Diagnostics in SLR

stats544-1-slr Applied simple linear regression

Gunnar Stefansson

July 18, 2019

Introduction

Evaluate model assumption: $Y_i \sim n(\beta_0 + \beta_1 x_i, \sigma^2)$, independent.

- Linearity
- Independence
- Normality
- Constancy of variance

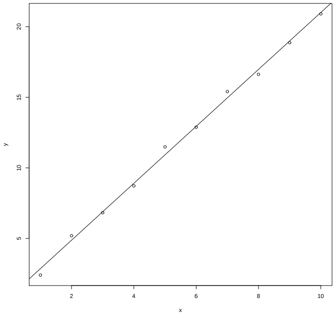


Figure : Simulated data

See influence.measures in R.

Residuals

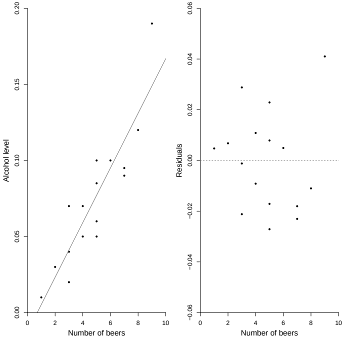
The first step in most diagnostic analyses is to compute the residuals

Residuals

The vertical distance from our measurements to the regression line are called the **residuals** and are denoted with $\hat{e}$. The size of the residuals can be calculated with

$$\hat{e}_i = y_i - \hat{y}_i$$

Points above the regression line have a positive residue but points below it have a negative.



Diagnostics based on residuals

Diagnostics for residuals include tests for normality and constancy of variance.
Semistudentized residuals $(e_i/\sqrt(MSE))$ are commonly used but studentized $e_i/\sqrt{(MSE)(1-h_{ii})}$
would obviously be better.

Verifying the distribution

There are several ways to verify that the residuals follow a normal distribution:

- Kolmogorov-Smirnov test
- Normal probability plot

Constancy of variance

If the variance is constant, then e^2 should not show a trend in any independent variable.
Simple test: Regress e^2 on x and test in usual manner.
Slightly more advanced: Breusch-Pagan test takes properties of e^2 into account.

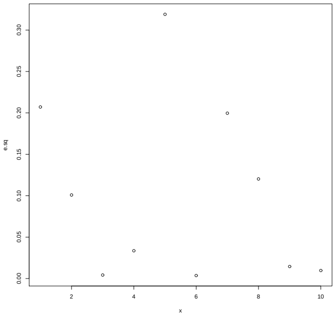


Figure : Base model with correct assumptions

Verifying linearity

Basic:

- Plot residuals against x-variable
- Look for pattern

Later:

- Test for autocorrelation
- Multiple regression: Add a quadratic term
- Lack-of-fit tests (replace x by a factor)

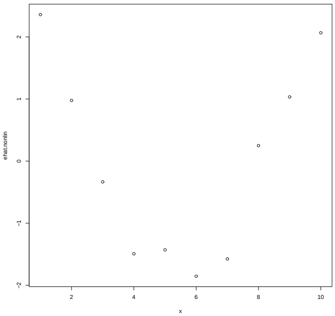


Figure : Residuals vs independent variable. Error in linearity assumption.

Tests are approximate

Testing for normality etc is only approximate

Most of the tests used for diagnostics are only approximate.

The Kolmogorov-Smirnov test is derived under the assumption that the distribution is fully specified under the null hypothesis. However, the residuals in OLS are computed after fitting a model and hence they are not independent.

Similarly when plotting e^2 against x.

Note that exact tests exist, but these simple approximate tests are often adequate.

Outliers

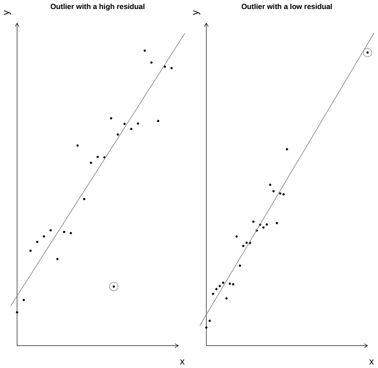


Figure : Outliers and their residuals.

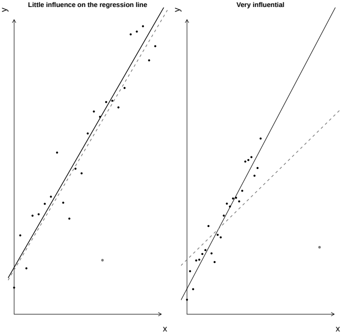
Outliers and influential cases

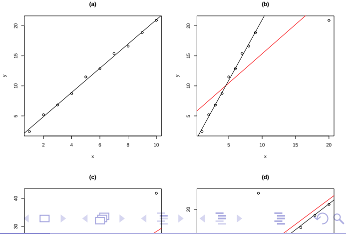
It is in particular important to search for outliers or influential cases in the x or y-measurements.

Typically use residuals and/or hat matrix:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

Methods for this will be introduced.





Same example as before - insert outliers in different locations and investigate effects.

Outliers in y - consider deleted residuals

Outliers can be considered a particular deviation from normality
Can base analysis on the concept

$$\frac{Y_h - (\hat{\beta}_0 + \hat{\beta}_1 x_h)}{\hat{\sigma}_{Y_h - \hat{Y}_h}} \sim t_{n-2}$$

i.e. use the deleted residual:

$$d_i = y_i - \hat{y}_{i(i)}$$

Computing deleted residuals

In principle, compute deleted residuals or studentized deleted residuals through fitting model without i'th observations, compute fitted, $\hat{y}_{i(i)}$, and compute $d_i = y_i - \hat{y}_{i(i)}$, $t_i = d_i / s_{d_i}$.
Simpler

$$t_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right]^{\frac{1}{2}}$$

Can use Bonferroni test with $t_{1-\alpha/(2n),n-p-1}$

Autocorrelation

Autocorrelation refers to correlation between Y_i and Y_{i+1}.
Only makes sense if i is "time".

Cooks distance

Measures total effect of i'th on all predictions

$$D_i = \frac{\sum_j \left(\hat{y}_j - \hat{y}_{i(i)}\right)^2}{pMSE}$$

Leverage values

Hat matrix $H = X(X'X)^{-1}X'$ so $\hat{y} = Hy$ and $\hat{e} = (I - H)y$ with $\Sigma_{\hat{e}} = \sigma^2(I - H)$ and $V(\hat{e}_i) = \sigma^2(1 - h_{ii})$.
h_{ii}=leverage values. $\sum_{i=1}^{n} h_{ii} = p \quad 0 \leq h_{ii} \leq 1$. Average h_{ii} is p/n so e.g. $2p/n$ is "large", or use rules of thumb such as 0.2 or 0.5 as "large" values.

Influential observations, DFFITS

Influential observations:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_i h_{ii}}} = t_i \left(\frac{h_{ii}}{1 - hii} \right)^{\frac{1}{2}}$$