Regression Diagnostics

Rakhi Singh

Indian Institute of Technology Bombay

Yesterday's observations

We were working on surgical unit data: 8 dependent variables and 1 independent variable.

Yesterday's observations

We were working on surgical unit data: 8 dependent variables and 1 independent variable.

We suspected that observation 5 may be an outlier in Y.

Yesterday's observations

We were working on surgical unit data: 8 dependent variables and 1 independent variable.

We suspected that observation 5 may be an outlier in Y.

We also suspected that observation 28 and 38 may be outliers in X.

Yesterday's observations

We were working on surgical unit data: 8 dependent variables and 1 independent variable.

We suspected that observation 5 may be an outlier in Y.

We also suspected that observation 28 and 38 may be outliers in X.

While studying whether any of these 3 "possible" outliers may be influential, we found

(i) on DFFits, observation 5 and 28 seems to be influential

(ii) on DFBetas, observation 5 seems to be influential

(iii) What happens on Cook's distance?

Yesterday's observations

We were working on surgical unit data: 8 dependent variables and 1 independent variable.

We suspected that observation 5 may be an outlier in Y.

We also suspected that observation 28 and 38 may be outliers in X.

While studying whether any of these 3 "possible" outliers may be influential, we found

(i) on DFFits, observation 5 and 28 seems to be influential

(ii) on DFBetas, observation 5 seems to be influential

(iii) What happens on Cook's distance?

We are going to learn 7 new things today; from each of the 7 rows.

Question 1: Judging based on Cook's Distance

Cook's:

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

Rule of thumb:

Calculate the percentile of $F_{(p,n-p)}$. If the percentile value is less than 20th percentile, then no serious departure. If more than 50th percentile, then definitely an outlier.

This is a question specifically for row 6

Question 1: Judging based on Cook's Distance

Cook's:

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

Rule of thumb:

Calculate the percentile of $F_{(p,n-p)}$. If the percentile value is less than 20th percentile, then no serious departure. If more than 50th percentile, then definitely an outlier.

This is a question specifically for row 6

cook= cooks.distance(mod)

Question 1: Judging based on Cook's Distance

Cook's:

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

Rule of thumb:

Calculate the percentile of $F_{(p,n-p)}$. If the percentile value is less than 20th percentile, then no serious departure. If more than 50th percentile, then definitely an outlier.

This is a question specifically for row 6

```
cook= cooks.distance(mod)
round(cook,2)[5]
```

Question 1: Judging based on Cook's Distance

Cook's:

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

Rule of thumb:

Calculate the percentile of $F_{(p,n-p)}$. If the percentile value is less than 20th percentile, then no serious departure. If more than 50th percentile, then definitely an outlier.

This is a question specifically for row 6

```
cook= cooks.distance(mod)
round(cook,2)[5]
```

0.52

Question 1: Judging based on Cook's Distance

Cook's:

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

Rule of thumb:

Calculate the percentile of $F_{(p,n-p)}$. If the percentile value is less than 20th percentile, then no serious departure. If more than 50th percentile, then definitely an outlier.

This is a question specifically for row 6

cook= cooks.distance(mod)

round(cook,2)[5]

0.52

pf(0.52,9,45)= 14 pctl

Question 1: Judging based on Cook's Distance

Cook's:

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

Rule of thumb:

Calculate the percentile of $F_{(p,n-p)}$. If the percentile value is less than 20th percentile, then no serious departure. If more than 50th percentile, then definitely an outlier.

This is a question specifically for row 6

```
cook= cooks.distance(mod)
round(cook,2)[5]
```

0.52

pf(0.52,9,45)= 14 pctl

So, we don't consider even observation 5 as an outlier.

Question 2: How to judge multicollinearity

Multicollinearity is a way to see if the independent variables are correlated with each other.

$$(VIF)_k = \frac{1}{1-R_k^2}; k = 1, \ldots, p-1.$$

Rule of thumb:
$VIF > 10$ serious multicollinearity; $2 < VIF \leq 10$ intermediate

This is a question specifically for row 1

Question 2: How to judge multicollinearity

Multicollinearity is a way to see if the independent variables are correlated with each other.

$$(VIF)_k = \frac{1}{1-R_k^2}; k = 1, \ldots, p-1.$$

Rule of thumb:
$VIF > 10$ serious multicollinearity; $2 < VIF \leq 10$ intermediate

This is a question specifically for row 1

Install package "CAR"
vif(mod)

Question 2: How to judge multicollinearity

Multicollinearity is a way to see if the independent variables are correlated with each other.

$$(VIF)_k = \frac{1}{1-R_k^2}; k = 1, \ldots, p-1.$$

Rule of thumb:
$VIF > 10$ serious multicollinearity; $2 < VIF \leq 10$ intermediate

This is a question specifically for row 1

Install package "CAR"
vif(mod)

x1	x2	x3	x4	x5	x6	x7	x8
2.01	1.311	1.79	3.025	1.134	1.131	1.396	1.453

Question 2: How to judge multicollinearity

Multicollinearity is a way to see if the independent variables are correlated with each other.

$$(VIF)_k = \frac{1}{1-R_k^2}; k = 1, \ldots, p-1.$$

Rule of thumb:
$VIF > 10$ serious multicollinearity; $2 < VIF \leq 10$ intermediate

This is a question specifically for row 1

Install package "CAR"
vif(mod)

x1	x2	x3	x4	x5	x6	x7	x8
2.01	1.311	1.79	3.025	1.134	1.131	1.396	1.453

modx= lm(x2∼ x1+x3+x4+x5+x6+x7+x8)
summary(modx)$r.squared = 0.2370387
1/(1-0.2370387) = 1.310682
No serious multicollinearity.

Question 3: Model Selection using AIC criterion and through forward, backward and stepwise methods

$$(AIC)_p = nlog(SSE_p) - nlog(n) + 2p$$

Rule of thumb:
Smaller the AIC, better the model.

Question 3: Model Selection using AIC criterion and through forward, backward and stepwise methods

$$(AIC)_p = nlog(SSE_p) - nlog(n) + 2p$$

Rule of thumb:
Smaller the AIC, better the model.
Use lny as dependent variable

Question 3: Model Selection using AIC criterion and through forward, backward and stepwise methods

$$(AIC)_p = nlog(SSE_p) - nlog(n) + 2p$$

Rule of thumb:
Smaller the AIC, better the model.
Use lny as dependent variable

This is a question specifically for row 2

Question 3: Model Selection using AIC criterion and through forward, backward and stepwise methods

$$(AIC)_p = nlog(SSE_p) - nlog(n) + 2p$$

Rule of thumb:
Smaller the AIC, better the model.
Use lny as dependent variable

This is a question specifically for row 2

mod= lm(lnY $\sim$ x1+x2+x3+x4+x5+x6+x7+x8)
step(mod, direction="both", k=2)
step(mod, direction= "backward", k=2)
step(mod, direction="forward", k=2)

Question 4: Model Selection using SBC criterion

$$(SBC)_p = nlog(SSE_p) - nlog(n) + log(n)p$$

Rule of thumb:
Smaller the SBC, better the model. Generally, SBC helps in getting more parsimonious models.

Question 4: Model Selection using SBC criterion

$$(SBC)_p = nlog(SSE_p) - nlog(n) + log(n)p$$

Rule of thumb:
Smaller the SBC, better the model. Generally, SBC helps in getting more parsimonious models.
Use lny as dependent variable

Question 4: Model Selection using SBC criterion

$$(SBC)_p = nlog(SSE_p) - nlog(n) + log(n)p$$

Rule of thumb:
Smaller the SBC, better the model. Generally, SBC helps in getting more parsimonious models.
Use lny as dependent variable

This is a question specifically for row 5

Question 4: Model Selection using SBC criterion

$$(SBC)_p = nlog(SSE_p) - nlog(n) + log(n)p$$

Rule of thumb:
Smaller the SBC, better the model. Generally, SBC helps in getting more parsimonious models.
Use lny as dependent variable

This is a question specifically for row 5

mod= lm(lnY ∼ x1+x2+x3+x4+x5+x6+x7+x8)
step(mod, direction="both", k=log(nrow(data)))
step(mod, direction= "backward", k=log(nrow(data)))
step(mod, direction="forward", k=log(nrow(data)))

Question 5: Model Selection using C_p criterion

$$C_p = \frac{SSE_p}{MSE(X_1,\ldots,X_{P-1})} - (n - 2p)$$

Rule of thumb:
The model with smaller C_p and C_p close to p are better.

Question 5: Model Selection using C_p criterion

$$C_p = \frac{SSE_p}{MSE(X_1,\ldots,X_{P-1})} - (n - 2p)$$

Rule of thumb:
The model with smaller C_p and C_p close to p are better.
Use lny as dependent variable

Question 5: Model Selection using C_p criterion

$$C_p = \frac{SSE_p}{MSE(X_1,\ldots,X_{P-1})} - (n - 2p)$$

Rule of thumb:
The model with smaller C_p and C_p close to p are better.
Use lny as dependent variable

This is a question specifically for row 3

Question 5: Model Selection using C_p criterion

$$C_p = \frac{SSE_p}{MSE(X_1,\ldots,X_{P-1})} - (n - 2p)$$

Rule of thumb:

The model with smaller C_p and C_p close to p are better.

Use lny as dependent variable

This is a question specifically for row 3

step(mod, direction="both", scale=(summary(mod)$sigma)^2)

Question 5: Model Selection using C_p criterion

$$C_p = \frac{SSE_p}{MSE(X_1,\ldots,X_{P-1})} - (n - 2p)$$

Rule of thumb:
The model with smaller C_p and C_p close to p are better.
Use lny as dependent variable

This is a question specifically for row 3

```
step(mod, direction=”both”, scale=(summary(mod)$sigma)^2)
library(leaps)
x<-model.matrix(mod)[,-1]
y<-data$lnY
leaps(x,y,nbest=3)
matrix(c(a$size-1,a$Cp, a$which),ncol=10)
```

Question 6: Model Selection using adjusted R-squared, R-squared, PRESS criteria

$$R^2_{a,p} = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SSTO}$$

Rule of thumb:

The model with higher R^2, $R^2_{a,p}$ and lower $PRESS_p$ are better.

Question 6: Model Selection using adjusted R-squared, R-squared, PRESS criteria

$$R^2_{a,p} = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SSTO}$$

Rule of thumb:

The model with higher R^2, $R^2_{a,p}$ and lower $PRESS_p$ are better.

Use lny as dependent variable

Question 6: Model Selection using adjusted R-squared, R-squared, PRESS criteria

$$R^2_{a,p} = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SSTO}$$

Rule of thumb:
The model with higher R^2, $R^2_{a,p}$ and lower $PRESS_p$ are better.
Use lny as dependent variable

This is a question specifically for row 4

Question 6: Model Selection using adjusted R-squared, R-squared, PRESS criteria

$$R^2_{a,p} = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SSTO}$$

Rule of thumb:

The model with higher R^2, $R^2_{a,p}$ and lower $PRESS_p$ are better.

Use lny as dependent variable

This is a question specifically for row 4

```
library(leaps)
x<-model.matrix(mod)[,-1]
leaps(x, lny, method="adjr2",nbest=3)
leaps(x, lny, method="r2",nbest=3)
```

Question 6: Model Selection using adjusted R-squared, R-squared, PRESS criteria

$$R^2_{a,p} = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SSTO}$$

Rule of thumb:

The model with higher R^2, $R^2_{a,p}$ and lower $PRESS_p$ are better.

Use lny as dependent variable

This is a question specifically for row 4

```
library(leaps)
x<-model.matrix(mod)[,-1]
leaps(x, lny, method="adjr2",nbest=3)
leaps(x, lny, method="r2",nbest=3)
mod2= lm(lny~ x1+x2+x3+x8)
h2ii=ls.diag(mod2)$hat
PRESS2=sum((mod2$residual/(1-h2ii))^2) 2.737771
```

Question 7: Model Validation

(i) Predict the observations in validation dataset using the model you built above and see the mean squared prediction errors. Should not differ greatly from MSE_p.

Question 7: Model Validation

(i) Predict the observations in validation dataset using the model you built above and see the mean squared prediction errors. Should not differ greatly from MSE_p.

(ii) Run the model with same variables and look if there are any major differences: sign change, magnitude of SS, etc.

This is a question specifically for row 7

Question 7: Model Validation

(i) Predict the observations in validation dataset using the model you built above and see the mean squared prediction errors. Should not differ greatly from MSE_p.

(ii) Run the model with same variables and look if there are any major differences: sign change, magnitude of SS, etc.

This is a question specifically for row 7

mod= lm(lny $\sim$ x1+x2+X3+X4+x5+x6+x7+x8)
predict(mod, data2)

Question 7: Model Validation

(i) Predict the observations in validation dataset using the model you built above and see the mean squared prediction errors. Should not differ greatly from MSE_p.

(ii) Run the model with same variables and look if there are any major differences: sign change, magnitude of SS, etc.

This is a question specifically for row 7

```
mod= lm(lny ∼ x1+x2+X3+X4+x5+x6+x7+x8)
predict(mod, data2)
data2=read.table(”CH09TA05.txt”, header=T)
modv1= lm(data2$lny ∼
data2$x1+data2$x2+data2$x3+data2$x5+data2$x6
+data2$x8)
summary(modv1)
summary(mod1)
```