Model Diagnostics and Remedial Measures in MLR

Jeffrey Woo

MSDS, University of Virginia

Outliers in the Predictors

When there are multiple predictors, outliers are more difficult to detect visually because of plotting limitations in multiple dimensions.

- Geometrically, a vector of k predictor values is an outlier **if it is far away from the center** of the predictor values in k-dimensional space.
- A common measure to detect outliers in the predictor space is called **leverage**.
- Observations with large leverages are more "important" in determining the regression equation.

Hat Matrix

The hat matrix is

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}', \tag{1}$$

where $\boldsymbol{X}$ is the design matrix. The diagonal elements of the hat matrix (1) are the leverages h_{ii}, for each observation. The vector of fitted values can be written as

$$\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}.$$

Detecting Outliers in Predictors

Properties of leverages:

- $h_{ii} = \boldsymbol{X_i'} (\boldsymbol{X'X})^{-1} \boldsymbol{X_i}$
- $0 \leq h_{ii} \leq 1$,
- $\sum_{i=1}^{n} h_{ii} = p$, where p is number of parameters.

The leverage of observation i, h_{ii}, is a measure of distance between the predictors of the ith observation and the mean of predictor values for all n.

Rule for outliers in predictors:
$h_{ii} > \frac{2p}{n}$ indicates outlying cases with regard to their predictors.

Residuals

Residuals can be written in vector form as

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$$

Since the variance-covariance matrix of $\boldsymbol{Y}$ is $\sigma^2\boldsymbol{I}$, the variance-covariance matrix of the ordinary residuals is

$$\boldsymbol{\sigma^2\{e\}} = \sigma^2(\boldsymbol{I} - \boldsymbol{H}).$$

Therefore, the variance of e_i is

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii}) \tag{2}$$

where h_{ii} is the ith element on the **main diagonal** of the hat matrix, and the covariance is

$$\sigma\{e_i, e_j\} = -h_{ij}\sigma^2 \text{ for } i \neq j. \tag{3}$$

Properties of Residuals

- (2) implies the variance of the residuals are not exactly constant.
- (2) also implies that observations with high leverage will have smaller residuals, on average.
- (3) implies the residuals have some correlation.

Note: If $n \gg p$, the entries in the hat matrix tends to 0. This means the variance of the residuals tend towards being constant, and the correlation between residuals tend towards 0.

Outliers in the Response

A refinement to make residuals more effective for **detecting outlying responses** is to measure the ith residual when the fitted regression is based on all of the observations except the ith one. **Externally studentized** residuals, denoted by

$$t_i = \frac{e_i}{\sqrt{\text{MSE}_{(i)}(1 - h_{ii})}}.$$

should be used to detect outliers in the response. If the absolute value of t_i is bigger than $t_{1-\alpha/2n;n-1-p}$, obseration i is outlying in the response.

Measures of Influence

	Formula	Influential if
Cook's D, D_i	$\frac{(\hat{\beta}_{(i)}-\hat{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\beta}_{(i)}-\hat{\beta})}{pMS_{res}}$; or $\frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}$	$> F_{0.5,p,n-p}$
$DFBETAS_{j,i}$	$\frac{\hat{\beta}_j-\hat{\beta}_{j(i)}}{\sqrt{S_{(i)}^2 C_{jj}}}$	magnitude $> 2/\sqrt{n}$
$DFFITS_i$	$\frac{\hat{y}_i-\hat{y}_{(i)}}{\sqrt{S_{(i)}^2 h_{ii}}}$; or $(\frac{h_{ii}}{(1-h_{ii})})^{1/2} t_i$	magnitude $> 2\sqrt{p/n}$

What to do with Influential Observations

- Influential observations usually have something interesting about them that make them "stand out" from the other observations.
- Fit the model with and without the influential observations and see how the models answer our questions of interest.
- Occasionally an observation is influential due to an error in the data entry.
- Rarely do I advocate deleting an influential data point. These observations must addressed.