Labbrapport i Statistik

Laboration 6

732G46

Mattias Hällgren, Michael Debebe

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2021-10-11

Innehåll

1 Introduktion

I denna laboration kommer ett dataset användas.
I Datasetet som består av ett OSU av 113 sjukhus från 338 sjukhus som har undersökts, kommer variabler såsom; medellängden på besöken, åldern, sannolikheten att få en infektion och medelsumman av antal sjukhussängar m.fl. att analyseras och användas i modeller. Det som skiljer denna laboration från den tidigare laborationen är att samtliga 113 observationer inte kommer att användas samtidigt.

Målet med denna laboration är att se användbarheten i hatmatriser för att hitta avvikande observationer i Y och X samt identifera observationerna som är inflytande på sig själva. Lära sig använda alternativ i R för att detektera avvikande observationer.

2 Databehandling

```
Commercial_properties <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/CH06PR18.t
Commercial_properties <-cbind(Commercial_properties[,1:3],Commercial_properties[,5])
colnames(Commercial_properties) <-c("y","x1","x2","x4")
```

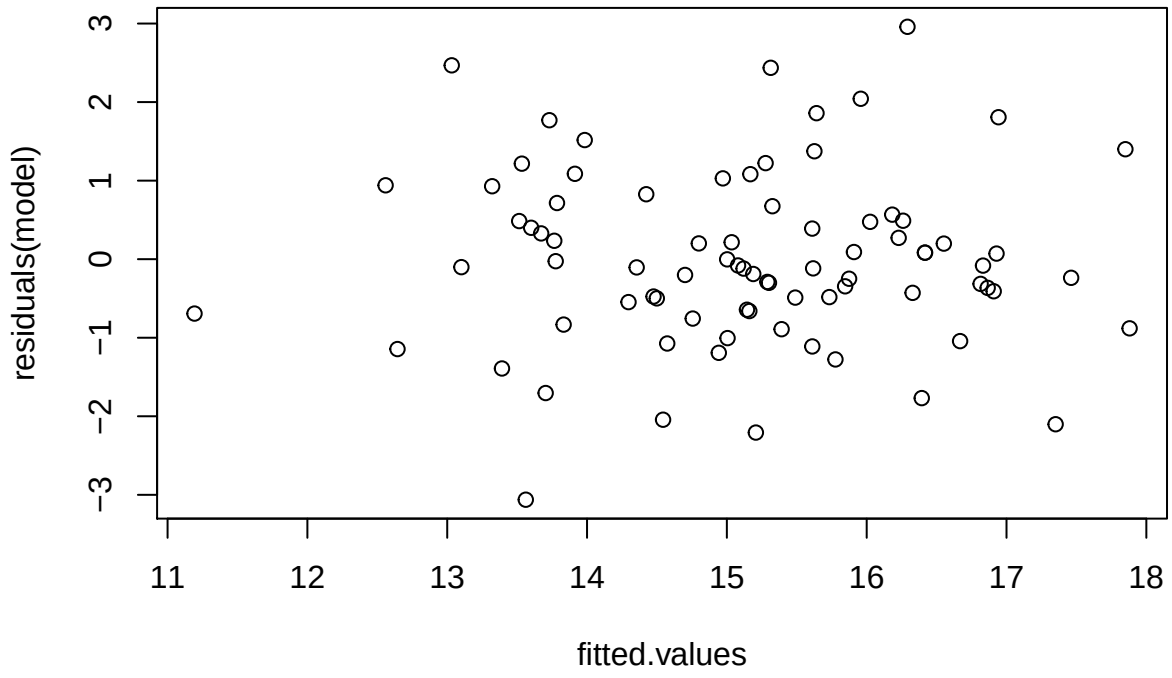
3 Uppgifter

3.1 Uppgift 1 Detecting outliers

```
model <-lm(y~x1+x2+x4,data = Commercial_properties)
```

3.1.1 a) Plot residuals against the fits.

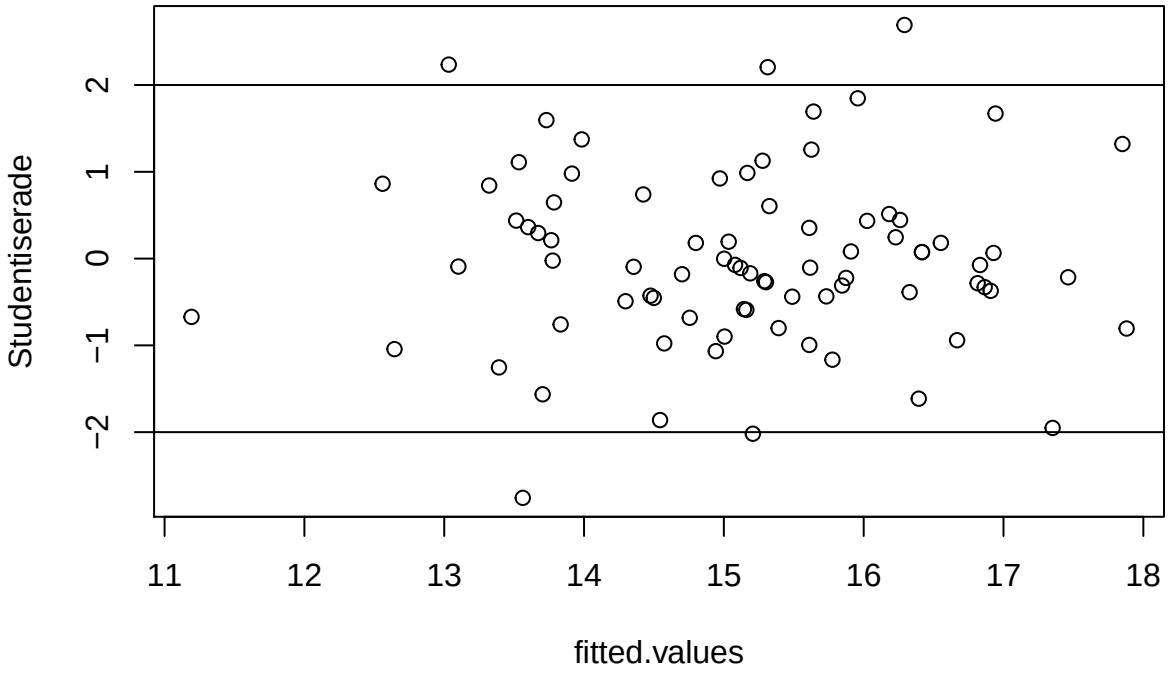
```
residualer <-residuals(model)
fitted.values <-model$fitted.values
plot(y=residuals(model),x=fitted.values)
```



Bland de vanliga residualerna kan vi se att det finns vissa observationer som ser "ensamma" ut, dessa är till exempel den första residualen i den vänstra delen av diagrammet eller den nedersta mellan 13 och 14, spridningen som syns i diagrammet kan klassas som okej, däremot är det svårt att svara på huruvida normalfördelat man kan anta residulaerna vara bara baserat på denna plot.

3.1.2 a) Plot studentized residuals against the fits

```
Studentiserade <-rstandard(model)
plot(y=Studentiserade,x=fitted.values)
abline(h=2)
abline(h=-2)
```

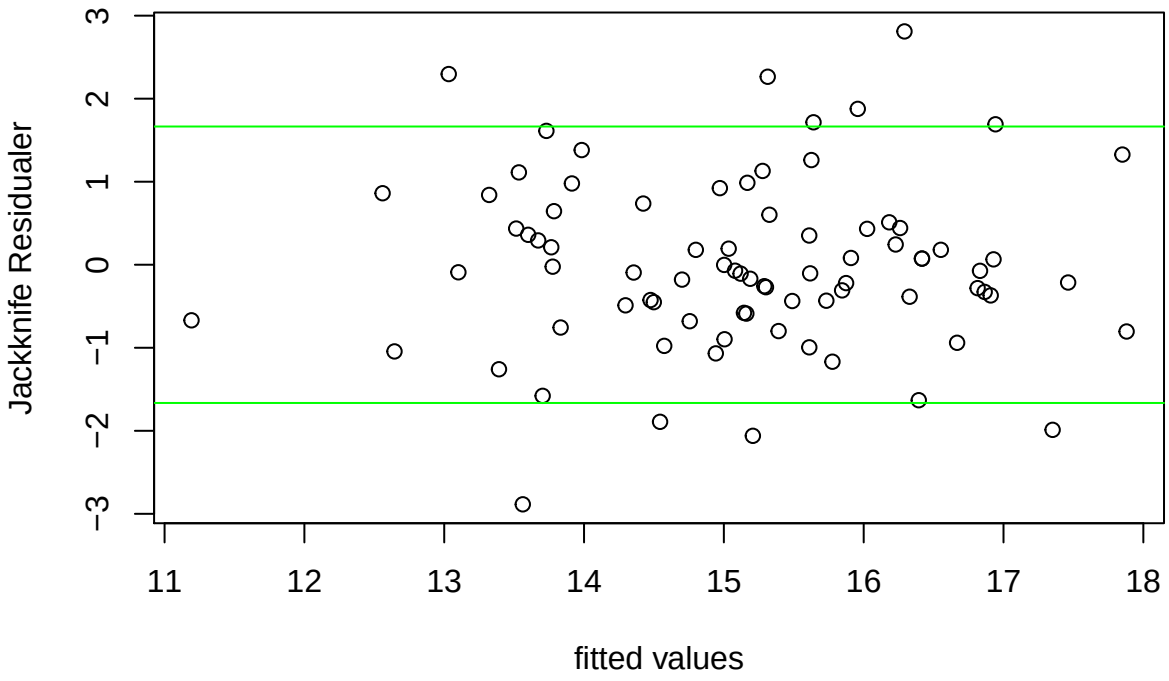


I detta diagram över studentiserade residualer gentemot anpassade värden får returneras ett diagram som är väldigt likt ovanstående, det visas alltså fortfarande ett relativt jämn spritt diagram. Ett fåtal avvikande observationer syns i diagrammet (Observation:6,42,62,63), 3 observationer som överstiger 2 och 1 observationer som ligger under -2, dessa bör underökas vidare.

Hänvisar till FL VECKA 6 om frågetecken kring varför observationer med ett värde högre än $|r_i| > 2$ bör undersökas vidare, väcks.

3.1.3 a) Plot studentized deleted (jackknifed) residuals against the fits.

```
jackknife <-studres(model)
plot(y=jackknife,x=fitted.values,xlab="fitted values",ylab="Jackknife Residualer")
alfa1 <-qt(1-0.10/2,df=77)
alfa2 <-qt(0.10/2,df=77)
abline(h=alfa2,col="green")
abline(h=alfa1,col="green")
```



I plotten visas Jackknife residulaerna gentemot de anpassade värdena, detta är definitivt den mest uppenbara plotten sett till ögat, däremot inte baserat på plotten utan i form av att två linjer på 10 % signifikansnivå lagts till. Sett till ögat så är det inte något som skiljer de tre plottarna åt.

För att uppmärksamma de små skillnaderna som med ögat kan vara svåra att se, visas nedan en tabell över de första 5 residualerna utav varje typ.

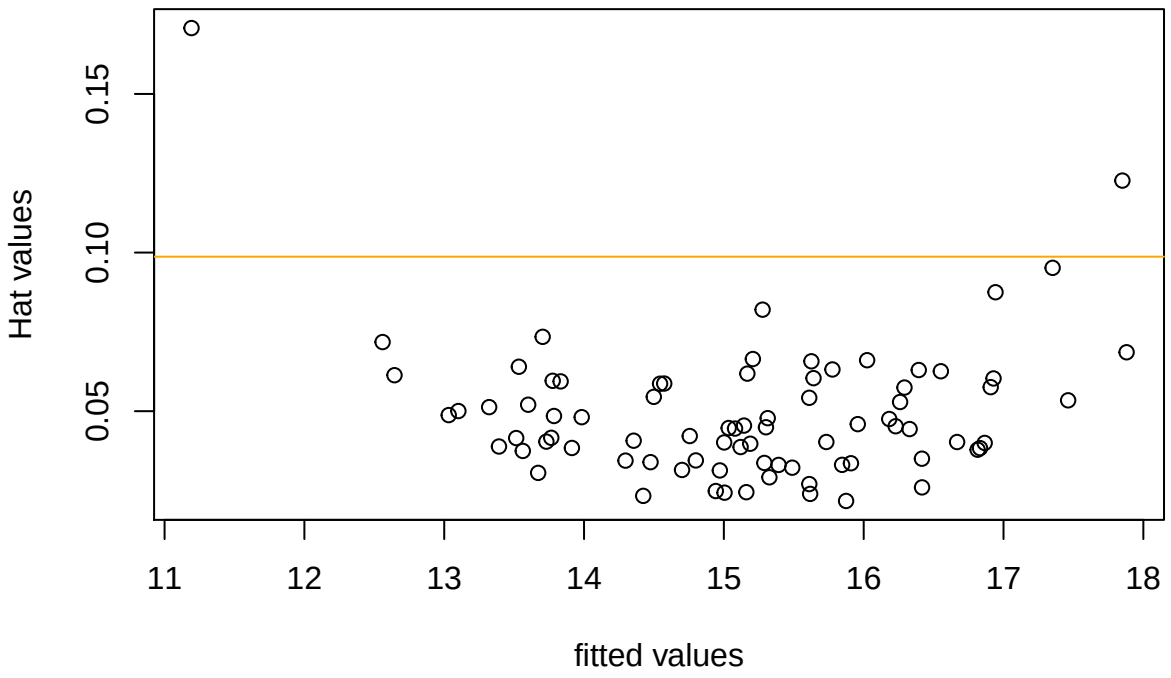
jackknife	residualer	Studentiserade
-0.9772723	-1.0735155	-0.9775576
-1.2587666	-1.3915546	-1.2540161
-0.6695076	-0.6925605	-0.6719194
-0.1071507	-0.1196843	-0.1078452
0.2926010	0.3280485	0.2943540

Precis som i tabellen visas finns det skillnader mellan de olika typerna av residualer, däremot kan dessa med ögat vara svåra att se, sett till hur små de är.

3.1.4 b) Identify any outlyng X observations using leverage values.If you get any outlier, go to the original data and try to interpret why the case is an outlier in X data.

Hattvärdena fås ut genom nedanstående formel

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



x
0.1707732
0.1226868

Formel räkna ut det "kritiska" leverage värdet.

$$h_i > \frac{2\cdot(k+1)}{n} \Rightarrow \frac{2\cdot(3+1)}{81} = 0.0987$$

Alltså är en observation en outlier om den överstiger detta värde.

Från denna utskrift får vi att det är två observationer som överstiger det kritiska värdet 0.0987, dessa observationer är observation 3 och 65. Resterande observationer har inte ett leverage-värde som överstiger det kritiska och har därför inte en inverkan på den

3.1.5 c) Identify whether a new observation with x1= 12, x2= 9.23, x4= 354884 is a substantial extrapolation beyond the range of data by computing its leverage.

Hattvärdet fås ut genom nedanstående formel

$$H = X\left(X'X\right)^{-1}X' = 0.064553$$

```
##     diag(H_0)
## 82 0.06455271
```

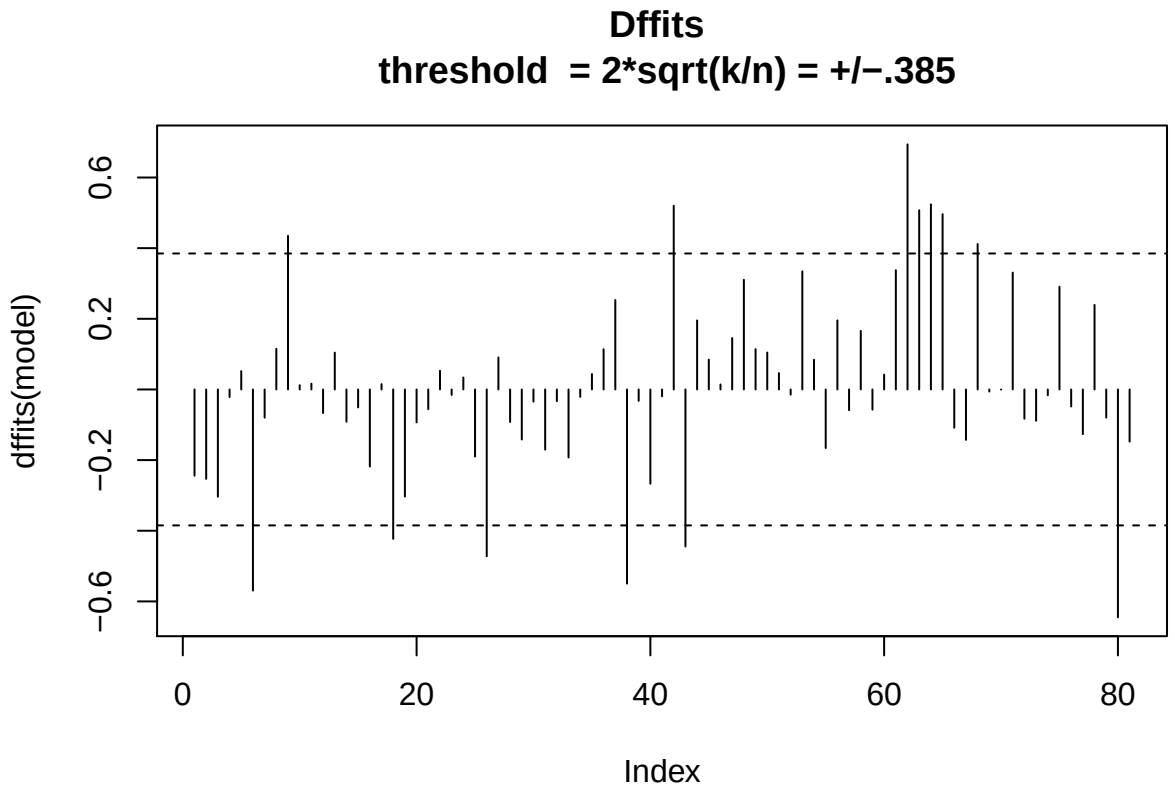
$$h_i > \frac{2 \cdot (k+1)}{n} \Rightarrow \frac{2 \cdot (3+1)}{82} = 0.09756$$

Det hatvärdet som returneras för vår nya observation är 0.064553 vilket är mindre än det kritiska värde på 0.09756, alltså skulle inte en ny observation med dessa värden ligga utanför området. MED ANDRA ORD, inte en substantial extrapolation.

3.2 Uppgift 2 Identifyng influential cases

3.2.1 a) Find the observations which are influential on their own fitted value

```
n <- nrow(Commercial_properties)
k <- length(model$coefficients)-1
cv <- 2*sqrt(k/n)
plot(dffits(model), type = 'h', main ="Dffits \n threshold  = 2*sqrt(k/n) = +/-.385")
abline(h = cv, lty = 2)
abline(h = -cv, lty = 2)
```

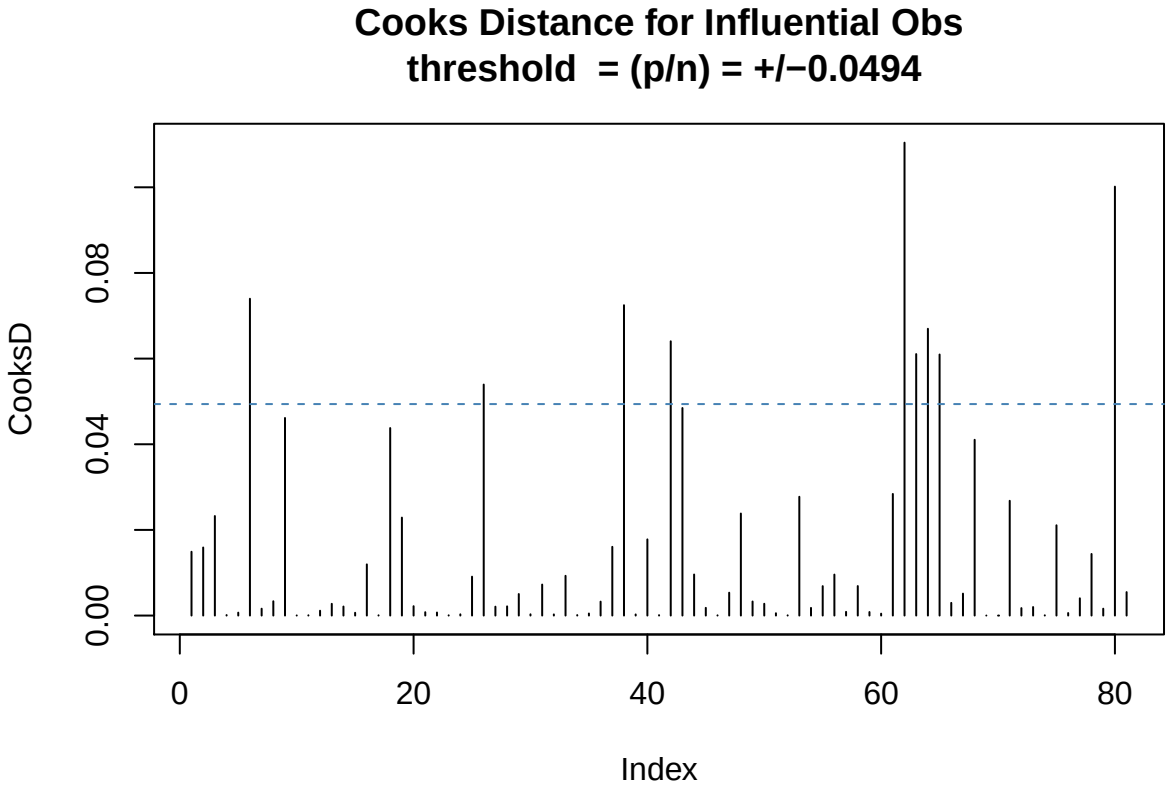


I diagrammet visar x-axeln index av varje observation i datasetet och y-axeln korresponderande Dffits värde. I detta fall är det 4 av de 81 observationerna som har ett inflytande på sitt egna anpassade värde, i detta fall är det en sån liten andel att man inte behöver undersöka de vidare, eventuellt att man skulle undersöka de grövsta observationer vidare.

```
##        62        64        42        63        65         9        68
## 0.6937149 0.5237856 0.5199975 0.5072672 0.4962222 0.4349704 0.4117845
```

3.2.2 b) Investigate if there are any observations that are influential on all fitted values

```
CooksD <- cooks.distance(model)
n <- 81
plot(CooksD, main = "Cooks Distance for Influential Obs\n threshold  = (p/n) = +/-0.0494",type="h")
abline(h = 4/n, lty = 2, col = "steelblue") # add cutoff line
```



Till skillnad från plotten i a) visas här värdena som har inflytande på alla anpassade värden till skillnad tidigare där ploten visar värden som har inflytande på sitt egna anpassade värde. De inflytelserika observationerna ökar markant från mitten av datasetet till slutet.

En annan stor skillnad här, är att dessa värden är dubbelt så många, vilket innebär att man borde undersöka vidare varför detta sker.

	CooksD	dffitsD
62	0.1104205	0.6937149
80	0.1001577	0.5237856
6	0.0739799	0.5199975
38	0.0724726	0.5072672
64	0.0669690	0.4962222
42	0.0640441	0.4349704
63	0.0610578	0.4117845
65	0.0609568	0.3376467
26	0.0539431	0.3343649

I tabellen visas observationerna som har inflytande på alla värden och deras observationsnummer, många av observationerna som har inflytande på sitt egna anpassade värde har också inflytande på samtliga anpassade värden. Värt o nämna är att observation 26 och 65 inte har inflytande på sitt egna anpasssade värde men kom med i tabellen.

3.2.3 Compare the observations found in b) and c) with the observations found in Assignment 1. Conclusions?

Jämförelse mellan de tidigare observationerna i leverage kan man se att DFFITS och Cook´s chart har betydligt mer observationer som är avvikande, samt att 7 av 9 observationer som klassats som avvikande i dffits återfinns i cooks disance, men att de observationer som är avvikande i leverage är också avvikande i de andra modellerna. Gällande Jackknife är det istället observation 42 än 64 som är matchande avvikande med DFFITS och Cook´s chart.

4 Lärdomar

Målet med denna laboration är att se användbarheten i hatmatriser för att hitta avvikande observationer i Y och X samt identifera observationerna som är inflytande på sig själva. Lära sig använda alternativ i R för att detektera avvikande observationer.

Lärdomarna som kom var återigen användningen av matriser inom regression, paketet Olsrr() kom till stor användning för att kunna visualisera och sen jobba därefter.