00
Average	765	795	800	710	767.50
Variance	450	50	200	1800	1821.43
Føtex/Coop					
Obs	2	2	2	2	8
Sum	1,660	1,680	1,700	1,570	6,610
Average	830	840	850	785	826.25
Variance	200	200	200	2450	1141.071
Lidl/Aldi					
Obs	2	2	2	2	8
Sum	1,530	1,540	1,500	1,480	6,050
Average	765	770	750	740	756.25
Variance	1,250	0	800	200	483.93
Total					
Obs	6	6	6	6	
Sum	4,720	4,810	4,800	4,470	
Average	786.67	801.67	800	745	
Variance	1,506.67	1,056.67	2,240	2,030	

Two-way ANOVA

Source	*SS*	*df*	*MS*	*F*	*F-value*	*F crit*
Factor A (seg./rows)	22,608.33	2	11,304.17	17.39	0.00	3.89
Factor B (cities/col.)	12,566.67	3	4,188.89	6.44	0.01	3.49
Interaction	3,758.33	6	626.39	0.96	0.49	3.00
Error	7,800.00	12	650			
I alt	46,733.33	23				

Initially, a descriptive analysis of each segment by city is provided. This is some kind of "averaging" of the analysis found under the analysis of randomized block design. The outcome from the ANOVA-table has the following interpretation: Factor A is strongly significant, so the segments are observed. Further, there are in general price differences among the cities. However, this is not due to the presence of the segments. The segments prevail among the cities and do not lead to a change in competition. Or stated alternatively: If for example "Coop" and "Føtex" is the most expensive place to buy the basket of consumer goods, then this is the case in all cities.

Moving to Megastat, the post-hoc or supplementary analysis is performed as well. This is a Tukey comparison as described under the simple ANOVA-analysis; see Chapter 11 in Bowerman Section 11.2. In Megastat use ***add-ins / Megastat / Analysis of Variance /Two-way ANOVA***. The dialog box looks as under one-way ANOVA. However, in Megastat some interesting plots can be produced see below.

Two factor ANOVA

Factor 2

Means:

Factor 1	Sønderborg	Aabenraa	Kolding	Ribe	
Netto/Fakta	765.0	795.0	800.0	710.0	767.5
Føtex/Coop	830.0	840.0	850.0	785.0	826.3
Lidl/Aldi	765.0	770.0	750.0	740.0	756.3
	786.7	801.7	800.0	745.0	783.3

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Factor 1	22,608.33	2	11,304.167	17.39	.0003
Factor 2	12,566.67	3	4,188.889	6.44	.0076
Interaction	3,758.33	6	626.389	0.96	.4884
Error	7,800.00	12	650.000		
Total	46,733.33	23			

Post hoc analysis for Factor 1

Tukey simultaneous comparison t-values (d.f. = 12)

		Lidl	Netto	Føtex
		756.3	767.5	826.3
Lidl	756.3			
Netto	767.5	0.88		
Føtex	826.3	5.49	4.61	

critical values for experiment wise error rate:

0.05	2.67
0.01	3.56

p-values for pair wise t-tests

		Lidl	Netto	Føtex
		756.3	767.5	826.3
Lidl	756.3			
Netto	767.5	.3948		
Føtex	826.3	.0001	.0006	

Post hoc analysis for Factor 2

Tukey simultaneous comparison t-values (d.f. = 12)

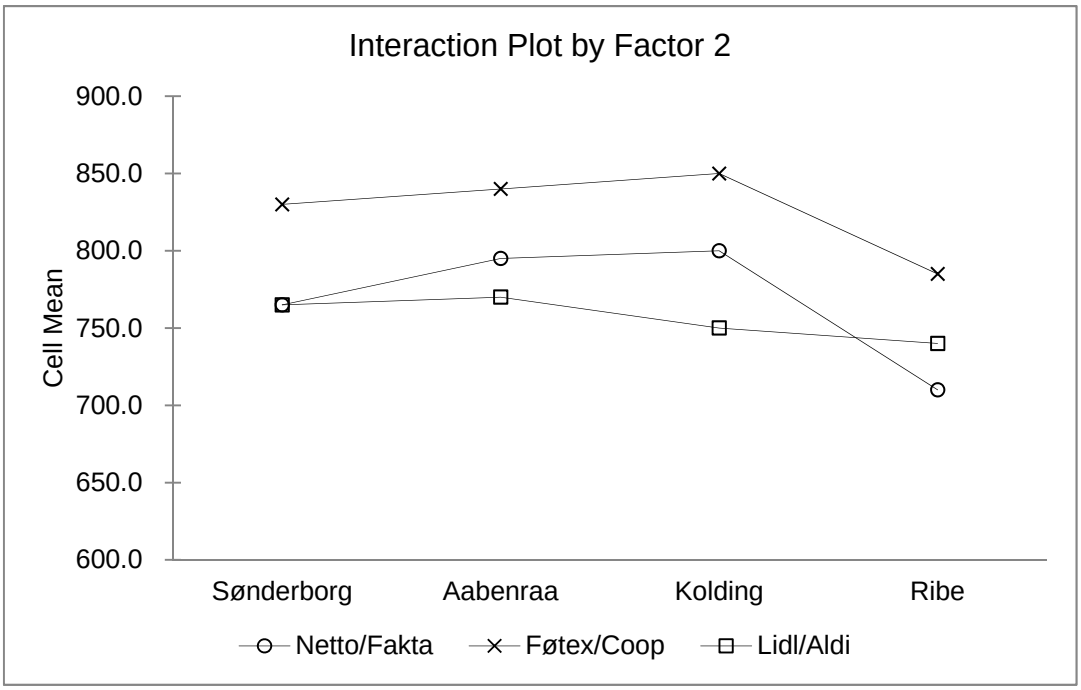
		Ribe	Sønderborg	Kolding	Aabenraa
		745.0	786.7	800.0	801.7
Ribe	745.0				
Sønderborg	786.7	2.83			
Kolding	800.0	3.74	0.91		
Aabenraa	801.7	3.85	1.02	0.11	

critical values for experiment wise error rate:

0.05	2.97
0.01	3.89

p-values for pair wise t-tests

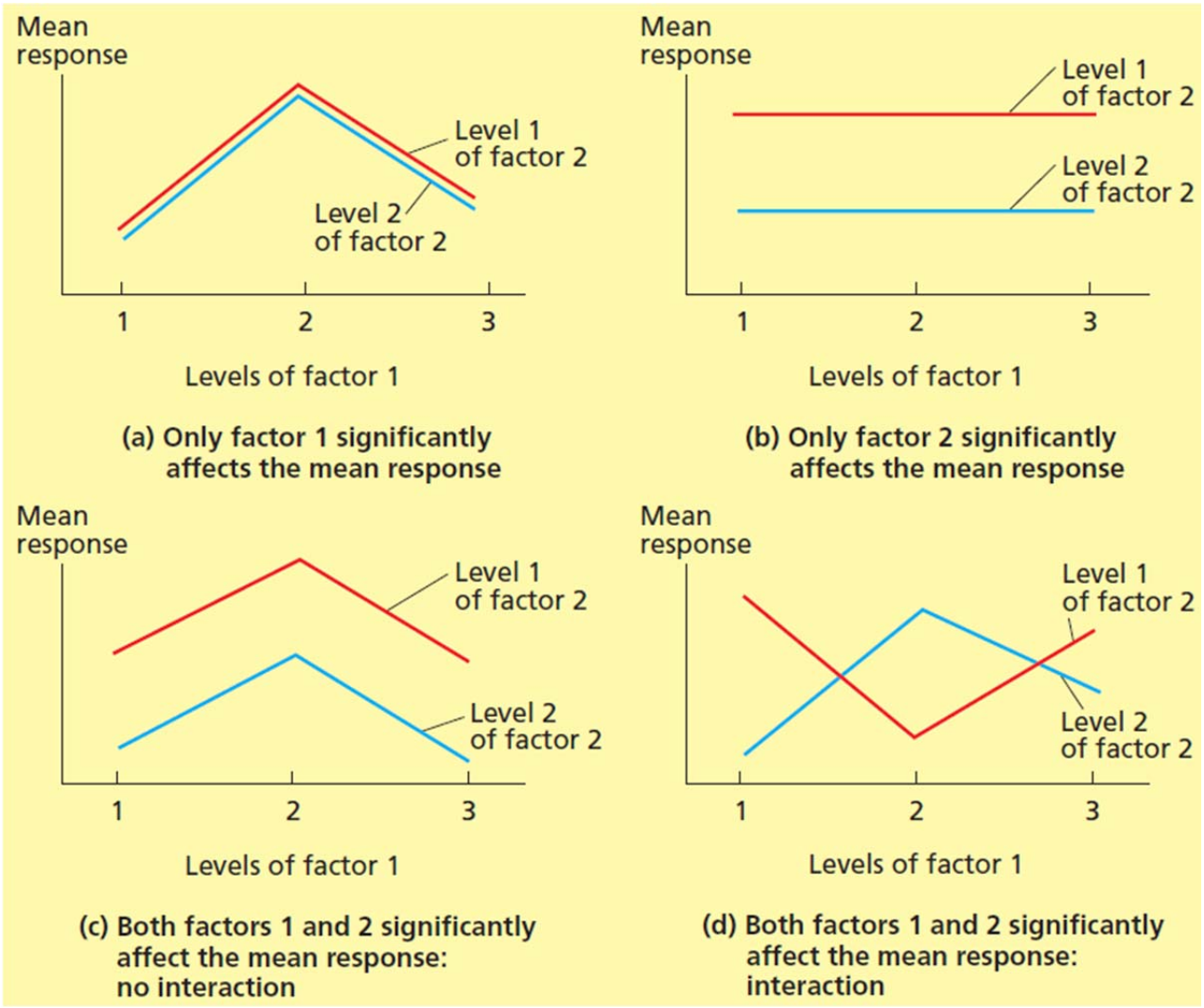
		Ribe	Sønderborg	Kolding	Aabenraa
		745.0	786.7	800.0	801.7
Ribe	745.0				
Sønderborg	786.7	.0152			
Kolding	800.0	.0028	.3829		
Aabenraa	801.7	.0023	.3283	.9117	



What is the interpretation of the graphic illustration? In Appendix D in Bowerman, a guide is provided in order to read the graphs. This guide is shown on the top of the next page. Four cases are considered. In the present case the lower left panel seems to be the most appropriate. In the cases in the upper part of the illustration the use of two-way ANOVA has no effect because the variables not are related. In the lower panel the case to the left shows a

situation where both factors have influence, but there is no significant interaction. This is the case in the final illustration shown in the bottom panel to the right. Here the interesting issue is that the lines are crossing, but still displays a systematic pattern. Observe that in the illustration above in Ribe, the segments of Lidl/Aldi and Netto/Fakta crosses. This implies that the price basket of the two Danish discount markets is cheaper than the similar basket of goods supplied by Lidl/Aldi. In this case an interaction among the two segments with the feature that discount is still the cheapest way to buy consumer goods. In interaction is only observed for Ribe and is therefore not overall significant. That would properly have been the case, if this feature also had been observed in one of the three other cities.

Different Possible Treatment Effects in Two-Way ANOVA



Nils Karl Sørensen ©
Advanced Tools Statistics

Set 7: The use of SPSS and Logistic Regression

by Nils Karl Sørensen

Outline

page

1. Working with SPSS

SPSS is a widely used software package. It covers all types of statistical analyses. It is specially designed for analyses of questionnaires. Time series analysis is not optimal covered by the package. SPSS has a long tradition, and it has been a part of statistical analysis since the days of the mainframe systems from the 1960ties and later.

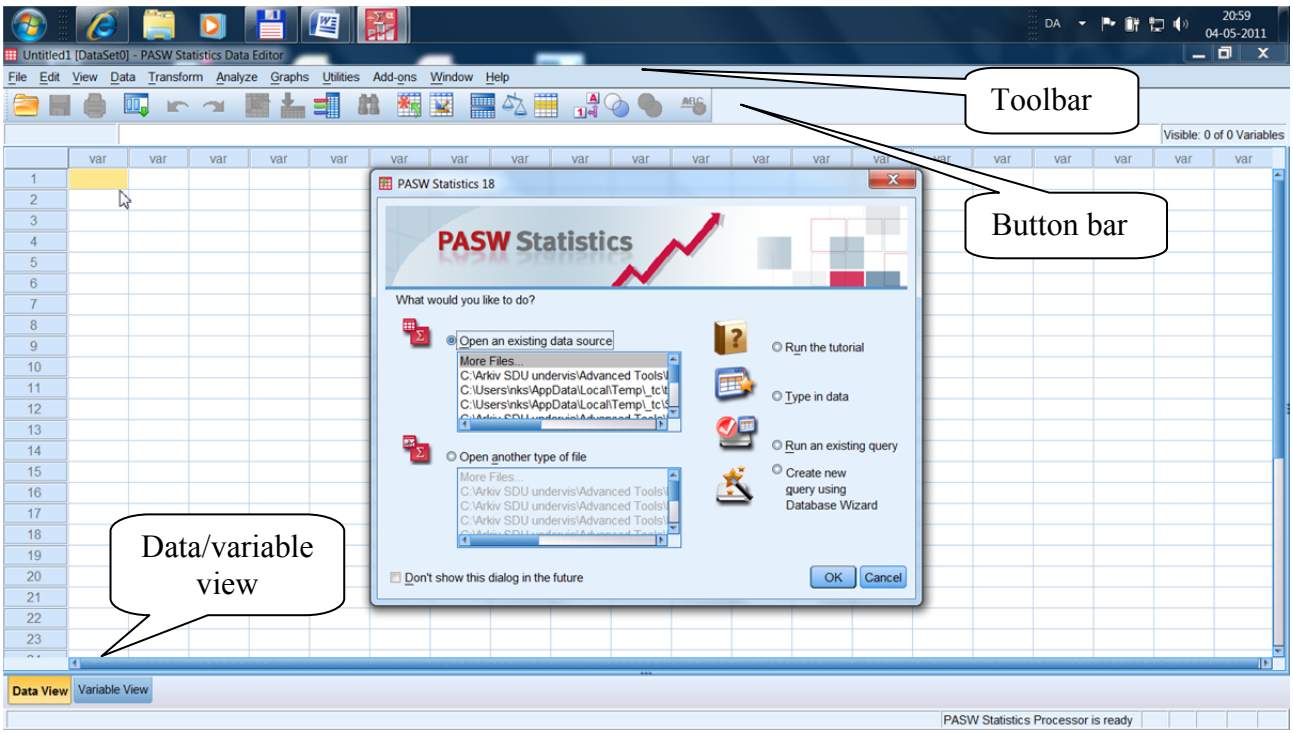
Nearly all books on marketing research uses and advocates SPSS. Interestingly, no books on statistics use SPSS for applications, examples etc. This makes teaching on SPSS a little bit special. SPSS has a very good distribution system, but the license is expensive for business users just as the case with SAS. SDU supports SPSS. Over the past two decades SPSS has faced increased competition and several alternatives has occurred. SPSS has met this challenge by introduction a new version nearly every year.

In recent versions a system of add-ins has been adopted. This has caused the price to decrease for the basic version, but here SPSS is not much better than the AnalysisToolPack by Excel/Microsoft. However, SDU has most of the add-ins, and they are being installed automatically when downloading the program from Backboard. SPSS run on Windows as well an on MAC. The license runs for a year, and has to be renewed every year at the end of June. Having the add-ins installed SPSS has many features. This is an advantage as well as a weakness of the package, because there are so many options, that the user easily gets confused.

A freeware alternative to SPSS is PSPP. Just provide a search on “PSPP” and follow the instructions! PSPP also supports Linux etc. A second alternative is **Winstat.** This is an add-in to Excel. This program is very competitive price set relative to SPSS, and with the analysis tool package loaded it is nearly as good.

The following is a very brief introduction to the package. SPSS has a very efficient help function that can answer nearly all questions. First get SPSS installed from the SDU system. Select if you are Windows or Mac user.

Having SPSS successfully installed click on the SPSS icon and obtain the start screen:



SPSS offers two possibilities for opening data namely in the SPSS data format *.sav or other formats like for example Excel *.xlsx.

On the screen shot observe also:

- The *data view / variable view.* This is a very handy feature. It allows you to type in data with very long names. For example a full question in a questionnaire. This is a good feature if you have for example 100 questions, and want to look on correlations or goodness-of-fit tests (notice, that the way data is organized is a little bit different and more efficient than the method applied by Excel).
- The *toolbar* has all relevant main menus. Three of them are worth mentioning namely: ***data*** for data handling; ***analyze*** with all relevant menus for statistical analyses, and finally the ***help*** function to the right.
- The *button bar* covers main function on opening of data, zoom and pivot functions etc.

The best way to learn the package is to load in a data set, and then flip around in the menus! So good luck!

2. Logistic Regression

(BO Section 14.12 or Appendix E in Bowerman 4e or older)

This term *logistic regression* often is confusing, and in fact also in SPSS, where several menus claim that they can perform the same thing. In the present context we refer to what is known as the *logistic regression model.* A binary logistic regression model implies that that the dependent variable has only two categories for example 0 = success and 1 = failure just as the case with a dummy variable.

Such a general linear probability model with *k* "regressors" labeled *x* can be written as

$$y = E(y) + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

But now compared to the "traditional" multiple regression model *y* is defined as:

$$y = \begin{cases} 1 \text{ if initial option is choosen } (\textit{buy a car}, \textit{vote yes}, \textit{drink "Carlsberg"}) \\ 0 \text{ if alternative option is choosen } (\textit{not buy a car}, \textit{vote no}, \textit{drink "Tuborg"}) \end{cases}$$

This model is complex to estimate by OLS. Only if *y* is divided into two equal large parts will the estimates be consistent. Otherwise the coefficients will be either upward or downward biased. The problem is that we have a non−linear model, and therefore we also need a non−linear estimator. This is a complex expression that is maximized (a little like the optimum method known from the course in mathematics). This method is called a log likelihood estimator.

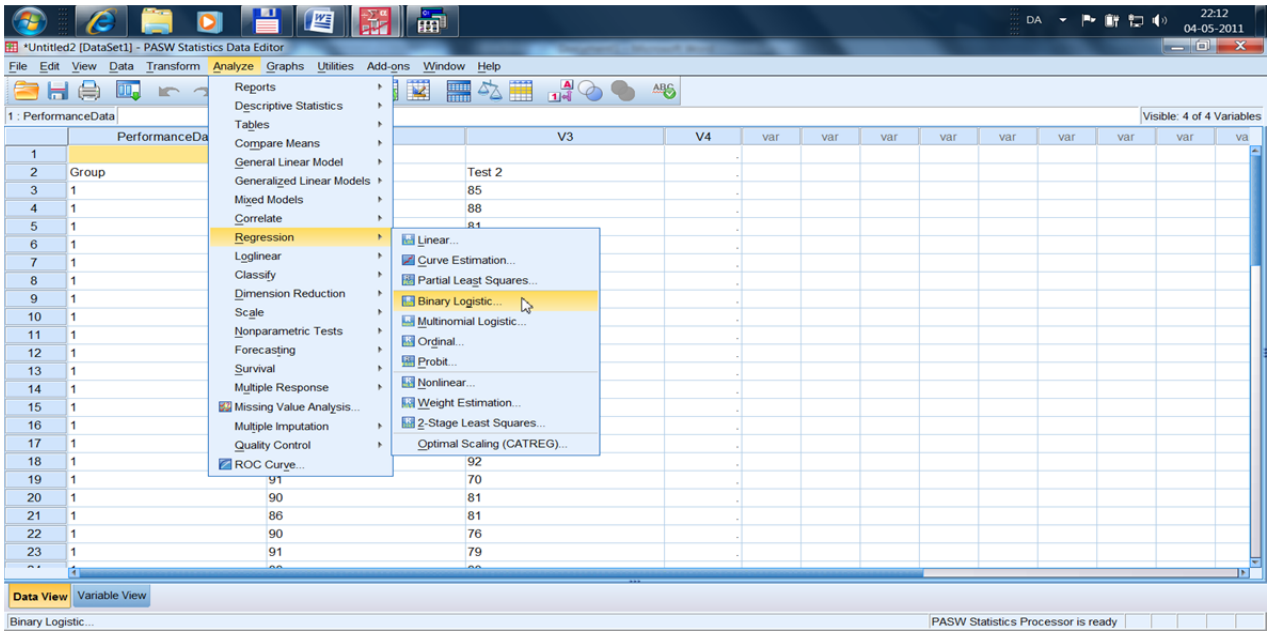
The interpretation of this expression is similar to the well known regression output. We illustrate with the example taken from Bowerman. A personal director of a firm has developed two tests to help determine whether potential employees would perform successfully in a particular position. To help estimate the usefulness of the tests, the director gives both tests to 43 employees that currently hold the position. If the employee is performing successfully we associate the value 1 for the *y*-variable and 0 if the employee is performing unsuccessfully. We label this variable *Group.* The data set has now the following set up:

Observations	*Group*	*Test 1*	*Test 2*
1	1	96	85
2	1	96	88
3	1	91	81
...	...	...	...
...	...	...	...
...	...	...	...
42	0	83	77
43	0	81	71

We want to estimate the following model by use of SPSS:

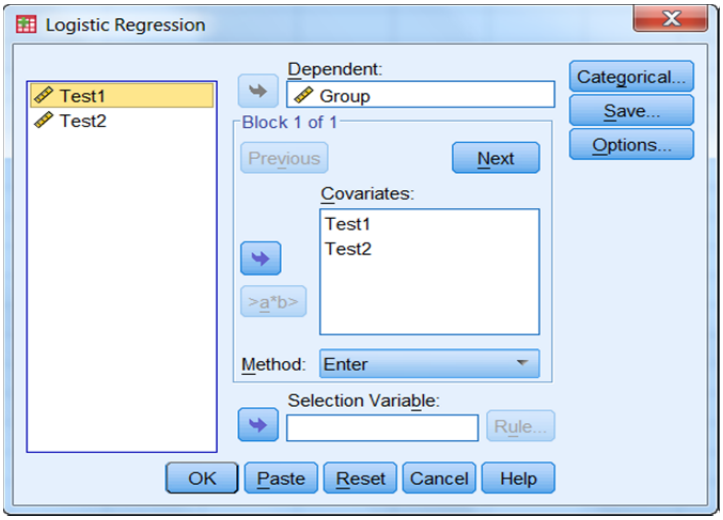
$$Group_i = E(Group) + \varepsilon_i = \beta_0 + \beta_1 Test1_i + \beta_2 Test2_i + \varepsilon_i$$

We open SPSS and load in the data from the Excel file labeled *Perftest.xls* from the data directory accompanying Bowerman. In SPSS select on the tool bar "Analyze" then "Regression" and then "Binary Logistic...".



In order to import your data from an Excel file into the data format written by SPSS, you have to manipulate a little with the data loaded in from Excel. Delete lines 1 and 2, and move into *variable view* (bottom left). Here give the variable the names in the equation above.

Under "Binary logistic..." the following menu appears:



As the dependent variable use *Group,* and as covariates (regressors) use the variables *Test1* and *Test2* respectively. Click then on "OK". A lot of things happen, but the final output looks very similar to Figure E.2 in Bowerman:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Test1	,481	,158	9,319	1	,002	1,618
	Test2	,165	,102	2,622	1	,105	1,179
	Constant	-55,981	17,434	10,311	1	,001	,000

a. Variable(s) entered on step 1: Test1, Test2.

What is interpretation? The "B"s are the coefficients. They are both positive. So a good test score means that you belong to the best group (1). What a surprise! The coefficient of "Test1" is the highest, so this is the most important test. The level of significance or the *p-value* is shown under the column "Sig". For "Test1" it is equal to 0.002. This is highly significant. For "Test2" it is equal to 0.105. This is higher than 0.05, so this variable is not significant. The conclusion is that only the first test is relevant for the overall performance of the employee.

How good is the model? In the SPSS output we find:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	27,886[a]	,519	,694

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Depending on the measure used the model explains between 0.519 and 0.694 of the variation in data. In logistic regression these values are frequently low, so this is pretty good.

What is the difference relative to the ordinary regression performed by use of Excel? In order to examine this issue consult the output by Excel below:

Regression Statistics	
Multiple R	0.73
R-squared	0.54
Adjusted R-square	0.51
Standard error	0.35
Observations	43

ANOVA

	df	*SS*	*MS*	*F-value*	*Signif*
Regression	2	5.75	2.87	23.22	0.00
Residual	40	4.95	0.12		
Sum	42	10.70			

	Coeffi-cient	*Standard error*	*t-stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-5.9291	0.9633	-6.1547	0.00	-7.8760	-3.9821
Test 1	0.0586	0.0112	5.2330	0.00	0.0360	0.0812
Test 2	0.0153	0.0100	1.5393	0.13	-0.0048	0.0354

Comparison reveals that the significance and signs of the coefficients is quite similar, but the size of the coefficients is very different. The model performed by the traditional OLS method is therefore biased and insufficient.