CISF 2017

Residual analysis, a glimpse to the past to understand the future.

Mattia Ivaldi

12/05/2017

Bari, III Conferenza Italiana Studenti di Fisica

Outline

1. Residuals
    - What?
    - Why?
    - How?
2. Detecting lack of fit
3. Detecting outliers and identifying influential observations
4. Detecting residual correlation
5. Time series
    - Definition
    - Model design
    - Regression
    - Forecasting
6. Discussion

Residuals: what?

Let (y_i, x_{ij}), $i = 1, ..., N$, $j = 1, ..., k$ be a dataset and take a general *linear* regression model f depending on $k + 1$ parameters (β_j and β_0) and unknown random error ε

$$f(x_j, \beta_j) = \beta_0 + \sum_{(i,j)=1}^{k} \beta_j \, x_j + \varepsilon \tag{1}$$

[Mendenhall W. and Sincich T. (2012). *A second course in regression analysis*. Boston, US-MA: Prentice Hall.]

Residuals: what?

Let (y_i, x_{ij}), $i = 1, ..., N$, $j = 1, ..., k$ be a dataset and take a general *linear* regression model f depending on $k + 1$ parameters (β_j and β_0) and unknown random error ε

$$f(x_j, \beta_j) = \beta_0 + \sum_{(i,j)=1}^{k} \beta_j \, x_j + \varepsilon \tag{1}$$

Using the dataset to obtain least squares (LS) estimators $\hat{\beta}_j$, such that $\hat{y}_i = f(x_{ij}, \hat{\beta}_j)$, the i^{th} residual is

$$e_i = y_i - \hat{y}_i \tag{2}$$

The i^{th} partial residual for the j^{th} independent variable is

$$e_i^* = e_i + \hat{\beta}_j \, x_{ij} \tag{3}$$

[Mendenhall W. and Sincich T. (2012). *A second course in regression analysis*. Boston, US-MA: Prentice Hall.]

Coefficient of determination

What if $\chi^2 = 0$?

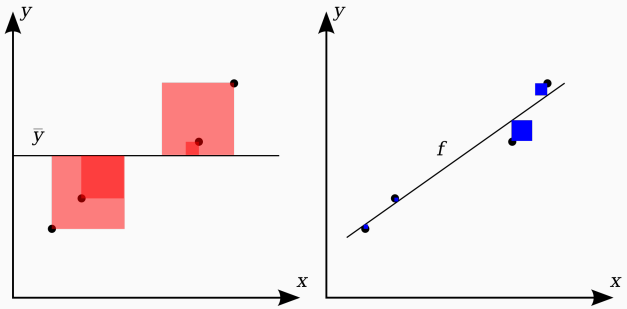
[Mendenhall W., Sincich T. (2012)]

Coefficient of determination

What if $\chi^2 = 0$? We can define the coefficient of determination as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} = 1 - \frac{SSE}{SST} \in [0, 1] \tag{4}$$

where $\bar{y}$ is the mean value of y_i. The better the linear regression fits the data in comparison to the mean value, the closer the value of R^2 is to 1. For instance $R^2 = 0.6$ means that the sum of squares of deviations of the y-values about their predicted values has been reduced 60% by using the least squares equation f, instead of $\bar{y}$, to predict y.



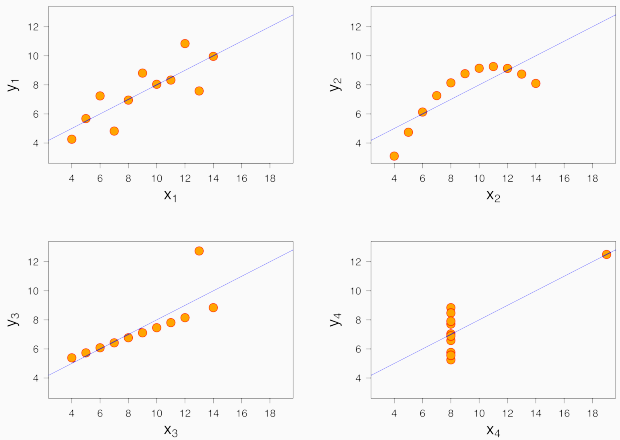
[Mendenhall W., Sincich T. (2012)]

Residuals: why?

- R^2 is monotone increasing with the number of variables included so it can always be increased by adding more variables into the model.

Residuals: why?

- R^2 is monotone increasing with the number of variables included so it can always be increased by adding more variables into the model.
- An R^2 close to one does not guarantee that the model fits the data well, because as Anscombe's quartet shows, a high R^2 can occur in the presence of misspecification of the functional form of a relationship or in the presence of outliers that distort the true relationship.



1. *Usefulness of graphs*

Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:

(1) numerical calculations are exact, but graphs are rough;
(2) for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;
(3) performing intricate calculations is virtuous, whereas actually looking at the data is cheating.

Anscombe's quartet: each dataset has same mean of x and y, same variance of x and y and $R^2 = 0.816$.

Picture courtesy of Schutz, licensed under CC-BY-SA 3.0.

[Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician.* **27** (1), 17–21.]

Residuals: how?

The residual analysis is based on two fundamentals properties of regression residuals.

1. The mean of the residuals is equal to 0.

$$E(e) = \sum_{i=1}^{k} e_i = \sum_{i=1}^{k} (y_i - \hat{y}_i) = 0 \tag{5}$$

2. The standard deviation of the residuals is equal to the standard deviation of the fitted regression model, s.

$$\sum_{i=1}^{k} e_i^2 = \sum_{i=1}^{k} (y_i - \hat{y}_i)^2 = SSE$$

$$s = \sqrt{\frac{SSE}{N - (k+1)}} \tag{6}$$

Residuals: how?

The residual analysis can be formed by the following steps:

- Detecting lack of fit
- Detecting heteroscedasticity
- Checking the normality assumption (properties 1. and 2.)
- Detecting outliers and identifying influential observations
- Detecting residual correlation

Detecting lack of fit

Assume that the model f (1) is correctly specified. The assumption $E(\varepsilon) = 0$ implies that $E(f) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$. Now, suppose an analyst hypothesizes a misspecified model $f = E_m(f) + \varepsilon$, $E_m(f) \neq E(f)$. Thus $\varepsilon = f - E_m(f)$ and $E(\varepsilon) = E(f) - E_m(f) \neq 0$.

Detecting lack of fit

Assume that the model f (1) is correctly specified. The assumption $E(\varepsilon) = 0$ implies that $E(f) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$. Now, suppose an analyst hypothesizes a misspecified model $f = E_m(f) + \varepsilon$, $E_m(f) \neq E(f)$. Thus $\varepsilon = f - E_m(f)$ and $E(\varepsilon) = E(f) - E_m(f) \neq 0$.

In order to detect lack of fit it is possible to plot the residuals against each of the independent variables or against the predicted value and look for trends, dramatic changes in variability, and/or more than 5% of residuals that lie outside $2s$ of 0.

An alternative method of detecting lack of fit in models with more than one independent variable is to construct a partial residual plot.

Detecting lack of fit

A supermarket chain wants to investigate the effect of price p on the weekly demand d for a house brand of coffee at its stores. Eleven prices were randomly assigned to the stores and were advertised using the same procedures. A few weeks later, the chain conducted the same experiment using no advertisements.

Detecting lack of fit

A supermarket chain wants to investigate the effect of price p on the weekly demand d for a house brand of coffee at its stores. Eleven prices were randomly assigned to the stores and were advertised using the same procedures. A few weeks later, the chain conducted the same experiment using no advertisements. Consider the model:

$$E_1(d) = \beta_0 + \beta_1\, p + \beta_2\, a \tag{7}$$

$$a = \begin{cases} 1, & \text{if advertisment used} \\ 0, & \text{if not} \end{cases} \tag{8}$$

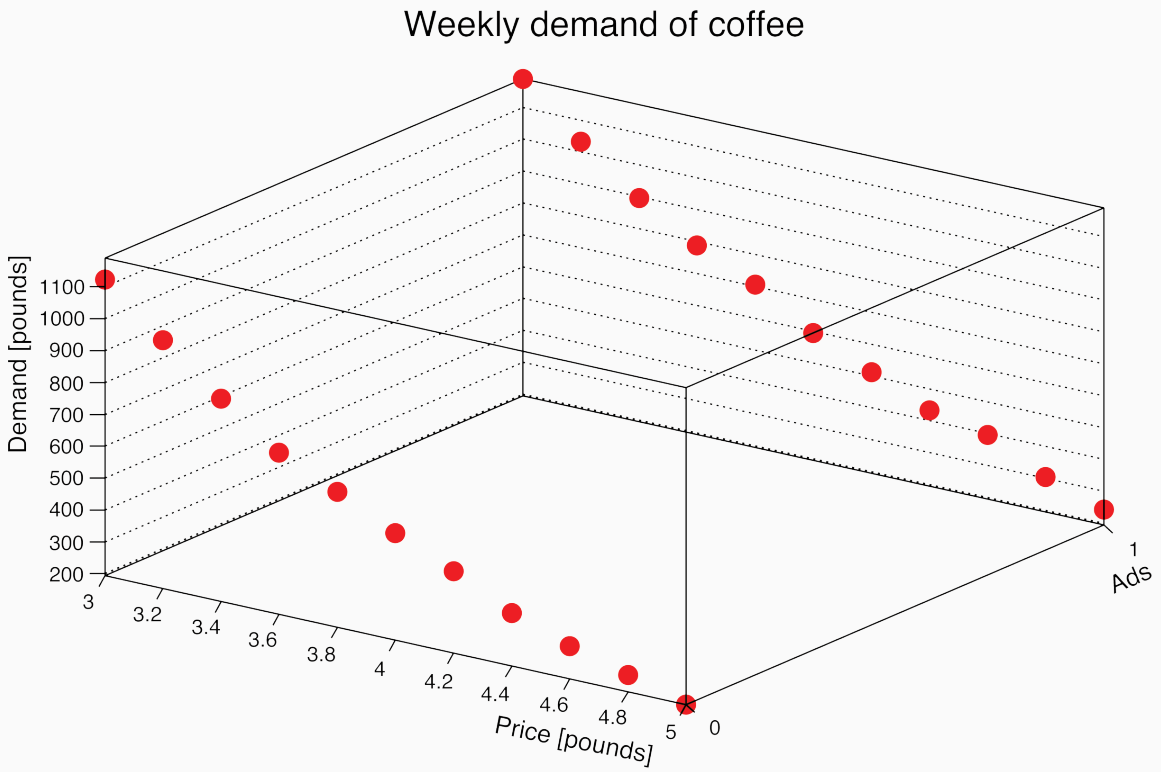
Is this model adequate for predicting weekly demand d?

Detecting lack of fit

The regression provides:

$$R^2 = 0.975 \quad s = 49.876\,£$$

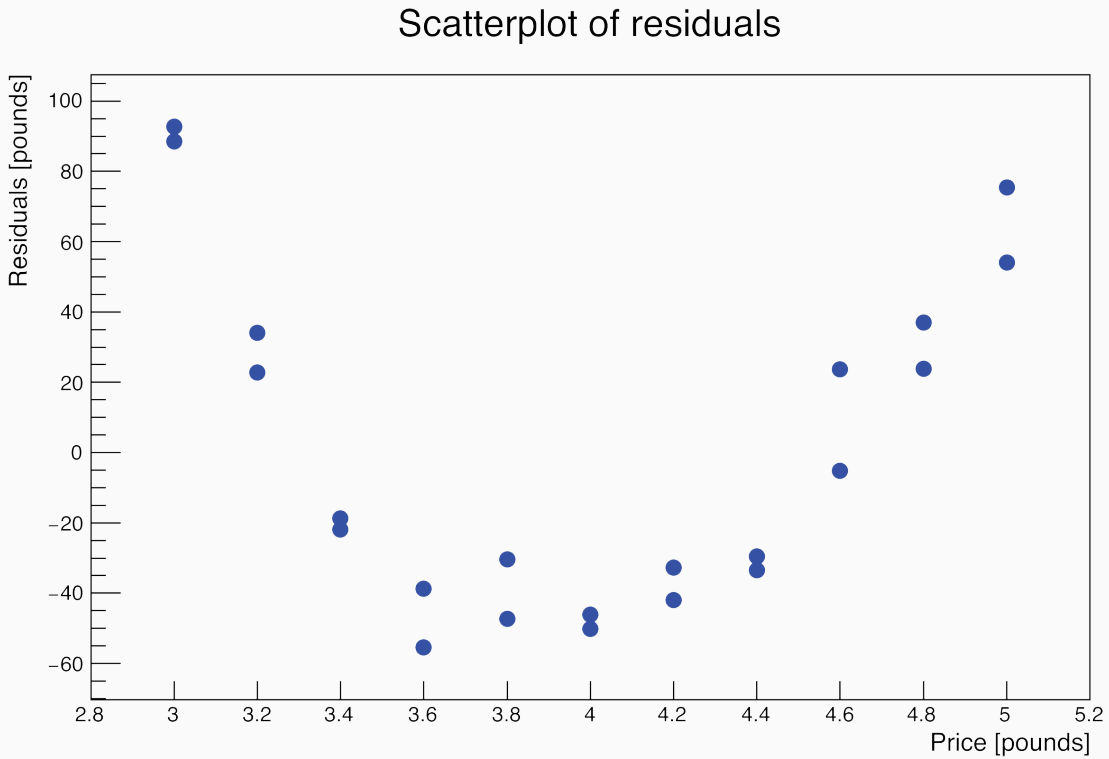
Is it really good?



Dataset from [Mendenhall W., Sincich T. (2012)]

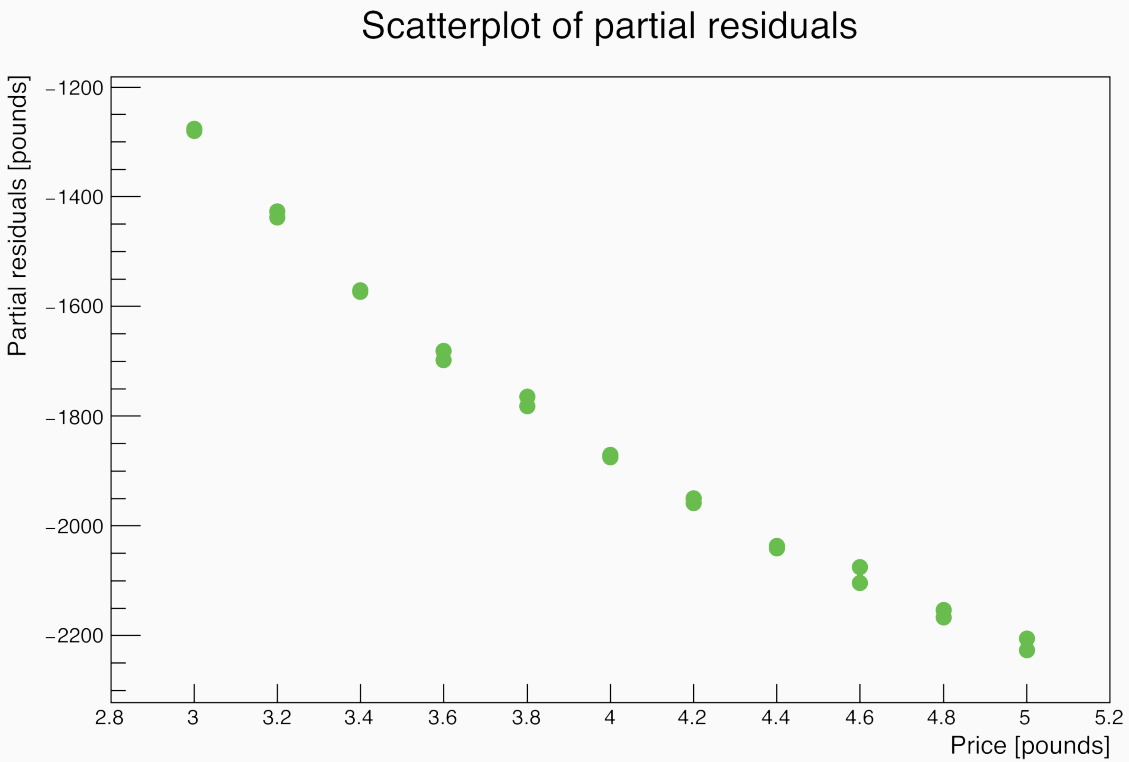
Detecting lack of fit

The residuals versus p plot reveals a clear parabolic trend, implying a lack of fit. Thus, the residual plot supports our hypothesis that the weekly demand-price relationship is curvilinear, not linear. A transformation of dependent variable is needed but the residual plot may not reveal the appropriate transformation.



Detecting lack of fit

The partial residual plot for the independent variable p also reveals a curvilinear trend, but, in addition, displays the correct functional form of weekly demand-price relationship. The curve is decreasing and approaching (but never reaching) 0 as p increases. This suggests that the appropriate transformation on price is either $1/p$ or e^{-p}.



Detecting lack of fit

Using the model:

$$E_2(d) = \beta_0 + \frac{\beta_1}{p} + \beta_2\, a \tag{9}$$

although R^2 increased only slightly (from 0.975 to 0.999), the model standard deviation s decreased significantly (from 49.876 to 11.097).

Whereas the model with untransformed price can predict weekly demand for coffee to within $2s = 2(50) = 100$ £, the transformed model can predict demand to within $2s = 2(11) = 22$ £.

Detecting outliers and identifying influential observations

If the i^{th} observation has $e_i \geq |3s|$ it is considered to be an outlier. Outliers are usually attributable to one of several causes and the measurement associated with the outlier may be invalid.

For example, the experimental procedure used to generate the measurement may have malfunctioned, the experimenter may have misrecorded the measurement, or the data may have been coded incorrectly for entry into the computer.

[Mendenhall W., Sincich T. (2012)]

Detecting outliers and identifying influential observations

Using the matricial notation it is possible to show that the LS estimators are given by the vector:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \tag{10}$$

The predicted values $\hat{y}_i$ can then be written as:

$$\hat{\mathbf{Y}} = \mathbf{X}\,\hat{\boldsymbol{\beta}} = \mathbf{X}\,(\mathbf{X}^\top \mathbf{X})^{-1}\,\mathbf{X}^\top\,\mathbf{Y} = \mathbf{H}\,\mathbf{Y} \tag{11}$$

where $\mathbf{H} \in \mathbb{R}^{n \times n}$ *puts the hat on* y and is therefore referred to as the *hat matrix*.

[Hoaglin D. C. and Welsch R. E. (1978). The Hat Matrix in Regression and ANOVA. *The American Statistician*. **32** (1), 17-22.]

Detecting outliers and identifying influential observations

Using the matricial notation it is possible to show that the LS estimators are given by the vector:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \tag{10}$$

The predicted values $\hat{y}_i$ can then be written as:

$$\hat{\mathbf{Y}} = \mathbf{X}\,\hat{\boldsymbol{\beta}} = \mathbf{X}\,(\mathbf{X}^\top \mathbf{X})^{-1}\,\mathbf{X}^\top\,\mathbf{Y} = \mathbf{H}\,\mathbf{Y} \tag{11}$$

where $\mathbf{H} \in \mathbb{R}^{n \times n}$ *puts the hat on* y and is therefore referred to as the *hat matrix*. Thus it is possible to write that

$$\hat{y}_i = \sum_{i=1}^{N} h_i\, y_i \tag{12}$$

The coefficients $h_i = \mathbf{x}_i^\top\,(\mathbf{X}^\top \mathbf{X})^{-1}\,\mathbf{x}_i$ are the diagonal elements of $\mathbf{H}$ and are referred to as the leverage of the i^{th} observation.

[Hoaglin D. C. and Welsch R. E. (1978). The Hat Matrix in Regression and ANOVA. *The American Statistician*. **32** (1), 17-22.]

Detecting outliers and identifying influential observations

The leverage h_i measures the influence of y_i on its predicted value $\hat{y}_i$. It is usually compared with the average leverage value of all N observations:

$$h^* = \frac{k+1}{N} \tag{13}$$

A rule of thumb for detecting influence with leverage states that the observed value of y_i is influential if:

$$h_i > 2\,h^* \tag{14}$$

[Mendenhall W., Sincich T. (2012)]

Detecting outliers and identifying influential observations

A more refined way to identify influential observation is to measure the overall influence an outlying observation has on the LS estimators.

[Cook D. R. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*. **19** (1), 15-18.]

Detecting outliers and identifying influential observations

A more refined way to identify influential observation is to measure the overall influence an outlying observation has on the LS estimators. If $\mathbf{e} = (\mathbb{1} - \mathbf{H})\mathbf{Y}$ is the residuals vector, the Cook's distance for the i^{th} observation is calculated as:

$$D_i = \frac{e_i^2}{s^2\,(k+1)}\left[\frac{h_i}{(1-h_i)^2}\right] = \frac{e_i^2}{\mathbf{e}^\top\mathbf{e}}\left[\frac{1-h^*}{h^*}\right]\left[\frac{h_i}{(1-h_i)^2}\right] \qquad (15)$$

D_i is a summary measure of the distance between $\hat{\beta}_i$ and β_i. A large value of D_i indicates that the observed value has strong influence on the estimated LS coefficients. The values of D_i can be compared to the values of the F-distribution with $\nu_1 = k+1$ and $\nu_1 = N - (k+1)$ degrees of freedom.

[Cook D. R. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*. **19** (1), 15-18.]

Detecting residual correlation

Positive (negative) serial (or auto-) correlation is serial correlation in which a positive (negative) error for one observation increases the chances of a positive (negative) error for another observation.

[Durbin J. and Watson G.S. (1950). Testing for Serial Correlation in Least Squares Regression I. *Biometrika*. **37** (3/4), 409-428.]

[Kanji G. K. (2006). *100 statistical tests*. London, UK: SAGE Publications.]

Detecting residual correlation

Positive (negative) serial (or auto-) correlation is serial correlation in which a positive (negative) error for one observation increases the chances of a positive (negative) error for another observation.

The Durbin-Watson (DB) test allows the detection of residual correlation. It is based on the residual model $e_i = \phi\, e_{i-1} + u_i$, where ϕ is the autocorrelation parameter and $u \sim \mathcal{N}(0, \sigma^2)$. When one is concerned with positive autocorrelation the alternatives are given as: $H_0 : \phi \leq 0, \ , H_1 : \phi > 0$. Here H_0 implies that error terms are uncorrelated or negatively correlated, while H_1 implies that they are positively autocorrelated.

[Durbin J. and Watson G.S. (1950). Testing for Serial Correlation in Least Squares Regression I. *Biometrika*. **37** (3/4), 409-428.]

[Kanji G. K. (2006). *100 statistical tests*. London, UK: SAGE Publications.]

Detecting residual correlation

The test statistic is

$$d = \frac{\sum_{i=2}^{N}(e_i - e_{i-1})^2}{\sum_{i=1}^{N} e_i^2} \in [0, 4] \tag{16}$$

Let be α the significance of the test, if $d < d_{L,\alpha}$ there is statistical evidence that the residuals are positively autocorrelated, if $d > d_{U,\alpha}$ there is no statistical evidence that the residuals are positively autocorrelated and if $d_{L,\alpha} < d < d_{U,\alpha}$ the test is inconclusive. In order to test the negative autocorrelation the test statistic is $(4-d)$.

[Durbin J. and Watson G.S. (1950)]

[Kanji G. K. (2006)]

Time series – definition

A time series (TS) is a collection of data obtained by observing a response variable at periodic points in time. In many practical applications the objective is to forecast some future value or values of the series. To obtain forecasts, some type of model must be used to describe the TS.

[Mendenhall W., Sincich T. (2012)]

Time series – definition

A time series (TS) is a collection of data obtained by observing a response variable at periodic points in time. In many practical applications the objective is to forecast some future value or values of the series. To obtain forecasts, some type of model must be used to describe the TS. One of the most widely used models is the additive model:

$$y_t = T_t + C_t + S_t + R_t \tag{17}$$

where the *secular* (T_t) trend is the tendency of the TS to increase o decrease over a long period of time, the *cyclical* (C_t) and *seasonal* (S_t) variation are oscillating patterns with different period. The *residual* effect (T_t) is not systematic, it may be attributed to unpredictable influences and it represents the random error component of a TS.

[Mendenhall W., Sincich T. (2012)]

Time series – model design

The TS model design can be summarize in the choose of the deterministic and the residual component. The deterministic component should be chosen as the most suitable accordingly to the dataset. For example, the model of a monthly series might be:

$$E(y_t) = \beta_0 + \beta_1 \cos\left(\frac{\pi}{6}t\right) + \beta_2 \sin\left(\frac{\pi}{6}t\right) \tag{18}$$

[Mendenhall W., Sincich T. (2012)]

Time series – model design

The TS model design can be summarize in the choose of the deterministic and the residual component. The deterministic component should be chosen as the most suitable accordingly to the dataset. For example, the model of a monthly series might be:

$$E(y_t) = \beta_0 + \beta_1 \cos\left(\frac{\pi}{6}t\right) + \beta_2 \sin\left(\frac{\pi}{6}t\right) \tag{18}$$

The appropriate form of the residual component must take in account the presence of autocorrelation in the residuals. If residuals at times t and $(t+m)$ are correlated we can define the m^{th}-order autoregressive error model:

$$R_t = \sum_{i=1}^{m} \phi_i \, R_{t-i} + \gamma_t \quad \phi_i \in (-1, 1) \tag{19}$$

where the white noise γ_t is uncorrelated with any and all other residual components.

[Mendenhall W., Sincich T. (2012)]

Residuals: why?

- R^2 is monotone increasing with the number of variables.
- An R^2 close to one does not guarantee that the model fits the data well.
- If the dataset is a TS, although the LS estimators $\hat{\beta}_j$ remain unbiased even if the residuals are autocorrelated, the standard errors given by LS theory are usually smaller than the true standard error when the residuals are positively autocorrelated. Consequently, t-values will usually be inflated and will lead to a higher Type I error rate.

Thus, the application of LS techniques to time series can produce misleading statistical test results that result in overoptimistic evaluations of a model's predictive ability.

Time series – regression

Consider the simple general model:

$$y_t = \beta_0 + \beta_1 t + R_t, \quad R_t = \phi\, R_{t-1} + \gamma_t \tag{20}$$

A modification of the LS method is required.

[Fuller W. A. (1996). *Introduction to statistical time series.* Toronto, CA: Wiley-Interscience.]

Time series – regression

Consider the simple general model:

$$y_t = \beta_0 + \beta_1 t + R_t, \quad R_t = \phi\, R_{t-1} + \gamma_t \tag{20}$$

A modification of the LS method is required. First, we multiply the model by ϕ at time $(t-1)$ and take the difference between the two equations. Then we apply the transformation $y_t^* = y_t - \phi\, y_{t-1}$, $t^* = t - \phi\,(t-1)$, $\beta_0^* = \beta_0\,(1-\phi)$ to obtain:

$$y_t^* = \beta_0^* + \beta_1\, t^* + \gamma_t \tag{21}$$

[Fuller W. A. (1996). *Introduction to statistical time series.* Toronto, CA: Wiley-Interscience.]

Time series – regression

Consider the simple general model:

$$y_t = \beta_0 + \beta_1 t + R_t, \quad R_t = \phi\, R_{t-1} + \gamma_t \tag{20}$$

A modification of the LS method is required. First, we multiply the model by ϕ at time $(t-1)$ and take the difference between the two equations. Then we apply the transformation $y_t^* = y_t - \phi\, y_{t-1}$, $t^* = t - \phi\,(t-1)$, $\beta_0^* = \beta_0\,(1-\phi)$ to obtain:

$$y_t^* = \beta_0^* + \beta_1\, t^* + \gamma_t \tag{21}$$

Now it is possible to apply the LS method and the estimator of the original intercept can be calculated by:

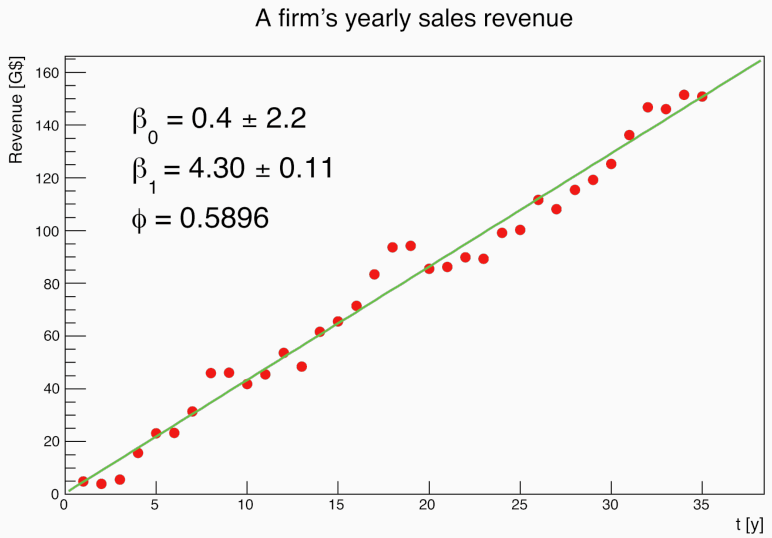
$$\hat{\beta}_0 = \frac{\hat{\beta}_0^*}{1-\phi} \tag{22}$$

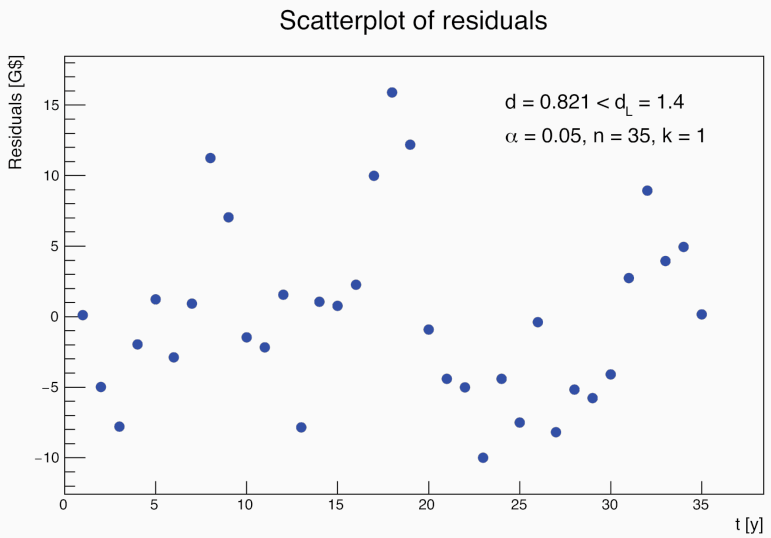
Where the knowing of ϕ and the adjustments for the values at $t = 1$ – the values of y_t^* and t^* can be calculated only for $t \geq 2$ – is required (very hard but possible!).

[Fuller W. A. (1996). *Introduction to statistical time series.* Toronto, CA: Wiley-Interscience.]

Time series – regression

For example consider the data on annual sales (G$) for a firm in each of its 35 years of operation. We want to forecast the revenue in the following years: the first-order linear model $y_t = \beta_0 + \beta_1\, t + \varepsilon$ seems to fit the data very well, since $R^2 = 0.98$ but the DB test shows the presence of positive autocorrelation. An autoregressive error model must be take in account to evaluate the set $(\hat{\beta}_0, \hat{\beta}_1, \hat{\phi})$.





Dataset from [Mendenhall W., Sincich T. (2012)]

Time series – forecasting

From the dataset (y_n, t_n) we now want to forecast the value of:

$$\begin{aligned} y_{n+1} &= \beta_0 + \beta_1\, t_{n+1} + R_{n+1} \\ &= \beta_0 + \beta_1\, t_{n+1} + \phi\, R_n + \gamma_{n+1} \end{aligned} \tag{23}$$

Time series – forecasting

From the dataset (y_n, t_n) we now want to forecast the value of:

$$\begin{aligned} y_{n+1} &= \beta_0 + \beta_1\, t_{n+1} + R_{n+1} \\ &= \beta_0 + \beta_1\, t_{n+1} + \phi\, R_n + \gamma_{n+1} \end{aligned} \tag{23}$$

Setting γ_{n+1} to its expected value of 0 we obtain:

$$\begin{aligned} F_{n+1} &= \hat{\beta}_0 + \hat{\beta}_1\, t_{n+1} + \hat{\phi}\, \hat{R}_n \\ &= \hat{\beta}_0 + \hat{\beta}_1\, t_{n+1} + \hat{\phi}\, [y_n - (\hat{\beta}_0 + \hat{\beta}_1\, t_n)] \end{aligned} \tag{24}$$

Time series – forecasting

From the dataset (y_n, t_n) we now want to forecast the value of:

$$\begin{aligned} y_{n+1} &= \beta_0 + \beta_1\, t_{n+1} + R_{n+1} \\ &= \beta_0 + \beta_1\, t_{n+1} + \phi\, R_n + \gamma_{n+1} \end{aligned} \tag{23}$$

Setting γ_{n+1} to its expected value of 0 we obtain:

$$\begin{aligned} F_{n+1} &= \hat{\beta}_0 + \hat{\beta}_1\, t_{n+1} + \hat{\phi}\, \hat{R}_n \\ &= \hat{\beta}_0 + \hat{\beta}_1\, t_{n+1} + \hat{\phi}\, [y_n - (\hat{\beta}_0 + \hat{\beta}_1\, t_n)] \end{aligned} \tag{24}$$

The two-step-ahead is:

$$\begin{aligned} F_{n+2} &= \hat{\beta}_0 + \hat{\beta}_1\, t_{n+2} + \hat{\phi}\, \hat{R}_{n+1} \\ &= \hat{\beta}_0 + \hat{\beta}_1\, t_{n+2} + (\hat{\phi})^2\, \hat{R}_n \\ &= \hat{\beta}_0 + \hat{\beta}_1\, t_{n+2} + (\hat{\phi})^2\, [y_n - (\hat{\beta}_0 + \hat{\beta}_1\, t_n)] \end{aligned} \tag{25}$$

This iterative process leads to the general – for the model (20) – m-step-ahed forecast:

$$F_{n+m} = \hat{\beta}_0 + \hat{\beta}_1\, t_{n+m} + (\hat{\phi})^m\, [y_n - (\hat{\beta}_0 + \hat{\beta}_1\, t_n)] \tag{26}$$

For example we obtain:

$$F_{36} = 155.2 \text{ G\$}$$
$$F_{37} = 159.4 \text{ G\$}$$
$$F_{38} = 163.7 \text{ G\$}$$

Does the potential for error increases as the distance into the future increases?

Discussion

- Given an experimental dataset and a model $f = E(f) + \varepsilon$, the residuals can be computed as the difference of the observed and the predicted values of the dependent variables. Because residuals estimate the true random error, they can be used to check the regression assumptions.
- The residuals allows to easily detect a misspecification of the model, the presence of influential observation and the presence of autocorrelation.
- The residuals provide a useful tool in the time series study and forecasting.
- In general, the residual analysis is a powerful and easy-to-use instrument to perform regression analysis, also by the educational side.

References

[1] Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician.* **27** (1), 17–21.

[2] Cook D. R. (1977). Detection of Influential Observation in Linear Regression. *Technometrics.* **19** (1), 15-18.

[3] Durbin J. and Watson G.S. (1950). Testing for Serial Correlation in Least Squares Regression I. *Biometrika.* **37** (3/4), 409-428.

[4] Hoaglin D. C. and Welsch R. E. (1978). The Hat Matrix in Regression and ANOVA. *The American Statistician.* **32** (1), 17-22.

[5] Mendenhall W. and Sincich T. (2012). *A second course in regression analysis.* Boston, US-MA: Prentice Hall.

[6] Fuller W. A. (1996). *Introduction to statistical time series.* Toronto, CA: Wiley-Interscience.

[7] Kanji G. K. (2006). *100 statistical tests.* London, UK: SAGE Publications.

Igrajte za nju
Našu voljenu
Nek jače kuca to
Srce vatreno

Detecting heteroscedasticity

One of the assumptions necessary for the validity of regression inferences is that the error term ε have constant variance σ^2 for all levels of the independent variable(s). Variances that satisfy this property are called homoscedastic. Unequal variances for different settings of the independent variable(s) are said to be heteroscedastic. Plots of the residuals against the predicted values of the response will frequently reveal the presence of heteroscedasticity.

When data fail to be homoscedastic, the reason is often that the variance of the response y is a function of its mean $E(y)$. In this case we can often satisfy the LS assumption of homoscedasticity by transforming the response to some new response that has a constant variance. These are called variance-stabilizing transformations. For example, if y is a count that follows a Poisson distribution, the square root transform $\sqrt{y}$ can be shown to have approximately constant variance.

[Mendenhall W., Sincich T. (2012)]

Checking the normality assumption

The simplest way to determine whether the data violate the assumption of normality is to construct a frequency or relative frequency distribution for the residuals. This visual check is not foolproof because we are lumping the residuals together for all settings of the independent variables. It is conceivable (but not likely) that the distribution of residuals might be skewed to the left for some values of the independent variables and skewed to the right for others. Combining these residuals into a single relative frequency distribution could produce a distribution that is relatively symmetric.

Another graphical technique is to construct a normal probability plot where the residuals are graphed against the expected values of the residuals under the assumption of normality. A linear trend suggests that the normality assumption is nearly satisfied, whereas a nonlinear trend indicates that the assumption is most likely violated.

[Mendenhall W., Sincich T. (2012)]

Non-normality and heteroscedasticity

Non-normality of the distribution of the random error ε is often accompanied by heteroscedasticity. Both these situations can frequently be rectified by applying the variance-stabilizing transformations. For example, if the relative frequency distribution of the residuals is highly skewed to the right (as it would be for Poisson data), the square-root transformation on y will stabilize (approximately) the variance and, at the same time, will reduce skewness in the distribution of the residuals. Thus, for any given setting of the independent variables, the square-root transformation will reduce the larger values of y to a greater extent than the smaller ones. This has the effect of reducing or eliminating the positive skewness. For situations in which the errors are homoscedastic but non-normal, normalizing transformations are also available (see *Box–Cox approach*).

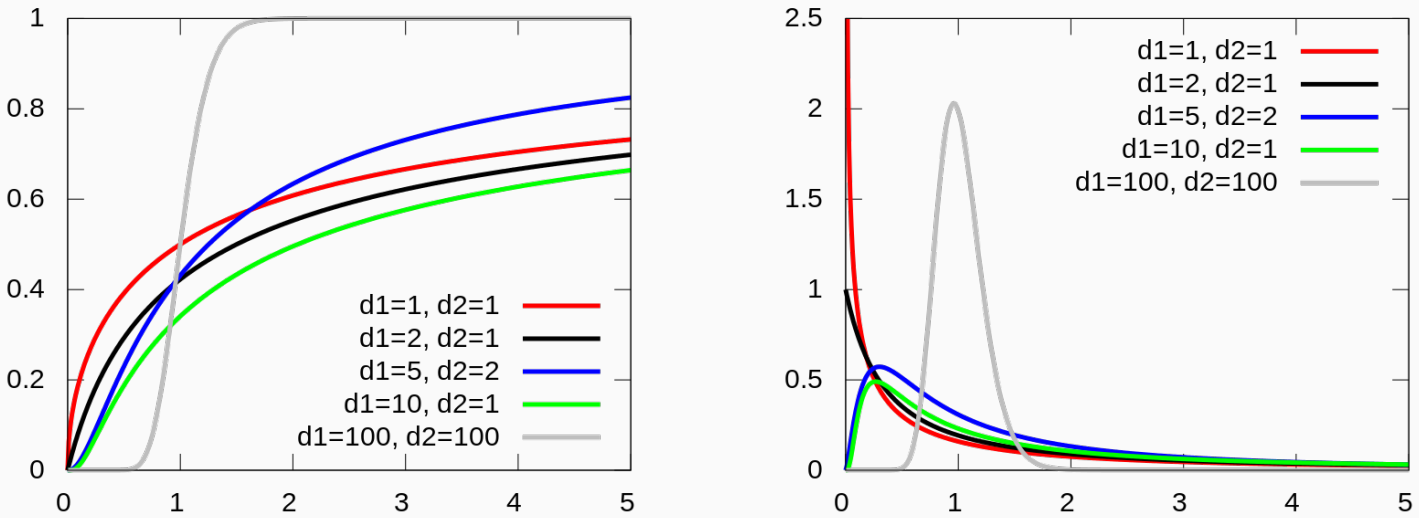
[Mendenhall W., Sincich T. (2012)]

F-distribution

If a random variable X has an Fisher-Snedecor distribution with parameters ν_1 and ν_2, we write $X \sim F(\nu_1, \nu_2)$.

$$f(x; \nu_1, \nu_2) = \frac{\sqrt{\frac{(\nu_1 x)^{\nu_1}\ \nu_2^{\nu_2}}{(\nu_1 x + \nu_2)^{\nu_1+\nu_2}}}}{x\,\mathrm{B}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)}, \ \Re(x) \geq 0 \tag{27}$$

$$F(x; \nu_1, \nu_2) = I_{\frac{\nu_1 x}{\nu_1 x + \nu_2}}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right), \ I_x(a, b) = \frac{\mathrm{B}(x; a, b)}{\mathrm{B}(a, b)} \tag{28}$$



PDF and CDF of the F-distribution. Picture courtesy of IkamusumeFan, licensed under CC-BY-SA 3.0.

Consider the model (20), we take:

$$\begin{aligned} y_t &= \beta_0 \quad + \beta_1\, t \qquad\quad + R_t \quad - \\ \phi\, y_{t-1} &= \phi\, \beta_0 + \phi\, \beta_1\, (t-1) + \phi\, R_{t-1} = \end{aligned}$$

$$y_t - y_{t-1} = \beta_0\, (1-\phi) + \beta_1\, [t - \phi\, (t-1)] + [R_t - \phi\, R_{t-1}] \qquad (29)$$

We apply the transformations $y_t^* = y_t - \phi\, y_{t-1}$, $t^* = t - \phi\, (t-1)$, $\beta_0^* = \beta_0\, (1-\phi)$ and we remember that $R_t = \phi\, R_{t-1} + \gamma_t$, obtaining:

$$y_t^* = \beta_0^* + \beta_1\, t^* + \gamma_t \qquad (30)$$

[Mendenhall W., Sincich T. (2012)]

Time series – forecasting

Forecast errors are traceable to three primary causes.

- The form of the model may change at some future time.
- The uncorrelated residual γ_t with variance σ^2. The approximate 95% forecasting limits using TS models with first-order autoregressive residuals for the m-step-ahed forecast is:

$$F_{n+m} \pm 2\, s \sqrt{\sum_{i=1}^{m} \hat{\phi}^{\,2(i-1)}} \tag{31}$$

- The error of estimating the model parameters.

[Fuller W. A. (1996)]

[Mendenhall W., Sincich T. (2012)]

Graphs are essential to good statistical analysis.

F. J. Anscombe, 1973