Validating multiple regression models: Model diagnostics

(STATS545.3: Regression diagnostics)

Gunnar Stefansson

September 19, 2014

Introduction and overview

Diagnostics include: Same as in simple linear regression

Diagnostics for residuals (normality, y-outliers, constancy of variance)

Identifying X-outliers: Hat matrix

Identifying influential cases

Multicollinearity

Later: Testing for lack of fit

Diagnostics include the same as in simple linear regression, but more are possible in the multiple regression setting and more things can go wrong.

Partial regression plots: Regress on all x_1 and plot residual against residuals from regression of x_2 onto x_1. Provides indication of whether x_2 should be added and if so, how.

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

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

Note: One should always plot the residuals against $\hat{y}$ as well as against each independent variable.

Identifying y-outliers: 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 $V(\hat{e}_i) = \sigma^2(1-h_{ii})$.

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

Partial regressions

Partial regression plots: Regress on all x_1 and plot residual against residuals from regression of x_2 onto x_1. Provides indication of whether x_2 should be added and if so, how.

DFBETAS

DFBETAS - influence on coefficients

Multicollinearity

Multicollinearity...

Remedial measures

Need to improve model based on diagnostics
Error distribution: Transform data?
Unequal variances: Weighting or transformation, possibly variance function of x-variable(s)?
Outliers in y, or non-normality: Robust regression?
Nonlinear mean response: Loess or other smoothers (later) or polynomial?
Non-independence: Use variance-covariance matrix?
May need to abandon LS and go to ML

Correlated data

If $\mathbf{Y}$ are not independent, $\mathbf{\Sigma_Y} \neq \sigma^2\mathbf{I}$ then adjustments are needed.
The special case when the correlations are known is of particular significance. This case can be solved and is also quite common. Assume therefore that the variance-covariance matrix can be written in the form $\mathbf{\Sigma_Y} = \sigma^2\mathbf{B}$ with $\mathbf{B}>0$.
Use Cholesky factorisation to write $\mathbf{B} = \mathbf{T}'\mathbf{T}$ and $\mathbf{T} = \mathbf{U}^{-1}$.
Now define $\tilde{\mathbf{Y}} = \mathbf{U}'\mathbf{Y}$...

Further reading