L26 DFFITS and Cook's D

1. Statistics for identifying outliers

   (1) Residuals and deleted residuals

   The ith residual $R_i = \widehat{e}_i = y_i - \widehat{y}_i \sim N\left(0,\, (1-h_{ii})\sigma^2\right)$.

   The ith deleted residual $\text{PRESS}_i = \widehat{e}_{i(i)} = y_i - \widehat{y}_{i(i)} \overset{*}{=\!=} \frac{\widehat{e}_i}{1-h_{ii}} \sim N\left(0,\, \frac{\sigma^2}{1-h_{ii}}\right)$.

   (2) Studentized residuals and sudentized deleted residuals

   $\widehat{e}_i = y_i - \widehat{y}_i \sim N\left(0,\, (1-h_{ii})\sigma^2\right) \Longrightarrow \frac{\widehat{e}_i}{\sqrt{(1-h_{ii})\sigma^2}} \sim N(0,\, 1^2)$.

   The ith studentized residual $\text{STUDENT}_i = t_i = \frac{\widehat{e}_i}{\sqrt{(1-h_{ii})\, MSE}} \sim t(n-p)$.

   $\widehat{e}_{i(i)} = y_i - \widehat{y}_{i(i)} = \frac{\widehat{e}_i}{1-h_{ii}} \sim N\left(0,\, \frac{\sigma^2}{1-h_{ii}}\right) \Longrightarrow \frac{\widehat{e}_{i(i)}}{\sqrt{\sigma^2/(1-h_{ii})}} = \frac{\widehat{e}_i}{\sqrt{(1-h_{ii})\sigma^2}} \sim N(0,\, 1^2)$.

   The ith studentized deleted residual $\text{RSTUDENT} = t_{(i)} = \frac{\widehat{e}_i}{\sqrt{(1-h_{ii})\, MSE_{(i)}}} \sim t(n-1-p)$.

   Comment: Standardizing $\widehat{e}_i$ and $\widehat{e}_{i(i)}$ lead to the same $\frac{\widehat{e}_i}{\sqrt{(1-h_{ii})\sigma^2}} \sim N(0,\, 1^2)$.

   In studentizing $\widehat{e}_i$, σ^2 is replaced by MSE that resulted in

   $$\text{STUDENT}_i = \frac{\widehat{e}_i}{\sqrt{(1-h_{ii})\, MSE}} \sim t(n-p).$$

   In studentizing $\widehat{e}_{i(i)}$, σ^2 is replaced by $MSE_{(i)}$ that resulted in

   $$\text{RSTUDEHT}_i = \frac{\widehat{e}_i}{\sqrt{(1-h_{ii})\, MSE_{(i)}}} \sim t(n-1-p).$$

2. DFFITS

   (1) Difference of fitted values

   The ith difference of fitted values

   $$\begin{aligned} \widehat{y}_i - \widehat{y}_{i(i)} &= \widehat{y}_i - y_i + y_i - \widehat{y}_{i(i)} = -\widehat{e}_i + \widehat{e}_{i(i)} = -\widehat{e}_i + \tfrac{1}{1-h_{ii}}\widehat{e}_i \\ &= \tfrac{h_{ii}}{1-h_{ii}}\widehat{e}_i \sim \tfrac{h_{ii}}{1-h_{ii}} N(0,\, (1-h_{ii})\sigma^2) = N\left(0,\, \tfrac{h_{ii}^2}{1-h_{ii}}\sigma^2\right). \end{aligned}$$

   So $\widehat{y}_i - \widehat{y}_{i(i)} = \begin{cases} \frac{h_{ii}}{1-h_{ii}}\widehat{e}_i \\ h_{ii}\widehat{e}_{i(i)} \end{cases} = \frac{h_{ii}}{1-h_{ii}}\widehat{e}_i \sim N\left(0,\, \frac{h_{ii}^2}{1-h_{ii}}\sigma^2\right)$

   (2) Two studentized difference of fitted values

   Standardizing $\widehat{y}_i - \widehat{y}_{i(i)}$, $\quad \frac{\widehat{y}_i - \widehat{y}_{i(i)}}{\sqrt{h_{ii}^2\sigma^2/(1-h_{ii})}} = \frac{\widehat{e}_i}{\sqrt{(1-h_{ii})\sigma^2}} \sim N(0,\, 1^2)$.

   Two studentized $\widehat{y}_i - \widehat{y}_{i(i)}$, $\quad \frac{\widehat{e}_i}{\sqrt{(1-h_{ii})\, MSE}} \sim t(n-p)$ and $\frac{\widehat{e}_i}{\sqrt{(1-h_{ii})\, MSE_{(i)}}} \sim t(n-1-p)$ are studentized residual and studentized deleted residual.

   Comment: Standardizing $y_i - \widehat{y}_i$, $y_i - \widehat{y}_{i(i)}$ and $\widehat{y}_i - \widehat{y}_{i(i)}$ lead to the same $\frac{\widehat{e}_i}{\sqrt{(1-h_{ii})\sigma^2}} \sim N(0,\, 1^2)$ with two studentized forms: Studentized residual and Studentized deleted residual.

(3) DFFITS

Define $\text{DFFITS}_i = \frac{\widehat{y}_i - \widehat{y}_{i(i)}}{\sqrt{h_{ii}\, MSE_{(i)}}}$.

Then $\text{DFFITS}_i = \sqrt{\frac{h_{ii}}{MSE_{(i)}}}\, \frac{\widehat{e}_i}{1-h_{ii}} = \sqrt{\frac{h_{ii}}{MSE_{(i)}}}\, \widehat{e}_{i(i)} = \sqrt{\frac{h_{ii}}{1-h_{ii}}}\, t_{(i)}$.

So $\text{DFFITS}_i \sim \sqrt{\frac{h_{ii}}{1-h_{ii}}}\, t(n-1-p)$

By the rule of thumb, if $|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}}$, then the ith observation is an outlier.

Ex1: Express DFFITS_i through $\widehat{e}_i$, $\widehat{e}_{i(i)}$ and $t_{(i)}$.

$\text{DFFITS}_i = \sqrt{\frac{h_{ii}}{MSE_{(i)}}}\, \frac{\widehat{e}_i}{1-h_{ii}} = \sqrt{\frac{h_{ii}}{MSE_{(i)}}}\, \widehat{e}_{i(i)} = \sqrt{\frac{h_{ii}}{1-h_{ii}}}\, t_{(i)}$.

3. Cook's D

(1) Squared norm of difference of fitted vectors
$\|\widehat{y} - \widehat{y}_{(i)}\|^2 = \sum_{j=1}^{n} \left(\widehat{y}_j - \widehat{y}_{j(i)}\right)^2$.

(2) Cook's D
Define $D_i = \frac{\|\widehat{y} - \widehat{y}_{(i)}\|^2}{p\, MSE}$

(3) Appication
Large D_i indicates that the ith observation is an outlier.
By the rule of thumb, the cut-off value for D_i is $\frac{4}{n}$.

Ex : SAS

```
proc reg;
   model y=x1 x2;
   output out=b P=P R=R H=H STUDENT=STUDENT
          PRESS=PRESS RSTUDENT=RSTUDENT
          DFFITS=DFFITS COOKD=COOKD;
   run;
proc print;
   run;
```

L27 Multicollinearity

1. Multicollinearity

   (1) An assumption
   Based on sample $y \sim N(X\beta, \sigma^2 I_n)$ from a regression model, β can be estimated by its MLE and LSE $\widehat{\beta} = (X'X)^{-1}X'y \sim N\left(\beta, \sigma^2(X'X)^{-1}\right)$. To get this estimator there is an underlying assumption: $X'X$ is non-singular. This assumption has many different but equivalent forms. For example
   (i) The columns of $X \in R^p$ are linearly independent;
   (ii) No column of X is a linear combination of others;
   (iii) $|X'X| > 0$.

   (2) Multicollinearity

   $$\begin{array}{rcl} \text{There is multicollinearity in } X & \overset{def}{\Longleftrightarrow} & \text{The columns of } X \text{ are almost linearly dependent} \\ & \Longleftrightarrow & \text{One column of } X \text{ is almost a LC of others} \\ & \Longleftrightarrow & |X'X| > 0 \text{ is almost 0.} \end{array}$$

   Let $X'X = Q\Lambda Q'$ be the eigenvalue decomposition of $X'X$ where the columns of Q are eigenvectors of $X'X$, Q is an orthogonal matrix ($Q' = Q^{-1}$); $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$, $\lambda_i > 0$, $i = 1, ..., p$, are eigenvalues of $X'X$. Because $|X'X| = \lambda_1 \cdots \lambda_p$, the multicollinearity is reflected as the values of $\lambda_i > 0$ close to 0.

   (3) Consequence of multicollinearity
   If there is multicollinearity in X, then the total variance in $\widehat{\beta} \sim N(0, \sigma^2(X'X)^{-1})$,

   $$\begin{array}{rcl} \sum_{i=1}^p \text{var}(\widehat{\beta}_i) & = & \text{tr}\left(\text{Cov}(\widehat{\beta})\right) = \text{tr}[\sigma^2(X'X)^{-1}] = \sigma^2\text{tr}[(Q\Lambda Q')^{-1}] \\ & = & \sigma^2\text{tr}(Q\Lambda^{-1}Q') = \sigma^2\text{tr}(\Lambda^{-1}) = \sigma^2\left(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_p}\right) \end{array}$$

   will be vary large. So even though $\widehat{\beta}$ is an unbiased estimator, but it is not stable.

2. Detecting multicollinearity

   (1) Sample correlation coefficient method

   ```
   proc corr;
     var x1 x2 x3;
     run;
   ```

   calculates and presents sample correlation coefficient matrix $\begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix}$. Because $|r_{ij}| \leq 1$ and $|r_{ij}| = 1 \Longleftrightarrow x_i = ax_j + b$, the value of $|r_{ij}|$ close to 1 indicates that x_i and x_j are almost correlated consequently there is multicollinearity in X.

   (2) Variance inflation factor method
   Columns of X are almost linearly dependent if and only if one column, say x_i, is almost a linear combination of others. Thus regression model $x_i = \beta_0 + \sum_{j \neq i} \beta_j x_j + \epsilon$ will produce the coefficient of determination R_i^2 will value close to 1. Consequently $\text{VIF}_i = \frac{1}{1-R_i^2}$ called the variance inflation factor will also have large value.

```
proc reg;
  model y=x1 x2 x3/vif;
  run;
```

calculates and presents VIF_i, $i = 1, 2, 3$. If one $\text{VIF}_i > 2.5$, then we conclude the problem of multicollinearity in X.

3. Ridge regression: A remedy

(1) Ridge regression
With eigenvalue decomposition $X'X = Q\Lambda Q'$, the LSE and MLE of β is

$$\widehat{\beta} = (X'X)^{-1}X'y = (Q\Lambda Q')^{-1}X'y.$$

Small values of $\lambda_i > 0$ in $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$ caused multicollinearity. An naive and intuitive idea is to add $K = \text{diag}(k_1, ..., k_p)$ where $k_i > 0$ for all i to Λ to have a new estimator $\widehat{\beta}(K) = [Q(\Lambda + K)Q']^{-1}X'y$.
The diagonal elements of Λ form the ridge of Λ, the proposed approach elevates the ridge and hence is called a ridge regression.

(2) Compare performance of estimators
Let $\widehat{\eta}$ be an estimator for $\eta \in R^k$. Then $E\|\widehat{\eta} - \eta\|^2$ is the mean squared error. When comparing two estimators, the one with smaller mean squared error is the better one.

$$\begin{aligned} E\|\widehat{\eta} - \eta\|^2 &= E[(\widehat{\eta} - \eta)'(\widehat{\eta} - \eta)] \\ &= E\left\{[\widehat{\eta} - E(\widehat{\eta}) + E(\widehat{\eta}) - \eta]'[\widehat{\eta} - E(\widehat{\eta}) + E(\widehat{\eta}) - \eta]'\right\} \\ &= E\left\{[\widehat{\eta} - E(\widehat{\eta})]'[\widehat{\eta} - E(\widehat{\eta})]\right\} + \|E(\widehat{\eta}) - \beta\|^2 \\ &= \textstyle\sum_{i=1}^k \text{var}(\widehat{\eta}_i) + \|E(\widehat{\eta}) - \beta\|^2 = \text{Toal variance in } \widehat{\eta} + (\text{Bias})^2. \end{aligned}$$

(3) Performance of $\widehat{\beta}(K)$

(i) Ridge estimator is a biased estimator

$$\begin{aligned} E[\widehat{\beta}(K)] &= E\left\{[Q(\Lambda + K)Q']^{-1}X'y\right\} \\ &\neq E[(Q\Lambda Q')^{-1}X'y] = E[(X'X)^{-1}X'y] = (X'X)^{-1}X'X\beta = \beta. \end{aligned}$$

(ii) Ridge estimator reduces the total variance

$$\begin{aligned} \textstyle\sum_i \text{var}[\widehat{\beta}_i(K)] &= \text{tr}[\text{Cov}(\widehat{\beta}(K))] \\ &= \text{tr}\{\sigma^2[Q(\Lambda + K)^{-1}Q']^{-1}X'X[Q(\Lambda + K)^{-1}Q']\} \\ &= \sigma^2\text{tr}[(\Lambda + K)^{-1}\Lambda(\Lambda + K)^{-1}] = \sigma^2 \textstyle\sum_i \frac{\lambda_i}{(\lambda_i + k_i)^2} \\ &< \sigma^2 \textstyle\sum_i \frac{1}{\lambda_i} = \sum_i \text{var}(\widehat{\beta}_i) \end{aligned}$$

(iii) Optimization
It has been shown that there are k_i, $k = 1, ..., p$, such the mean squared error of ridge estimator is less than that of MLE of β. However these k_i depend on parameters and must be estimated.