STAT 420: Methods of Applied Statistics

Model Diagnostics — Outliers

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course website: https://sites.google.com/site/teazrq/teaching/STAT420

Department of Statistics
University of Illinois at Urbana-Champaign
March 25, 2017

Detecting Outliers

- Previously we learned how to check the normality and constant variance assumptions.
- Sometimes there are observations that are distant from others, and we want to detect and remove them from the data.
- Usually these are observations that are “far away” from the regression line, and have large influence on the estimated $\widehat{\beta}$.
- We will introduce some statistical methods for detecting them.

An important notation change

- In previous lecture notes, we use p to denote the number of predictors, and $p+1$ is the number of parameters in a linear model if an intercept term is used.
- From now on, we don't want to discuss the two cases (with or without intercept) separately. So, we will use p as the total number of parameters, regardless of whether an intercept term is used or not.
- This is mainly for simplification of the notation.

Cook's Distance as a Measure of Influence

- Cook's distance, D_i is a measure of the distance between predicted values when all of the observations are used versus when a given observation i is omitted.
- Let $\widehat{\boldsymbol{\beta}}_{(-i)}$ denote the OLS estimate when the i^{th} observations is deleted.
- This new parameter estimate should produce a new set of predicted values:

$$\widehat{\mathbf{y}}_{(-i)} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{(-i)}$$

- This should be different from the original predicted values:

$$\widehat{\mathbf{y}}_{(-i)} - \widehat{\mathbf{y}} = \mathbf{X}\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}}\right)$$

- Hence, the squared distance is

$$\left(\widehat{\mathbf{y}}_{(-i)} - \widehat{\mathbf{y}}\right)^{\mathsf{T}}\left(\widehat{\mathbf{y}}_{(-i)} - \widehat{\mathbf{y}}\right) = \left(\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}}\right)^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}}\right)$$

Cook's Distance as a Measure of Influence

- It is interesting that this squared distance can be simplified into something that does not require refitting of the model:
- The Cook's distance (1977), D_i is

$$
\begin{aligned}
D_i &= \frac{\big(\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}}\big)^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{X} \big(\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}}\big)}{p \widehat{\sigma}^2} \\
&= \frac{e_i^2}{p \widehat{\sigma}^2} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right]
\end{aligned}
$$

  where all components in the second line are from the original model:
  - e_i is the residual of the ith subject
  - h_{ii} is the ith diagonal element in the "hat matrix" $\mathbf{H}$
  - $\widehat{\sigma}^2$ is the estimated variance of residual
- The second equality is due to Beckman and Trussel (1974).

Cook's distance: remove subject

```
> fit = lm(Work ~ FPC, data = ATT)
# Calculate Cook's distance for subject 58
# Refit by removing the subject
> ATTi = ATT[-58,]
> fiti = lm(Work ~ FPC, data = ATTi)
# Calculate Di using the first approach
> X = cbind(1, ATT$FPC)
> diff = X %*% (fit$coefficients - fiti$coefficients)
> sigma2 = sum(fit$residuals^2)/(nrow(ATT)-2)
> t(diff) %*% diff / 2 / sigma2
           [,1]
[1,] 0.05399123
```

Cook's distance

- Example: Calculate the Cook's distance D_i statistic for all subjects in the ATT data, and identify the subject with the largest D_i
- The subject ID is:

A : 58 B : 24 C : 16

Cook's distance using $\mathbf{H}$

- The easier and default way is to use the "hat matrix", because it does not involve refitting the model (computational cost is high).

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

- h_{ii} is also called leverage. Large values of h_{ii} are due to extreme values in $\mathbf{X}$.
- Recall the interpretation of $\mathbf{H}$, and by matrix operations, we have

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j .$$

  Hence a point with high leverage has the potential to greatly influence the fit.
- A "rule of thumb" is that leverages of more than $2\bar{h}_{ii}$ (some suggests $3\bar{h}_{ii}$) should be looked at more closely, where $\bar{h}_{ii} = n^{-1}\sum_i h_{ii}$

More about $\mathbf{H}$

- Some properties about $\mathbf{H}$ should be noted: $0 \leq h_{ii} \leq 1$.
- This is due to the fact that $\mathbf{H}$ is idempotent, i.e., $\mathbf{H}^2 = \mathbf{H}$. Hence $h_{ii} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$
- Hence $h_{ii} \geq 0$ because $\sum_{j=1}^{n} h_{ij}^2 \geq 0$.
- $h_{ii} \leq 1$ can be shown by contradiction: if $h_{ii} > 1$, then $\sum_{j \neq i} h_{ij}^2$ has to be negative.

Cook's distance: use H

```
> H = X %*% solve(t(X) %*% X) %*% t(X)
> h58 = H[58, 58]
> (h58 / (1 - h58)^2)* fit$residuals[58]^2 / 2/ sigma2
        58
0.05399123
```

Cook's distance

- Example: Calculate the Cook's distance D_i statistic for all subjects in the ATT data using the hat matrix approach.
- The subject that has the smallest D_i:

A : 74 B : 57 C : 79

Build-in approaches

```
# leverage of all subjects
> hat(X, intercept =FALSE)
# Di of all subjects
> cooks.distance(fit)

> hat(X, intercept =FALSE)[58]
[1] 0.0113612
> cooks.distance(fit)[58]
        58
0.05399123
> which.max(cooks.distance(fit))
24
24
```

Detecting outliers

- In practice, any observation with $D_i > 0.5$ could be influential. > 1 is quite likely to be influenzal. Or, if it stands out from the other D_i values, it is almost certainly influential.
- Cook's distance D_i is related to the F test statistic with degrees of freedom $(p, n-p)$.
- In practice, one compares D_i with $F_{0.5,\,p,\,n-p}$, and investigate any observation above that threshold.

Difference in fits (DFFITS)

- DFFITS is another practically useful approach.

$$\text{DFFITS}_i = \frac{\widehat{y}_i - \widehat{y}_{(-i)}}{\sqrt{\widehat{\sigma}^2_{(-i)} h_{ii}}}$$

  where $\sigma^2_{(-i)}$ is the MSE estimated by removing subject i.

- "Rule of thumb": if DFFITS_i is greater than $2\sqrt{\frac{p+1}{n-p-1}}$, its worth investigating.

Solutions

- Warning: an observation is influential does NOT mean that it should be removed immediately!!
- Iteratively reweighting the observations may reduce their influence (rlm in MASS package).
- Sometimes we may consider giving up the squared error (ℓ_2 norm) in OLS and use alternative approaches to deal with outliers
  - Use least absolute residuals instead of squared loss (VGAM package)
  - Use quantile regression (quantreg package)
- Sometimes we may also consider giving up the linear functional form and use nonparametric regressions such as random forests (randomForest package), generalized additive model (gam package), kernel method (loess function), etc.