Model diagnostics and remedies. III

Miaoyan Wang

Department of Statistics
UW Madison

Outliers & Influence

- Some residuals may be much larger than others which can affect the overall fit of the model.
- This may be evident of an outlier: a point where the model has very poor fit.
- This can be caused by many factors and such points should not be automatically deleted from the dataset.
- An observation is called influential if its deletion leads to major changes in the fitted regression.
- Not all outliers are influential.

Dropping an observation

- $A_{\cdot(i)}$ indicates i-th observation was not used in fitting the model
- For example: $\hat{Y}_{j(i)}$ is the regression function evaluated at the j-th observations predictors BUT the coefficients $(\hat{\beta}_{0,(i)}, \dots, \hat{\beta}_{p-1,(i)})$ were fit after deleting i-th row of data.
- Basic idea: if $\hat{Y}_{j(i)}$ is very different than $\hat{Y}_j$ (using all the data), then i is an influential point for determining $\hat{Y}_j$.

Different residuals

- Ordinary residuals: $e_i = Y_i - \hat{Y}_i$ or $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ where $H = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the hat matrix.
- The **standardized residuals (a.k.a. internally studentized residuals)** are defined as

$$r_i = \frac{e_i}{\hat{SD}(e_i)} = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}},$$

  for $\hat{\sigma}^2 = \text{MSE}$ and $i = 1, \ldots, n$. (`rstandard`)
- The **studentized residuals (a.k.a. externally studentized)** are defined as

$$t_i = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1 - h_{ii})}},$$

  where $\hat{\sigma}^2_{(i)}$ is MSE with the ith observation deleted for $i = 1, \ldots, n$. (`rstudent`)

Crud outlier detection test

- If the studentized residuals are large: observation may be an outlier.
- Problem: if n is large, if we "threshold" at $t_{1-\alpha/2,n-p-1}$ we will get many outliers "by chance" even if model is correct.
- Solution: Bonferroni correction, threshold at $t_{1-\alpha/(2n),n-p-1}$.

Bonferroni correction for multiple comparison

If we are doing many T (or other) tests, say the number of tests $m > 1$, we can control overall false positive rate at α by testing each one at level α/m.

- Proof:

$$
\begin{aligned}
&\mathbb{P}(\text{at least one false positive}) \\
&= \mathbb{P}(\cup_{i=1}^{m} |T_j| \geq t_{1-\alpha/(2m),n-p-1}) \\
&\leq \sum_{i=1}^{m} \mathbb{P}(|T_j| \geq t_{1-\alpha/(2m),n-p-1}) \\
&= \sum_{i=1}^{m} \frac{\alpha}{m} = \alpha.
\end{aligned}
$$

Identifying Outlying Y Observations

- How to identify outlying observations in Y?
- Outliers may involve large residuals and often have large impact on the model fit.
- Main idea: The ith observation is an outlier in Y if t_i is large.
- Under H_0 : If observation i is not an outlier in Y, then the studentized residual

$$t_i = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1 - h_{ii})}} \sim t_{n-p-1}.$$

- The decision rule is to reject H_0 if $|t_i^*| > t_{n-p-1, 1-\frac{\alpha}{2n}}$. by the Bonferroni adjustment for n multiple comparisons.
- For most n and p, $t_{n-p-1,1-\frac{\alpha}{2n}}$ at the $\alpha = 5\%$ level is greater than 3. Thus a rule of thumb is that if $|t_i^*| > 3$, investigate observation i as possible outlier.

Identifying Outlying X Observations

- How to identify outlying observations in X?
- Main idea: The ith observation is an outlier in X if h_{ii} is large.
- The leverage h_{ii} is a measure of the distance between $\boldsymbol{X}_i$ and the means of the $\{\boldsymbol{X}_i\}_{i=1}^n$.
- If the ith is an outlier in X with a high leverage h_{ii}, it can influence the fitted response $\hat{Y}_i$.
- Rule of thumb: If $h_{ii} > 2p/n$, then observation i is considered to be an outlier in X.

Identifying Influential Observations

- General strategy to measure influence: for each observation, drop it from the model and measure "how much does the model changes"?
- Consider 3 measures:
  1. DFFITS
  2. Cook's distance
  3. DFBETAS
- No diagnostics identify all possible problems.
  For example, leave-one-out methods do not address multiple influential observations.
- Use common sense.
  Delete suspected influential observations and refit to determine whether there is a large change in the model fit.

DFFITS

- DFFITS measures the influence of the i-th observation on the fitted value $\hat{Y}_i$.

Definition of DFFITS

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}, \quad \text{where} \quad i = 1, \ldots, n.$$

- This quantity measures how much the regression function changes at the i-th observation when the i-th varible is deleted.
- For small/medium datasets: value of 1 or greater is considered "suspicious".
- For large dataset: value of $2\sqrt{p/n}$.

Cook's Distance

- Cook's distance measures the influence of the i-th observation on all n fitted values.

Definition of Cook's distance

$$D_i = \frac{\sum_{i'=1}^{n}(\hat{Y}_{i'} - \hat{Y}_{i'(i)})^2}{p\hat{\sigma}^2}, \quad \text{where} \quad i = 1, \ldots, n.$$

- This quantity measures how much the entire regression function changes when the i-th variable is deleted.
- It is useful to compare D_i against $F_{p,n-p}$. If the percentile is near or more than 50%, then the ith observation may be influential.
- A general rule of thumb: If $D_i > 1$, investigate the ith observation as possibly influential.

DFBETAS

- DFBETAS measures the influence of the i-th observation on the fit of the regression coefficient β_k.

Definition of DFBETAS

$$\text{DFBETAS}_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\hat{\sigma}^2_{(i)}(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}}$$

$$\text{where} \quad i = 1, \ldots, n, \quad j = 0, \ldots, p-1.$$

- This quantity measures how much the coefficients change when the i-th variable is deleted.
- For small/medium datasets: absolute value of 1 or greater is "suspicious".
- For large dataset: value of $2/\sqrt{n}$.