Model Adequacy Checking – Part II

Johns Hopkins Engineering

625.461 Statistical Models and Regression

Module 7 – Lecture 7C

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Examination of Studentized Residuals

A logical procedure is to examine the studentized residuals

$$r_i = \frac{e_i}{\sqrt{MS_{\text{Res}}(1-h_{ii})}}, \quad i = 1, 2, \ldots, n$$

$\text{Var}(r_i) = 1$ regardless of the location of $\mathbf{x}_i$ when the form of the model is correct.
Examination of the studentized residuals is generally recommended.

Examination of Studentized Residuals in Simple Regression

In the simple linear regression scenario,

$$r_i = \frac{e_i}{\sqrt{MS_{\text{Res}}\left[1-\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)\right]}}, \quad i=1,2,\ldots,n$$

When x_i is close to the midpoint, the estimated standard deviation of e_i will be large.

Examination of Studentized Residuals in Simple Regression

Conversly, when x_i is near the extreme ends of the range of the x data, the estimated standard deviation of e_i will be small.

When the sample size n is really large, the effect of $(x_i - \overline{x})^2$ will be relatively small, so in big data sets, studentized residuals may not differ dramatically from standardized residuals.

Residuals for Checking Predicted or Fitted Value

3. PRESS Residuals

Examine $y_i - \hat{y}_{(i)}$ where $\hat{y}_{(i)}$ is the fitted value of the *i*th response based on all observations except the *i*th one.

The logic behind this is that if the *i*th observation y_i is really unusual, the regression model based on all observations may be overly influenced by this observation. That is, $\hat{y}_i$ could be very similar to $y_i \Rightarrow e_i$ will be small (hard to detect the outlier).

Residuals for Checking Predicted or Fitted Value

If the ith observation is deleted, then $\hat{y}_{(i)}$ cannot be influenced by that observation, so the resulting residual should be likely to indicate the presence of the outlier.

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

This prediction error calculation is repeated for each observation $i = 1, 2, \ldots, n$. These prediction errors are called PRESS residuals.

$$e_{(i)} = \frac{e_i}{1-h_{ii}}, \quad i = 1, 2, \ldots, n$$

PRESS Residuals

Residuals associated with points for which h_{ii} is large will have large PRESS residuals, which will generally be high influence points.

Generally a large difference between the ordinary residual and the PRESS residual will indicate a point where the model fits the data well, but a model built without that point predicts poorly.

PRESS Residuals

$$\mathrm{Var}[e_{(i)}] = \mathrm{Var}\left[\frac{e_i}{1-h_{ii}}\right] = \frac{1}{(1-h_{ii})^2}[\sigma^2(1-h_{ii})] = \frac{\sigma^2}{1-h_{ii}}$$

The standardized PRESS residual is

$$\frac{e_{(i)}}{\sqrt{\mathrm{Var}[e_{(i)}]}} = \frac{e_i/(1-h_{ii})}{\sqrt{\sigma_i^2(1-h_{ii})}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$$

If we use MS_{Res} to estimate σ^2, then it is just the studentized residual.

R-Student

The standardized PRESS residual is

$$\frac{e_{(i)}}{\sqrt{\mathrm{Var}[e_{(i)}]}} = \frac{e_i/(1-h_{ii})}{\sqrt{\sigma_i^2(1-h_{ii})}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$$

Estimate σ^2 based on a data set with the ith observation removed. That is, use

$$S_{(i)}^2 = \frac{(n-p)MS_{\mathrm{Res}} - e_i^2/(1-h_{ii})}{n-p-1}$$

R-Student

The R-student (externally studentized residual) is given by

$$t_i = \frac{e_i}{\sqrt{S_{(i)}^2 (1 - h_{ii})}}, \quad i = 1, 2, \ldots, n$$

If the ith observation is influential, then $S_{(i)}^2$ can differ significantly from MS_{Res}, and thus the R-student statistic will be more sensitive to this point for detecting as an outlier.

Ex 4.1 (page 135) 9th data point – outlier?

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING