

# Multiple Linear Regression V (Diagnosis)

Sara López-Pintado

*Department of Statistics*

*Rutgers University*

Fall, 2005

Objective: describe methods for checking the adequacy of a regression model

- detecting improper functional forms
- finding outliers and influential observations
- multicollinearity

- Added variable plot

- It consists in considering the marginal role of a predictor variable  $X_k$ , given that the other predictor variables are already in the model.
- Regress the response variable  $Y$  and the predictor variables  $X_k$  under consideration against the other predictor variables in the regression model and the residuals are obtained.
- Plot these residuals against each other.

## - Outliers

- Identifying outlying  $Y$  observations (Studentized deleted residuals)
- Identifying outlying  $X$  observations (Hat matrix, Leverage values)
- Identifying influential cases (DFFITS, Cook's Distance and DFBETAS)

- Identifying outlying  $Y$  observations: use studentized deleted residuals
- Residuals and Semistudentized Residuals

$$e_i = Y_i - \hat{Y}_i$$

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

- Studentized Residual

$$r_i = \frac{e_i}{s\{e_i\}},$$

where

$$s\{e_i\} = \sqrt{MSE(1 - h_{ii})}$$

- Deleted Residual

$$\begin{aligned}d_i &= Y_i - \hat{Y}_{i(i)} \\ &= \frac{e_i}{1 - h_{ii}}\end{aligned}$$

variance of Deleted Residual

$$\begin{aligned}s^2(d_i) &= MSE_{(i)}(1 + X_i'(X'_{(i)}X_{(i)})^{-1}X_i) \\ &= \frac{MSE_{(i)}}{1 - h_{ii}}\end{aligned}$$

where  $X_i$  is the  $X$  observation vector for the  $i$ th case,  $MSE_{(i)}$  is the mean square error when the  $i$ th case is omitted in fitting the regression function, and  $X_{(i)}$  is the  $X$  matrix with the case deleted.

$$\frac{d_i}{s\{d_i\}} \sim t_{n-p-1}$$

- Studentized Deleted Residuals

$$\begin{aligned}t_i &= \frac{d_i}{s\{d_i\}} \\ &= \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}}\end{aligned}$$

Another way of expressing the studentized deleted residual is

$$t_i = e_i \left[ \frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2} \right]^{1/2}$$

Use the studentized deleted residual for testing for outliers (use Bonferroni critical value)

- Identifying outlying  $X$  observations: Hat matrix and Leverage values

Hat matrix  $H = X(X'X)^{-1}X'$  can be useful for identifying outlying  $X$  observations. Let  $h_{ii}$  the diagonal elements of the hat matrix:

$$0 \leq h_{ii} \leq 1, \quad \text{and} \quad \sum_{i=1}^n h_{ii} = p$$

$h_{ii}$  (called leverage) can be considered as a distance between the  $i$ th case and the means of the  $X$  values all  $n$  cases.

A high value of  $h_{ii}$  implies that the  $i$ th observation is important for determining  $\hat{Y}_i$  :

1. The fitted values  $\hat{Y}_i$  is a linear combination of the observed  $Y$  values:

$$\hat{Y} = HY$$

2. The variance of  $e_i$  depends on  $h_{ii}$  :

$$\text{var}(e) = \sigma^2(1 - h_{ii})$$

A leverage value  $h_{ii}$  is considered large if it is twice as large as the mean leverage value ( $\bar{h} = \frac{p}{n}$ )

Use Hat matrix to identify hidden extrapolation:

$$h_{new,new} = X'_{new}(X'X)^{-1}X_{new}$$

- Identifying influential cases: DFFITS, Cook's Distance, and DFBETAS
- DFFITS

$$\begin{aligned}(DFFITS)_i &= \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}} \\ &= t_i \left( \frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}\end{aligned}$$

$\hat{Y}_{i(i)}$  predicted value for the  $i$ th case when the  $i$ th observation is omitted

$MSE_{(i)}$  mean square error when the  $i$ th case is omitted

An observation is influential if  $DFFITS$  exceeds 1 (for small/medium samples) and  $2\sqrt{p/n}$  for large data sets,

- Cook's distance

$$D_i = \frac{\sum_{j=1}^n (\hat{Y}_j - \hat{Y}_j)^2}{pMSE}$$

In matrix notation:

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})'(\hat{Y} - \hat{Y}_{(i)})}{pMSE}$$

An alternative way of expressing  $D_i$  is

$$D_i = \frac{e_i^2}{pMSE} \left[ \frac{h_{ii}}{(1 - h_{ii})^2} \right]$$

- DFBETAS (Influence on the Regression Coefficients)

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}},$$

where  $c_{kk}$  is the  $k$ th diagonal element of  $(X'X)^{-1}$ .

- Multicollinearity Diagnostics (Variance Inflation factor)

Ways of detecting multicollinearity:

1. Large changes in the estimated regression coefficients when a predictor variable is added or deleted.
2. Nonsignificant result on the individual tests on regression coefficients for important predictor variables.
3. Large correlation between pairs of predictor variables
4. Wide confidence intervals for the regression coefficients representing important predictor variables.

- Variance Inflation factor:

(how the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related)

$$\text{var}(b^*) = (\sigma^*)^2 r_{XX}^{-1}$$

$$\text{var}(b_k^*) = (\sigma^*)^2 (VIF)_k$$

$(VIF)_k$  is the variance inflation factor  $(VIF)$  for  $b_k^*$ .

$$(VIF)_k = (1 - R_k^2)^{-1} \quad k = 1, 2, \dots, p - 1$$

where  $R_k^2$  is the coefficient of determination when  $X_k$  is regressed on the  $p - 2$  other  $X$  variables in the model.

A maximum value of  $VIF$  greater than 10 indicates a problem with multicollinearity.