A Compendium of Reproducible Research about Descriptive Statistics and Linear Regression

Patrick Wessa

February 28, 2008

Abstract

This document can be used as an illustrated tutorial about Wessa.net and FreeStatistics.org. It illustrates how Reproducible Research can be integrated in Statistics Education based on a Compendium of Reproducible and Reusable Research. [1]

1 Introduction

This case study illustrates how free online statistical software (R modules) can be used in data analysis. This tutorial can be used at an undergraduate level - the prerequisites are as follows:

- a basic understanding of Descriptive Statistics
- an interest in scientific data analysis
- basic ICT skills

All statistical techniques are based on the online "e-Handbook of Statistical Methods" [16] that is available online as HTML and PDF documents. The structure of the e-Handbook is like a cookbook: it contains a list of statistical methods in alphabetical order. There is also an overview of all methods that is ordered by type of analysis. In my experience with students it is best to use the electronic PDF version of the e-Handbook because it can be easily browsed and searched. Each method is explained in simple words and with a minimum of mathematical background. In addition, the documents contains numerous illustrations and examples which may also be beneficial.

The case-based approach of this tutorial and the technology that allows to reproduce/reuse every aspect of the research, is consistent with the pedagogical paradigm of (social) constructivism. The fact that this tutorial has been designed as a compendium makes every aspect of the portayed research reproducible which empowers anyone with a healthy interest in statistics to interact (and conduct experiments) with the underlying statistical science.

From the statistical point of view, the following concepts are used in this tutorial:

[1] Acknowledgements: The Compendium Platform for Reproducible Research was partially funded by the OOF 2007/13 grant of the K.U.Leuven Association.

- Measures of Central Tendency (arithmetic mean and median)
- Explorative Data Analysis (such as Notched Boxplots)
- The importance of outliers
- Statistical Hypothesis Testing (one-sided and two-sided)
- Simple and Multiple Regression Analysis

A final word of caution: The models that are presented here are illustrative and simplistic in nature. Nevertheless the document should contain some interesting and challenging concepts for students at an undergraduate level.

2 How much should we pay for coffee?

2.1 Problem

We want to import Arabica coffee from Colombia and sell it in the USA. Therefore we would like to explore and estimate the relationship between the monthly time series Y_t (the US retail price in US cents per lb) and X_t (the price paid to growers in Colombia in US cents per lb). We suspect that the increase of prices paid to growers (on the long run) is substantially lower than the rise in retail prices in the USA.

2.1.1 Hypotheses

Hypothesis 1 Arabica coffee prices paid to growers in Colombia are constant. We define $H_0 : \alpha_X = c$ and $H_A : \alpha_X \neq c$ where c is any "positive constant number" for the following equation: $X_t = \alpha_X + e_t$ for $t = 1, 2, ...T$. Note: if the coffee prices are constant this implies that the constant α_X can be used to make predictions about X_t on the long run.

Hypothesis 2 Retail prices of Arabica coffee in the USA are constant. We define $H_0 : \alpha_Y = c$ and $H_A : \alpha_Y \neq c$ where c is any "positive constant number" for the following equation: $Y_t = \alpha_Y + e_t$ for $t = 1, 2, ...T$. Note: if the coffee prices are constant this implies that the constant α_Y can be used to make predictions about Y_t on the long run.

Hypothesis 3 The retail prices of Arabica coffee (in the USA) can be (partially) explained by the price paid to growers in Colombia. We define $H_0 : \beta = 0$ and $H_A : \beta > 0$ for the following equation: $Y_t = \alpha + \beta X_t + ... + e_t$ for $t = 1, 2, ...T$. Note: we use a one-sided hypothesis test because we expect β to be positive.

Hypothesis 4 The long run growth of retail prices of Arabica coffee (in the USA) cannot be explained by the prices paid to growers in Colombia. In other words, an additional trend variable is needed to account for the long run increase in US retail prices. We define $H_0 : \gamma = 0$ and $H_A : \gamma > 0$ for the following equation: $Y_t = \alpha + \beta X_t + \gamma t + ... + e_t$ for $t = 1, 2, ...T$. Note: we use a one-sided hypothesis test because we expect γ to be positive.

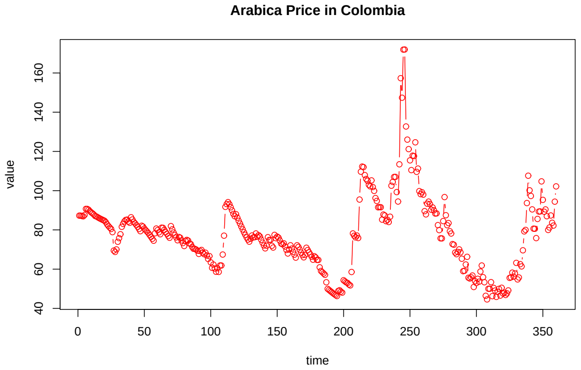


Figure 1: Run Sequence Plot - Arabica Coffee Price in Colombia
link of reproducible computation

2.2 Data

The following time series were obtained from the International Coffee Organization (ICO):

- The univariate time series of prices paid to growers in Colombia in US cents/lb (Figure 1) [1]. This time series is denoted X_t for $t = 1, 2, ...T$ where T is the number of observations.
- The univariate time series of retail prices in the USA in US cents/lb (Figure 2) [2]. This time series is denoted Y_t for $t = 1, 2, ...T$ where T is the number of observations.

Both time series range from January 1977 - December 2006 and can be viewed in a Google Spreadsheet (click to open Google Spreadsheet) [4].

As an alternative, the bivariate time series of Arabica Coffee in Colombia and the USA in US cents/lb (Figure 3) can be retrieved from the archive at FreeStatistics.org [3].

2.3 Analysis

Each hypothesis is investigated in turn, based on statistical models that are consistent with the theoretical concepts that are outlined in Chapter 1 (Explorative Data Analysis) [16].

2.3.1 Univariate Analysis of X_t

There are several ways to estimate α_X in $X_t = \alpha_X + e_t$ for $t = 1, 2, ...T$. First we investigate if the arithmetic mean could be an appropriate choice for estimating α_X. We know that the arithmetic mean can be very sensitive with respect to outliers. Therefore, we compute the winsorized and trimmed (arithmetic) mean of X_t (click to reproduce).

The arithmetic mean of X_t is 77.54 US cents/lb [5] when all observations are considered. However, if we “winsorize” the observations by making extreme

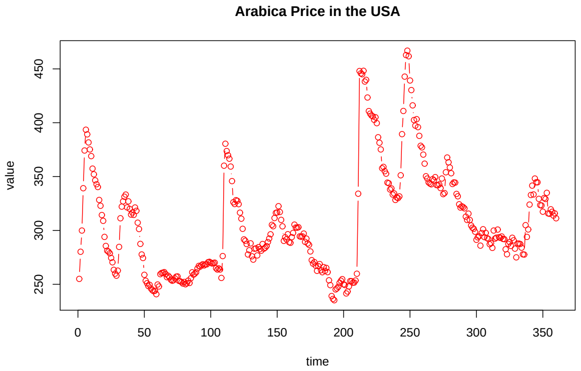


Figure 2: Run Sequence Plot - Arabica Coffee Price in the USA
link of reproducible computation

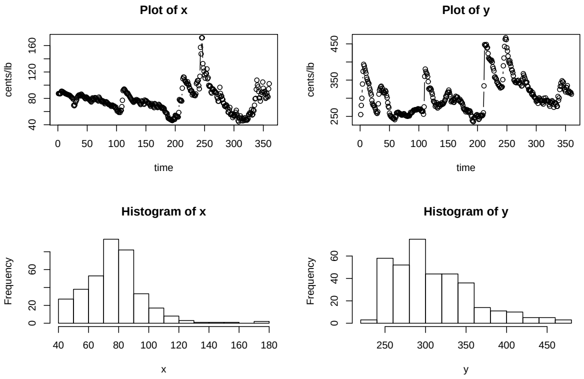


Figure 3: Run Sequence Plots and Histograms - Arabica Coffee Price in Colombia (x) and the USA (y)
link of reproducible computation

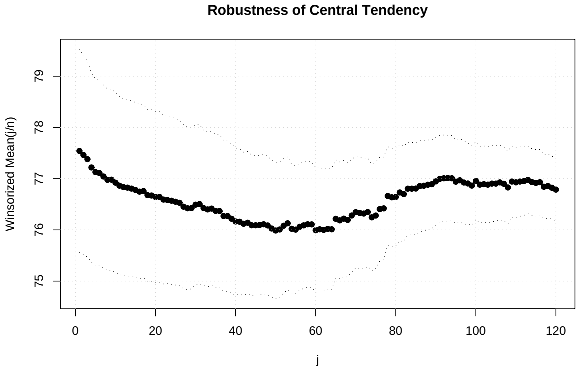


Figure 4: Robustness of Central Tendency - Arabica Coffee Price in Colombia
link of reproducible computation

values less extreme [10] then the result in Figure 4 is obtained. As we move along the x-axis (from left to right) we observe how the winsorized mean decreases as more extreme values are made less extreme. If we "winsorize" about 60 observations on both sides of the distribution ($j = 60$), the arithmetic mean is approximately 76 cents/lb. If we consider the hypothesis $H_0 : \alpha_X = 77.54$ and $H_A : \alpha_X \neq 77.54$ then we observe that 77.54 falls outside of the 95% confidence interval when $j = 60$. Hence, we reject the nullhypothesis that the arithmetic mean is equal to 77.54 at the 5% type I error level. In other words, the extreme values in X_t cause the arithmetic mean to be biased and therefore the arithmetic mean is probably not a good choice for α_X.

How about other measures of central tendency? It is often suggested that the median should be used instead of the arithmetic mean. The reseaon is that the median is not sensitive to outliers. Computation [5] shows that the median is 76.78 cents/lb which is smaller than the original arithmetic mean of 77.54 cents/lb. When the median and the arithmetic mean are unequal then we may conclude that the distribution is skewed (not symmetric). We now have to determine if the median and arithmetic mean are equal or significantly different. Unfortunately, hypothesis tests about the median are not easily derived. However, we could use notched boxplots (a tool from Explorative Data Analysis) combined with the Blocked Bootstrap method to further investigate this question.

The wessa.net website hosts an R module called "Blocked Bootstrap Plot - Central Tendency" which compares three measures of central tendency (arithmetic mean, median, and midrange) by means of simulation for any univariate time series (see also 1.3.3.4. Bootstrap Plot [16]). The idea is to simulate a (large) number of instances of the original time series by blocked bootstrapping. Then the three measures of central tendency are computed for each simulated series. The results can be summarized by the use of notched boxplots that allow us to compare them.

Instead of using the default R module we make some changes by "editing" the underlying R source code (click to reproduce) [6]:

- we eliminate the midrange computation (because it is of no interest for

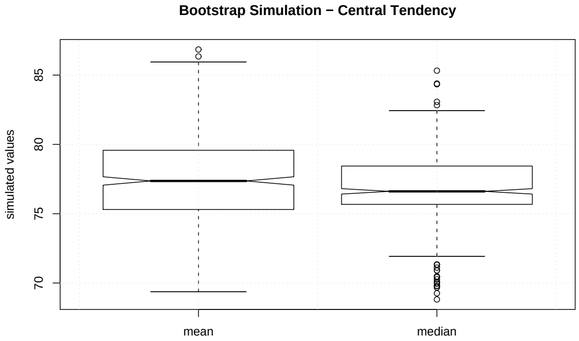


Figure 5: Notched Boxplots of Blocked Bootstrap - Arabica Coffee Price in Colombia

link of reproducible computation

our investigation)

- we add a table that displays the confidence intervals of the notched boxplot procedure as implemented in the R language
- we deleted all pictures with the exception of the notched boxplots (this is the only picture of interest for our investigation)

The result in Figure 5 clearly shows that the notches of both boxplots do not overlap (this can be verified in the table with confidence bounds [6]). The conclusion is that the median of simulated arithmetic means is different from the median of simulated medians. If the median and the mean of X_t are unequal then we may fairly assume the distribution of X_t to be skewed (compare this result with the histogram of Figure 3).

Let us continue the analysis with the median as estimator for α_X. We are interested to know if we can make fairly appropriate predictions of X_t for $t = T+1, T+2, ..., T+K$ where K is the forecast horizon. More formally the forecast is defined by $F_t \equiv X_t - e_t = \alpha_X$ for $t = T+1, T+2, ...T+K$ (where $\alpha_X = median(X_t)$ for $t = 1, 2, ...T$).

To test the hypothesis that the median is constant we need to address two questions:

1. is the median idependent of the month (seasonality)?
2. is the median constant on the long run?

The R module called “Mean Plot” allows us to investigate the median of periodic and sequential subseries as described in 1.3.3.20 Mean Plot [16]. The mean plot of X_t (click to reproduce) [7] clearly answers both questions:

1. there is no indication of seasonality in the time series (see Figure 6). Hence, the median does not depend on the month.
2. there is clear evidence of a trend-like behaviour in the time series under investigation (see Figure 7). Hence, the median is not constant on the long run.

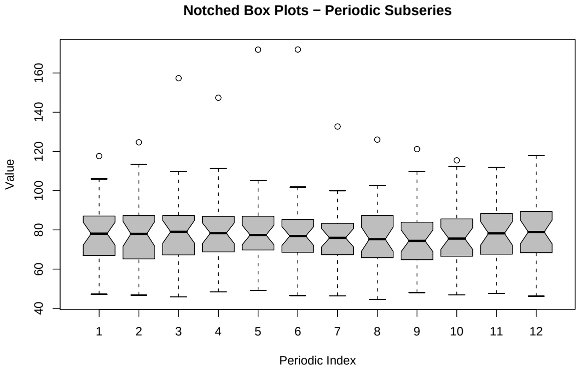


Figure 6: Notched Boxplots - Periodic Subseries - Arabica Coffee Price in Colombia

link of reproducible computation

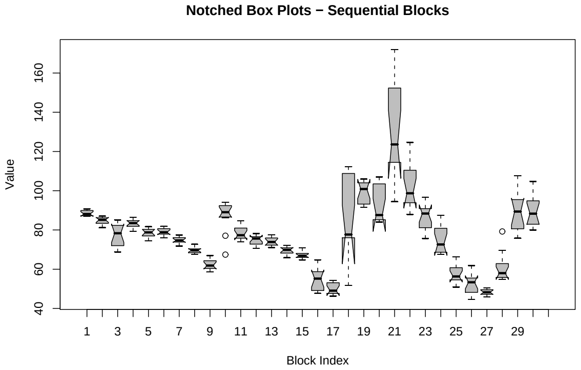


Figure 7: Notched Boxplots - Sequential Subseries - Arabica Coffee Price in Colombia

link of reproducible computation

Figure 7 is of particular interest in our analysis. It is clear from the analysis that the medians in sequential years are significantly different. Therefore the use of the median as a predictor ($F_t = \alpha_X$) is problematic when it is computed over the entire observation period. It looks like the recent past deserves more weight in computing α_X than the distant past.

The answer to Hypothesis 1 is clear. The null hypothesis $H_0 : \alpha_X = c$ is rejected and the alternative hypothesis $H_A : \alpha_X \neq c$ accepted. The model $X_t = \alpha_X + e_t$ for $t = 1, 2, ...T$ and the forecasts generated by $F_t = \alpha_X = c$ for $t = T + 1, T + 2, ...T + K$ are unacceptable.

Another way to investigate the proposed model is by means of the so-called 4-Plot as described in 1.3.3.32 4-Plot [16]. The implementation of the 4 Plot in wessa.net has 6 charts and is called "Univariate Explorative Data Analysis" (the Densityplot and Autocorrelation Function have been added) [13]. The underlying assumptions about the model can be investigated [8] with this R module and reveals that (click to reproduce):

- the time series does not behave like a constant (c.q. $X_t = \alpha_X + e_t$ is not appropriate)
- the variance is not constant
- the time series is not normally distributed (but skewed to the right)
- the time series is autocorrelated (and even contains high order autocorrelation)

2.3.2 Univariate Analysis of Y_t

To repeat the analysis that was performed in the previous section we simply need to substitute X_t for Y_t. There are several ways to estimate α_Y in $Y_t = \alpha_Y + e_t$ for $t = 1, 2, ...T$. First we investigate if the arithmetic mean could be an appropriate choice for estimating α_Y. We know that the arithmetic mean can be very sensitive with respect to outliers. Therefore, we compute the winsorized and trimmed (arithmetic) mean of Y_t (click to reproduce).

Assignment Fill in the blanks in the following text. The arithmetic mean of Y_t is ____ US cents/lb [9] when all observations are considered. If we "winsorize" more than 50 observations on both sides of the distribution, the arithmetic mean is significantly lower. In other words, the extreme values in Y_t cause the arithmetic mean to be biased and therefore the arithmetic mean is probably not a good choice for α_Y. As we move along the x-axis of the winsorized or trimmed mean chart, the computed mean (slowly) converges towards ____ US cents/lb which is the ________ (select from this list: midrange, midmean, harmonic mean, median, mode, geometric mean).

Assignment Investigate Y_t and answer the following questions:

1. Is the distribution of Y_t skewed? Use at least two different techniques to answer this question.
2. Does Y_t contain seasonality?
3. Are the forecasts generated by $F_t = \alpha_Y = c$ for $t = T + 1, T + 2, ...T + K$ acceptable?

Hint: make sure to blog (archive) your computations and include the hyperlinks of the archived computations in your document.

2.3.3 Regression Model 1

The third hypothesis under investigation focuses on the question if US retail prices of Arabica coffee can be explained by the price paid to growers in Colombia. We defined $H_0 : \beta = 0$ and $H_A : \beta > 0$ for the following equation: $Y_t = \alpha + \beta X_t + ... + e_t$ for $t = 1, 2, ...T$. Logically we expect that retail prices will reflect production costs which leads to the conclusion that the parameter β should be positive.

The hypothesis can be investigated by the use of the so-called 6-Plot (1.3.3.33 6-Plot [16]) which is basically an EDA-based graphical representation of the simple linear regression model. This analysis has been implemented in wessa.net with 9 charts (instead of 6) and is named “Linear Regression Graphical Model Validation” [14].

The result of this analysis [11] clearly shows that the null hypothesis $H_0 : \beta = 0$ is rejected in favor of the alternative $H_A : \beta > 0$ (click to reproduce). The economic interpretation is that the cost of production is reflected in the retail prices. The estimated value of the regression parameter $\hat{\beta} \simeq 1.84$ implies that an increase of 1 US cent/lb of X_t is multiplied by a factor 1.84 to get the estimated USA retail price $\hat{Y}_t$.

The result of this regression estimation should be subjected to careful “model validation” (testing the underlying model assumptions). Within the framework of EDA this is preferably done with graphical analysis.

Assignment Answer the following questions based on the archived computation:

1. Does the scatterplot between X_t and Y_t reveal/suggest a linear relationship?
2. Is there any evidence for autocorrelation? [2]
3. Are the estimated residuals normally distributed?
4. Is there any evidence for the presence of outliers?
5. Overall, do you think the assumptions of the model are satisfied?

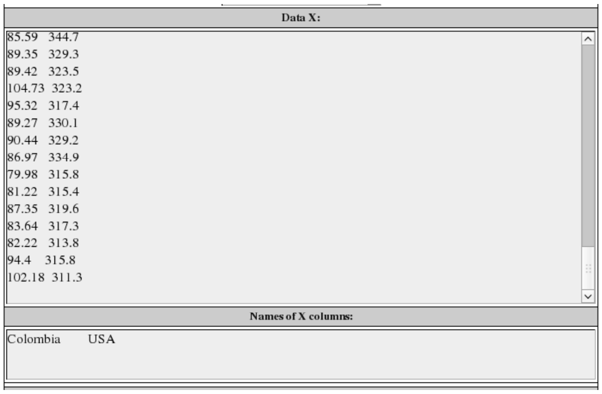
2.3.4 Regression Model 2

Let us suppose that the simple regression model of the previous section was misspecified (or incomplete) - in other words, let us suppose that a linear trend should have been added to the equation as defined in Hypothesis 4 of section 2.1.1.

We can’t use the 6-Plot (1.3.3.33 6-Plot [16]) from the previous section because we now have three variables instead of two. There are two exogenous (c.q. explanatory) variables (X_t and t) and one endogenous variable (Y_t). The newly added variable can be interpreted as a linear trend (or a long term effect of time). One might wonder if the addition of this new variable does make sense at all? Would it affect the estimated relationship between X_t and Y_t (represented by the parameter $\hat{\beta}$)?

We can find out the answer by making use of the multiple regression R module [15] which allows us to specify “multiple” (c.q. more than one) exogenous variables (hence the term “multiple regression”). In this case the data have to be entered as a multivariate dataset. The easiest way to do this is to copy the multivariate dataset from a spreadsheet and paste it into the Data X textbox

[2] You may not be familiar with the term autocorrelation. Therefore we could rephrase the question as follows: “Do prediction errors from the past contain any information that might help us to improve future predictions?” or “Is there any evidence that $\rho(e_t, e_{t-k}) \neq 0$ for $k = 1, 2, 3, ...$?”



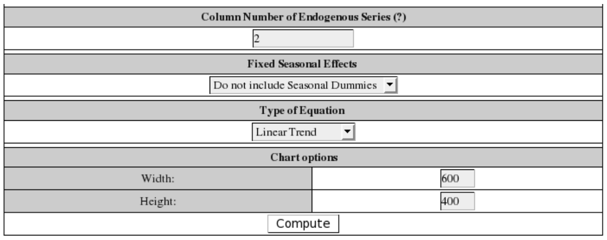
Data X:

```
85.59   344.7
89.35   329.3
89.42   323.5
104.73  323.2
95.32   317.4
89.27   330.1
90.44   329.2
86.97   334.9
79.98   315.8
81.22   315.4
87.35   319.6
83.64   317.3
82.22   313.8
94.4    315.8
102.18  311.3
```

Names of X columns:

Colombia	USA

Figure 8: Data Entry - Multiple Regression R module
link of R module



Column Number of Endogenous Series (?): 2

Fixed Seasonal Effects: Do not include Seasonal Dummies

Type of Equation: Linear Trend

Chart options

Width: 600

Height: 400

Figure 9: Parameters - Multiple Regression R module
link of R module

of the R module. You can use the Google Spreadsheet (click to open) that is available online.

The data should be pasted all at once: select the data range of all values in de spreadsheet and copy to the clipboard (with Ctrl-Insert). Then paste the contents of the clipboard into the "Data X" box of the R module. As a second step, we need to identiy the names of each column. We can do this by simply copying the contents of the cells (C1:D1) to the clipboard and pasting it into the "Names of X columns" box[3]. Figure 8 is an illustration (partial screendump) of the R module after the multivariate dataset was pasted into the application.

The model-specifying parameters of the multiple regression should be set correctly. First, we have to make sure that the "Column Number of Endogenous Series" field contains the number 2 because we want to define the second column as the endogenous series ($= Y_t$). Second, we select the option "Linear Trend" from the "Type of Equation" drop down list in order to make sure the software automatically generates an exogenous variable that represents time. An illustration of both model-specification parameters can be found in Figure 9.

The Multiple Regression R module produces a substantial amount of output. For the purpose of this tutorial however we focus exclusively on the aspects that are directly related to Hypothesis 4: is it necessary to include an additional trend variable to account for the long run increase in US retail prices? Based on the multiple regression analysis [12], the answer is affirmative[4]. The null hypothesis $H_0 : \gamma = 0$ is rejected in favor of the alternative $H_A : \gamma > 0$ as can be seen from the table with estimated regression parameters and respective p-values (click to reproduce).

The estimated parameters $\hat{\alpha} \simeq 135.7$, $\hat{\beta} \simeq 1.88$ (slightly higher than in regression 1) and $\hat{\gamma} \simeq 0.15$ can be substituted into the regression equation which yields: $\hat{Y}_t = 135.7 + 1.88X_t + 0.15t$ for $t = 1, 2, ...T$.

2.4 Important remark

As indicated in the introduction the analysis portrayed in this document is illustrative and simplistic in nature. More importantly, the underlying assumption of the multiple linear regression model are not satisfied. The multiple regression equation suffers from autocorrelation and heteroskedasticity which leads to inefficient estimation. It is highly likely that both problems are caused by a misspecification of the regression model (unobserved variables) which may lead to biased parameter estimates. As a consequence our conclusions about Hypothesis 3 and 4 should be revisited in the context of a more adequate (complete) model specification.

Assignment If you are familiar with autocorrelation and heteroskedasticity you may try to answer the following questions based on the archived computation:

[3] Note that the column names must not be too long nor contain any spaces!

[4] As with any regression model the validity of the hypothesis test depends on underlying assumptions. The answer that is provided here is conditional (c.q. given that the underlying assumputions are satisfied).

- Are the residuals of the regression model autocorrelated? How do you detect this?
- Is there any evidence for the presence of high-order autocorrelation?
- Does the model really suffer from heteroskedasticity? Use two different diagnostic tests to confirm your answer.
- Could the model be improved by taking into account seasonal effects? Hint: reproduce the model with monthly seasonal dummies.

References

[1] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL http://www.freestatistics.org/blog/date/2008/Jan/06/t1199574888xvrybp4pl7fyg1j.htm, Retrieved Sun, 06 Jan 2008 00:15:06 +0100

[2] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL http://www.freestatistics.org/blog/date/2008/Jan/06/t11995765663rnq0f6tqd9tvuc.htm, Retrieved Sun, 06 Jan 2008 00:42:50 +0100

[3] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL http://www.freestatistics.org/blog/date/2008/Jan/06/t11995770459lp663o7ht2jxj1.htm, Retrieved Sun, 06 Jan 2008 00:50:48 +0100

[4] http://spreadsheets.google.com/ccc?key=pV35do1d8bhGWVkmHcU2xRg&hl=en

[5] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL http://www.freestatistics.org/blog/date/2008/Jan/06/t1199655142l1mf45ba5ydsjhm.htm, Retrieved Sun, 06 Jan 2008 22:32:30 +0100

[6] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL http://www.freestatistics.org/blog/date/2008/Jan/07/t1199701558w9l4qhatxnfi326.htm, Retrieved Mon, 07 Jan 2008 11:26:04 +0100

[7] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL http://www.freestatistics.org/blog/date/2008/Jan/07/t119971350673encrjvyxgcqw7.htm, Retrieved Mon, 07 Jan 2008 14:45:24 +0100

[8] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL http://www.freestatistics.org/blog/date/2008/Jan/07/t1199715933ou67sm0izgd06ae.htm, Retrieved Mon, 07 Jan 2008 15:25:36 +0100

[9] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL http://www.freestatistics.org/blog/date/2008/Jan/19/t12007439926t0l73gl5ozrq14.htm, Retrieved Sat, 19 Jan 2008 12:59:59 +0100

[10] Borghers E. and P. Wessa, Descriptive Statistics - Central Tendency - Winsorized Mean, Office for Research Development and Education, URL http://www.xycoon.com/winsorized_mean.htm, Retrieved Mon, 07 Jan 2008 10:40 +0100

[11] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL http://www.freestatistics.org/blog/date/2008/Feb/26/t12040214435my4uqfrs09mgrk.htm, Retrieved Tue, 26 Feb 2008 11:24:09 +0100

[12] Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL http://www.freestatistics.org/blog/date/2008/Feb/26/t1204025214shjxx5mlpuxjln9.htm, Retrieved Tue, 26 Feb 2008 12:26:59 +0100

[13] Wessa P., (2007), Univariate Explorative Data Analysis (v1.0.5) in Free Statistics Software (v1.1.22-r4), Office for Research Development and Education, URL http://www.wessa.net/rwasp_edauni.wasp

[14] Wessa P., (2007), Linear Regression Graphical Model Validation (v1.0.2) in Free Statistics Software (v1.1.22-r4), Office for Research Development and Education, URL http://www.wessa.net/rwasp_linear_regression.wasp

[15] Wessa P., (2008), Multiple Regression (v1.0.25) in Free Statistics Software (v1.1.22-r4), Office for Research Development and Education, URL http://www.wessa.net/rwasp_multipleregression.wasp/

[16] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, Retrieved Mon, 07 Jan 2008 02:01 +0100.