STAC67: Regression Analysis

Lecture 20

Sohee Kang

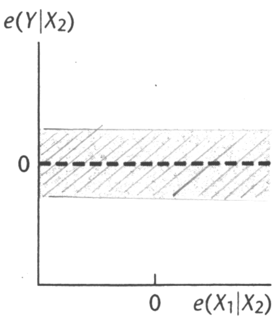
Mar. 24, 2021

Chapter 10

Building the Regession Model II: Diagnositics

10.1 Model Adequacy for Predictors - Added Variable 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
- Algorithm (assume plot for X_1, given X_2): / left over error, part of Y not explained by X_2 (residule)
  - Fit a regression of Y on X_2 $e(Y|X_2)$
  - Fit a regression of X_1 on X_2 $e(X_1|X_2)$ part of X_1 not explained by X_2
  - Plot $e(Y|X_2)$ vs. $e(X_1|X_2)$

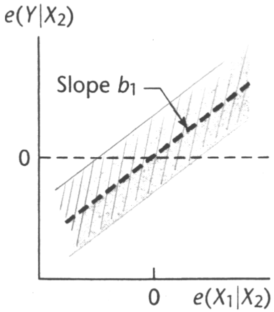


(a)

" X_1 contains no additional information for the prediction of Y beyond whats already provided by X_2 "

don't add X_1

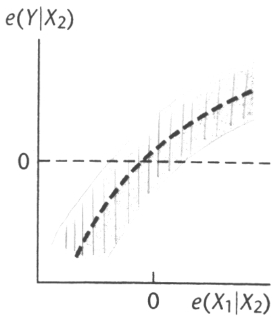
Added Variable Plot



(b)

" A part of X_1 not contained by X_2 is linearly related to Y not already explained by X_2 alone "

add X_1 as linear predictor



(c)

" A part of X_1 not contained by X_2 is curvilinearly related to Y not already explained by X_2 alone "

add X_1 as non-linear predictor

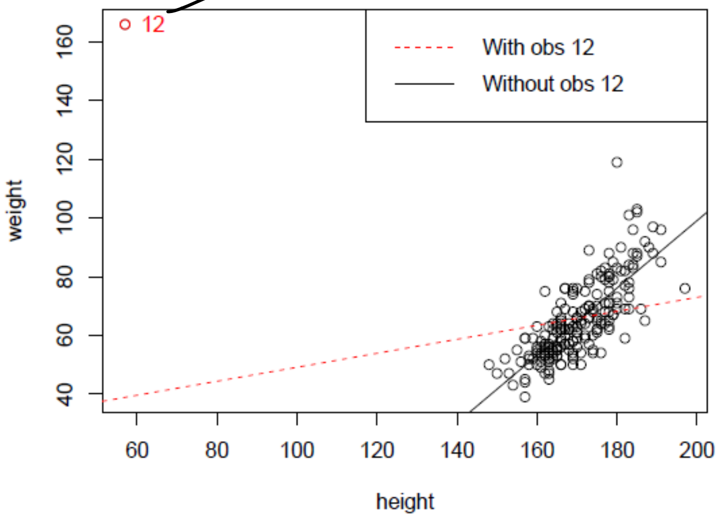
10.2 Outlying and influential observations

- **Outlying observation**: observation well separated from the remainder of the data
- May be outlying with respect to its Y value or X value
- May involve large residuals and have effects on the fitted least squares regression function
- **Influential observation**: observation having effects on the fitted least squares regression function

Outlying and influential observations

- The dataset contains the self-reported height and weight of 200 subjects. The subjects were men and women engaged in regular exercise. The scatterplot below shows the regression lines obtained including/excluding one outlier.

influential observations



Semistudentized residuals

- Detect outlying Y observations using the residuals

$$e_i = Y_i - \hat{Y}_i$$

or the semistudentized residuals

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

The projection matrix

- we defined the projection matrix

$$H = X(X'X)^{-1}X'$$

Studentized residuals

- We know that

$$\hat{\underset{\sim}{Y}} = X\hat{\beta} = HY$$

and

$$e = (I-H)Y$$

- We show that

idempotent

$$Var(\underset{\sim}{e}) = (I-H)' Var(Y)(I-H) = (I-H)\sigma^2$$

- Therefore, we have $\hat{Var}(e_i) = (1-h_{ii})\hat{\sigma}^2$

$$\hat{Cov}(e_i, e_j) = (-h_{ij})\hat{\sigma}^2$$

- studentized residual (residual divied by its standard error)

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

Deleted residuals

- Let $\hat{Y}_{i(i)}$ be the predicted value of the i-th observation when this observation was NOT used to fit the model.
- The deleted residuals are:

$$d_i = Y_i - \hat{Y}_{i(i)}$$

outlier was deleted

- We can show that an algebraically equivalent expression that does not require a recomputation of the fitted regression surface omitting the i-th observation is:

$$d_i = \frac{e_i}{1 - h_{ii}}$$

Studentized deleted residuals

- Using the same idea as for the studentized residuals, we obtain the studentized deleted residuals as:

$$t_i = \frac{di}{SE(d_i)} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

, where $MSE_{(i)}$ is the MSE of model without ith observation

- We can show that an algebraically equivalent expression is:

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

and that $t_i \sim t_{(n-p'-1)}$

T-dist

Test for outlying Y observation

- Outlying observations: observations with large studentized deleted residuals.
- **Test**: The i-th observation is considered an outlying Y observation if

$$|t_i| > t(1-\frac{\alpha}{2n}, n-p'-1)$$, reject

o.w. can not reject

(Bonferroni's Method)

divide by n since we do n tests for residual, we want to control type one error.

10.3 Identifying outlying X observations - Leveage Values

$H = X(X'X)^{-1}X'$

- We have $0 \le h_{ii} \le 1$, $\sum_{i=1}^{n} h_{ii} = p'$

  and h_{ii} is measure of distance between X values of the observation i and the center of the X-space. (the mean of X samples)

- A large h_{ii} indicates that observation i is far away from the center of all X observations.
- In this context, h_{ii} is called the **leverage** of observation i.

Diagnostic for outlying X observations

- **Diagnostic**: a leverage value h_{ii} is usually considered to be large if
  - Guideline 1: $h_{ii} > 2\frac{\sum h_{ii}}{n} = \frac{2p'}{n}$
  - Guideline 2: $h_{ii} > 0.5$

$0.5 = \frac{2p'}{n}$

$0.25n = p'$

10.4 Identifying Influential observations

- Determine whether an outlying Y or X observation is influential, i.e. if its exclusion causes major changes to the fitted regression model.
- Three measures of influence based on the omission of a single observation are:
    1. DFFITS
    2. Cook's Distance
    3. DFBETAS

DFFITS

- The measure of influence of observation i is given by:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

- The DFFITS can be computed using

$$DFFITS_i = t_i \left(\frac{h_{ii}}{1-h_{ii}} \right)^{\frac{1}{2}}$$

- **Guideline**: an observation *i* is influential if

1) if $|DFFITS_i| > 1$ for small or medium sized data

2) if $|DFFITS_i| > 2\sqrt{\frac{p'}{n}}$ for large data set

Cook's distance

- $DFFITS_i$ considers the influence of the i-th observation on the fitted value $\widehat{Y}_i$.
- Cook's distance considers the influence of the i-th observation on all n fitted values.
- Cook's distance is defined as

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p' \, MSE}$$

where $\hat{Y}_{j(i)}$ is the fitted value of $\hat{Y}_j$ with the ith observation deleted

- An equivalent expressin is:

$$D_i = \frac{e_i^2}{p' MSE} \cdot \frac{h_{ii}}{(1-h_{ii})^2}$$

Guideline:

- if D_i is less than 10th or 20th percentile of $F(p', n-p')$, then ith observation has little influence on the fitted val Y_i.
- if D_i is near 50th percentile of $F(p', n-p')$, then ith observation has strong influence on the fitted val Y_i.

DFBETAS

- Measure the influence of an observation on each regression coefficient.
- The influence of an observation i on the coefficient $\hat{\beta}_k$ is defined by

$$DFBETAS_k^{(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE(i)\,(X'X)^{-1}_{k+1,k+1}}}$$

— kth coefficient with ith observation removed

- **Guideline**: observation i is influential if

① if $|DFBETAS_k^{(i)}| > 1$ for small or medium data set

② if $|DFBETAS_k^{(i)}| > 2\sqrt{n}$ for large data set

10.5 Informal diagnostics for multicollinearity

- Indications of the presence of multicollinearity are given by the following informal diagnostics:
  1. Large change in the estimated regression coefficients when a predictor variable is added or deleted
  2. Nonsignificant results in the t-tests on the regression coefficients for important predictor variables
  3. Estimated regression coefficients with an opposite sign of that expected from theoretical considerations or prior experience, for instance
  4. Large standard error of the regression coefficients
  5. Small change in the coefficient of determination R2 when a predictor variable is added or deleted
  6. High correlation between the predictor variables

VIF

$k=3$

VIF_1 $\quad$ VIF_2 $\quad$ VIF_3

$X_1 \sim X_2 + X_3$ $\quad$ $X_2 \sim X_1 + X_3$ $\quad$ $X_3 \sim X_1 + X_2$

- Formal method of detecting the presence of multicollinearity
- The **variance inflation factor** VIF_k for the k-th regression parameter is

$$VIF_k = (1 - R^2_{(k)})^{-1}$$

where $R^2_{(k)}$ is coefficient of determination when X_k regress on the $(k-1)$ other predictors on the model (has nothing to do with Y)

- **Diagnostic**
- A largest VIF value larger than 10 is indicative of serious multicollinearity.
- A mean VIF value $\overline{VIF}$ is considerably bigger than 1.

Surgical unit example

- We selected the following model:

$$Y_i^* = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

where

- $Y^* = lnY$, with Y the survival time,
- X_1: blood clotting score,
- X_2: prognostic inidex
- X_3: enzyme function test score

Outlying observation

Exercise In the surgical unit example, we had observations for 54 patients and we found s = 0.249. Complete the following table. Are there outlying Y observations, and observations with high leverage among these four observations?

Case number	Y_i^*	$\widehat{Y}_i^*$	h_{ii}	e_i	t_i
1	6.544	6.565	0.026	-0.021	
5	7.759	7.270	0.123	0.489	
17	6.526	5.933	0.150	0.593	
38	5.893	6.241	0.290	-0.348	

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

$$tstat = t\left(1 - \frac{0.05}{54}, 49\right)$$

Measures of Influence

Exercise Complete the following table. Are there influential observations among these four observations? Hint: the 20th and 50th percentiles of a F(4,50) are 0.411 and 0.851, respectively.

Case number	Y_i^*	$\widehat{Y}_i^*$	h_{ii}	t_i	$DFFITS_i$	D_i
1	6.544	6.565	0.026	-0.086	-0.014	0
5	7.759	7.270	0.123	2.172	0.815	0.154
17	6.526	5.933	0.150	2.740	1.151	0.293
38	5.893	6.241	0.290	-1.687	-1.079	0.281

By DFFITs 17, 38, influencial

By Cook's, no influencial

Multicollinearity

Exercise

Using R, we computed the VIF and found

k	VIF_k
1	1.031
2	1.008
3	1.023

Is there indication of serious multicollinearity?

No, $\overline{VIF} \approx 1$

R codes

```
Surgic = read.table("Table9-1.txt", header=T)

fit = lm(lnY ~ X1 + X2 +X3, data=Surgic)
# Studentized deleted residuals
t = rstudent(fit)
alpha = 0.05
n = dim(Surgic)[1]
p.prime = length(coef(fit))
t.crit = qt(1-alpha/(2*n), n- p.prime-1)
round(t, 2)
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## -0.09 -0.51 -0.14 -0.75  2.17 -0.66  1.20  1.07  0.81  0.81 -0.99 -1.02 -0.50
##    14    15    16    17    18    19    20    21    22    23    24    25    26
## -0.49  0.50  0.47  2.74 -0.65 -1.14  0.97 -1.33 -0.87  2.34 -1.21 -1.33  1.16
##    27    28    29    30    31    32    33    34    35    36    37    38    39
## -0.28  1.19  1.08  0.37  0.41 -1.74 -0.86  0.54 -0.11 -0.52  0.43 -1.69  0.68
##    40    41    42    43    44    45    46    47    48    49    50    51    52
##  0.55  0.74  1.10 -0.23 -0.45 -2.03  0.47 -0.19  0.57 -0.20  0.23 -0.73  0.81
##    53    54
## -1.03 -1.51
```

```
t.crit
```

```
## [1] 3.526093
```

```
which(abs(t) > t.crit)
```

```
## named integer(0)
```

Leverage

```
# Outlying X observations
hii = hatvalues(fit)
round(hii, 2)
```

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 0.03 0.03 0.05 0.08 0.12 0.06 0.05 0.05 0.03 0.07 0.05 0.08 0.15 0.05 0.05 0.04
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
## 0.15 0.13 0.04 0.03 0.03 0.13 0.12 0.02 0.05 0.02 0.03 0.26 0.05 0.04 0.08 0.21
##   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
## 0.03 0.03 0.02 0.05 0.11 0.29 0.03 0.03 0.04 0.14 0.12 0.03 0.08 0.06 0.07 0.08
##   49   50   51   52   53   54
## 0.02 0.09 0.07 0.11 0.03 0.09
```

```
which(hii > 2*p.prime/n)
```

```
## 13 17 28 32 38
## 13 17 28 32 38
```

```
which(hii > 0.5)
```

```
## named integer(0)
```

Influential observations

```r
# Influence

DFFITTS = dffits(fit)
which(DFFITTS >1)
```

```
## 17
## 17
```

```r
D = cooks.distance(fit)
which(D >qf(0.2, p.prime, n-p.prime))
```

```
## named integer(0)
```

```r
DFBETAS= dfbetas(fit)
head(DFBETAS)
```

```
##    (Intercept)           X1           X2           X3
## 1  0.003261517 -0.007369944  0.001471285 -0.003242347
## 2 -0.062439696  0.034963560  0.015689099  0.042570990
## 3  0.008343962 -0.020971765  0.008740038 -0.008177859
## 4 -0.073456155 -0.016007025 -0.056296165  0.177318922
## 5 -0.577487411  0.489201023  0.003555701  0.635188484
## 6 -0.100939325 -0.009849634  0.139819627  0.023916577
```

```r
which(DFBETAS >1)
```

```
## integer(0)
```

Multicollinearity

```
\begin{verbatim}

> library(car)
> VIF = vif(fit)
> VIF
      X1       X2
1.025951 1.025951
> VIFbar = mean(vif(fit))
> VIFbar
[1] 1.025951

\end{verbatim}
```