Verifying the assumptions of SLR
(STATS545.3: Regression diagnostics)

Gunnar Stefansson

September 19, 2014

Introduction

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

- Linearity
- Independence
- Normality
- Constancy of variance

Start with diagnostics for $p = 2$.
See influence.measures in R.

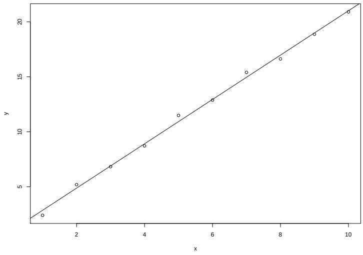


Figure: Simulated data

Residuals

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

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

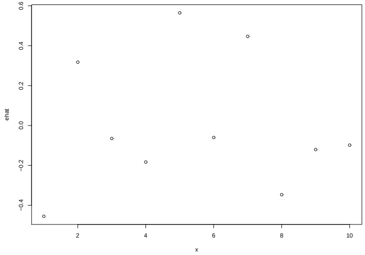


Figure: Residuals vs independent variable. No errors in model assumptions.

Verifying the distribution

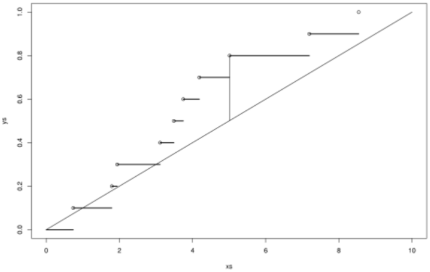
Kolmogorov-Smirnov: Compares data to a theoretical distribution

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} I_{[x_i,\infty)}(x) \text{ for } x \in \mathbf{R}$$

$$H_0 : P\left[X_i \leq x\right] = F(x) \text{ for } x \in \mathbf{R}.$$

The statistic:

$$D := \sup_x |F_n(x) - F(x)|.$$



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.

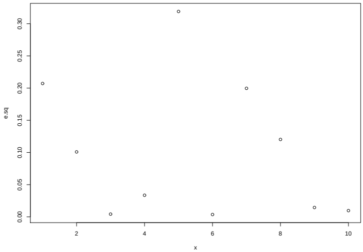
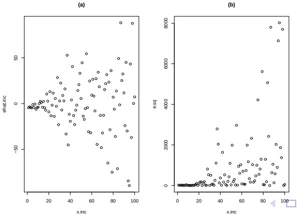


Figure: Base model with correct assumptions: e vs x.



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)

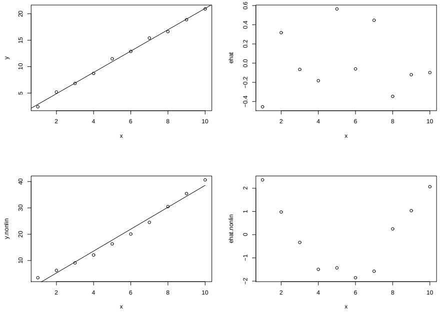


Figure: Residuals vs independent variable. Error in linearity assumption.

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.