Further diagnostics in SLR

(STATS545.3: Regression diagnostics)

Gunnar Stefansson

September 19, 2014

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.

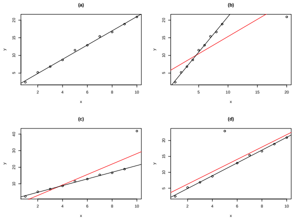

Figure: Effects of some outlier types on simple linear regression.

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.

Outliers in y - consider deleted residuals

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

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

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”.

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:

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

Cooks distance

Measures total effect of i'th on all predictions

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