Regression Model Building - Diagnostics

KNNL – Chapter 10

Model Adequacy for Predictors
Added Variable Plot (Partial Regression Plot)

- Graphical way to determine partial relation between response and a given predictor, after controlling for other predictors – shows form of relation between new *X* and *Y*
- May not be helpful when other predictor(s) enter model with polynomial or interaction terms that are not included
- Algorithm (assume plot for X_3, given X_1, X_2):
  - Fit regression of Y on X_1,X_2, obtain residuals = $e_i(Y|X_1,X_2)$
  - Fit regression of X_3 on X_1,X_2, obtain residuals = $e_i(X_3|X_1,X_2)$
  - Plot $e_i(Y|X_1,X_2)$ (vertical axis) versus $e_i(X_3|X_1,X_2)$ (horizontal axis)
- Slope of the regression through the origin of $e_i(Y|X_1,X_2)$ on $e_i(X_3|X_1,X_2)$ is the partial regression coefficient for X_3

Added Variable Plot

- If line is flat horizontal, then the additional variable is not helpful, controlling for the rest of the predictors.
- If the extra variable is important and has a linear relationship with *Y,* then we see a tilted straight-line scatterplot.
- If we see a curve, then we should add the extra variable with a transformation (like polynomial)

Outlying *Y* Observations – Studentized Residuals

Model Errors (unobserved):

$$\varepsilon_i = Y_i - \left(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}\right) \qquad E\left\{\varepsilon_i\right\} = 0 \qquad \sigma^2\left\{\varepsilon_i\right\} = \sigma^2 \qquad \sigma\left\{\varepsilon_i, \varepsilon_j\right\} = 0 \quad \forall i \neq j$$

Residuals (observed) where $h_{ij} = (i, j)^{th}$ element of $\underset{n\times n}{\mathbf{H}} = \mathbf{X(X'X)^{-1}X'}$:

$$e_i = Y_i - \hat{Y}_i = Y_i - \left(b_0 + b_1 X_{i1} + \ldots + b_{p-1} X_{i,p-1}\right)$$

$$E\left\{e_i\right\} = 0 \qquad \sigma^2\left\{e_i\right\} = \sigma^2\left(1 - h_{ii}\right) \qquad \sigma\left\{e_i, e_j\right\} = -h_{ij}\sigma^2 \quad \forall i \neq j$$

$$s^2\left\{e_i\right\} = MSE\left(1 - h_{ii}\right) \qquad s\left\{e_i, e_j\right\} = -h_{ij} MSE \quad \forall i \neq j$$

Semi-Studentized Residual (Residual divided by estimate of σ, trivial to compute):

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

Studentized Residual (Residual divided by its standard error, messier to compute):

$$r_i = \frac{e_i}{\sqrt{MSE\left(1 - h_{ii}\right)}}$$

Outlying *Y* Observations – Studentized Deleted Residuals

Deleted Residual (Observed value minus fitted value when regression is fit on the other n -1 cases):

$$d_i = Y_i - \hat{Y}_{i(i)} \qquad \hat{Y}_{i(i)} = b_{0(i)} + b_{1(i)} X_{i1} + \ldots + b_{p-1(i)} X_{i,p-1}$$

$b_{k(i)} \equiv$ regression coefficient of X_k when case i is deleted

Studentized Deleted Residual (makes use of having predicted i^{th} response from regression based on other n -1 cases):

$$t_i = \frac{d_i}{s\left\{d_i\right\}} \sim t\left(n-p-1\right)$$

$$s^2\left\{d_i\right\} = s^2\left\{\text{pred}_i\right\} = MSE_{(i)}\left[1 + \mathbf{X_i}'\left(\mathbf{X_{(i)}}'\mathbf{X_{(i)}}\right)^{\mathbf{-1}} \mathbf{X_i}\right] \qquad \mathbf{X_i}' = \begin{bmatrix} 1 & X_{i1} & \cdots & X_{i,p-1} \end{bmatrix}$$

Note: $SSE = \left(n-p\right)MSE = \left(n-p-1\right)MSE_{(i)} + \dfrac{e_i^2}{1-h_{ii}}$

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

Computed without re-fitting n regressions

Test for outliers (Bonferroni adjustment): Outlier if $\left|t_i\right| \geq t\left(1-\left(\dfrac{\alpha}{2n}\right), n-p-1\right)$

Outlying *X*-Cases – Hat Matrix Leverage Values

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X'X}\right)^{\mathbf{-1}}\mathbf{X'} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix} \qquad h_{ij} = \mathbf{x_i}'\left(\mathbf{X'X}\right)^{\mathbf{-1}}\mathbf{x_j} \qquad \mathbf{x_i} = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{i,p-1} \end{bmatrix}$$

Notes: $0 \le h_{ii} \le 1$ $\quad \sum_{i=1}^{n} h_{ii} = trace\left(\mathbf{H}\right) = trace\left(\mathbf{X}\left(\mathbf{X'X}\right)^{\mathbf{-1}}\mathbf{X'}\right) = trace\left(\mathbf{X'X}\left(\mathbf{X'X}\right)^{\mathbf{-1}}\right) = trace\left(\mathbf{I_p}\right) = p$

Cases with X-levels close to the “center” of the sampled X-levels will have small leverages. Cases with “extreme” levels have large leverages, and have the potential to “pull” the regression equation toward their observed Y-values. Large leverage values are > 2p/n (2 times larger than the mean)

$$\hat{\mathbf{Y}} = \mathbf{HY} \quad \Rightarrow \quad \hat{Y}_i = \sum_{j=1}^{n} h_{ij}Y_j = \sum_{j=1}^{i-1} h_{ij}Y_j + h_{ii}Y_{ii} + \sum_{j=i+1}^{n} h_{ij}Y_j$$

Leverage values for new observations: $h_{\text{new,new}} = \mathbf{X}_{\text{new}}^{'}\left(\mathbf{X'X}\right)^{\mathbf{-1}}\mathbf{X}_{\text{new}}$

New cases with leverage values larger than those in original dataset are extrapolations

Identifying Influential Cases – Fitted Values

Influential Cases in Terms of Their Own Fitted Values - DFFITS:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} \equiv \text{ \# of standard errors own fitted value is shifted when case included vs excluded}$$

Computational Formula (avoids fitting all deleted models):

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

Problem cases are >1 for small to medium sized datasets, $>2\sqrt{\frac{p}{n}}$ for larger ones

Influential Cases in Terms of All Fitted Values - Cook's Distance:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2}{pMSE} = \frac{\left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(\mathbf{i})} \right)' \left(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(\mathbf{i})} \right)}{pMSE} = \frac{e_i^2}{pMSE} \left[\frac{h_{ii}}{\left(1-h_{ii}\right)^2} \right]$$

Problem cases are $> F\left(0.50; p, n-p\right)$

Influence on the Regression Coefficients

A measure of the influence of the i^{th} case on each regression coefficient b_k is the difference between the estimated regression coefficient b_k based on all n cases and the regression coefficient $b_{k(i)}$ obtained when the i^{th} case is omitted:

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

where

$$(\boldsymbol{X'X})^{-1} = \begin{bmatrix} c_{00} & c_{01} & \cdots & c_{0,p-1} \\ c_{10} & c_{11} & \cdots & c_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p-1,0} & c_{p-1,1} & \cdots & c_{p-1,p-1} \end{bmatrix}$$

Problem cases when:

$$(DFBETAS)_{k(i)} > \frac{2}{\sqrt{n}}$$

Note: Cook's Distance can also be computed as aggregate measure on the the influence on all regression coefficients and is algebraically equivalent to the overall influence on all fitted values.

Multicollinearity - Variance Inflation Factors

- Problems when predictor variables are correlated among themselves
  - Regression Coefficients of predictors change, depending on what other predictors are included
  - Extra Sums of Squares of predictors change, depending on what other predictors are included
  - Standard Errors of Regression Coefficients increase when predictors are highly correlated
  - Individual Regression Coefficients are not significant, although the overall model is
  - Width of Confidence Intervals for Regression Coefficients increases when predictors are highly correlated
  - Point Estimates of Regression Coefficients are wrong sign (+/-)

Multicollinearity – Informal Diagnostics

- Large changes in the estimated regression coefficients when a predictor variable is added or deleted.
- Nonsignificant results in individual t-tests on the regression coefficients.
- Estimated regression coefficients with an algebraic sign that is the opposite of that expected.
- Large coefficients of simple correlation between pairs of predictor variables in the correlation matrix.
- Strong linear associations between pairs of predictors in the scatterplot matrix.

Variance Inflation Factor

Original Units for $X_1,...,X_{p-1},Y$: $\quad \boldsymbol{\sigma}^2\left\{\mathbf{b}\right\} = \sigma^2\left(\mathbf{X'X}\right)^{-1}$

Correlation Transformed Values: $X_{ik}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik}-\overline{X}_k}{s_k}\right) \qquad Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i-\overline{Y}}{s_Y}\right)$

$\boldsymbol{\sigma}^2\left\{\mathbf{b}^*\right\} = \left(\sigma^*\right)^2 \mathbf{r}_{\mathbf{XX}}^{-1} \qquad \sigma^2\left\{b_k^*\right\} = \left(\sigma^*\right)^2 \left(VIF\right)_k$

where: $\left(VIF\right)_k = \frac{1}{1-R_k^2}$

with $R_k^2 \equiv$ Coefficient of Determination for regression of X_k on the other $p-2$ predictors

$R_k^2 = 0 \Rightarrow \left(VIF\right)_k = 1 \qquad 0 < R_k^2 < 1 \Rightarrow \left(VIF\right)_k > 1 \qquad R_k^2 = 1 \Rightarrow \left(VIF\right)_k = \infty$

Multicollinearity is considered problematic wrt least squares estimates if:

$\max\left(\left(VIF\right)_1,...,\left(VIF\right)_{p-1}\right) > 10$ or if $\left(\overline{VIF}\right) = \frac{\sum_{k=1}^{p-1}\left(VIF\right)_k}{p-1}$ is much larger than 1