Cost of Healthcare data
Multiple linear regression modeling

Lecture/Practical 22

25/09/2021

Cost of health care data

- The cost of delivery of health care has become an important concern. These data were collected by Department of Health and Social Services of the state of New Mexico and cover 52 of the 60 licensed facilities in New Mexico in 1998.
- Specific definitions of the variables are given.
- The location of the facility is indicated whether it is the rural or non rural area

Variables in the Cost of health care data

Variable	Definition
RURAL	Rural home (1) and non-rural home (0)
BED	Number of Beds in home
MCDAYS	Annual medical in-patient days(hundreds)
TDAYS	Annual Total Patient Days (Hundreds)
PCREV	Annual Total Patient Care Revenue($100)
NSAL	Annual nursing salaries ($100)
FEXP	Annual Facilities Expenditure ($100)
NETREV	PCREV-NSAL-FEXP

- How do the hospital characteristics affect Patient Care Revenue? Use a multiple linear regression to determine the best fitted model?
- Check the multicollinearity of predictors using VIF criterion. Report the VIF. Does any predictor(s) seems to be highly correlated. If so, remove that predictor(s) one at a time and report.
- Obtain the cooks distance and comment on it.
- Obtain the standardized residuals. Remove any outliers/influential observations one at a time. Remove max 10% of the data if any outliers present.

- Check the assumptions of regression using residual versus fitted plot. Report and interpret it.
- Does all variables significant? If not remove insignificant predictor(s) and rerun the regression.
- Report the final model obtained along with R-square. Interpret the same.

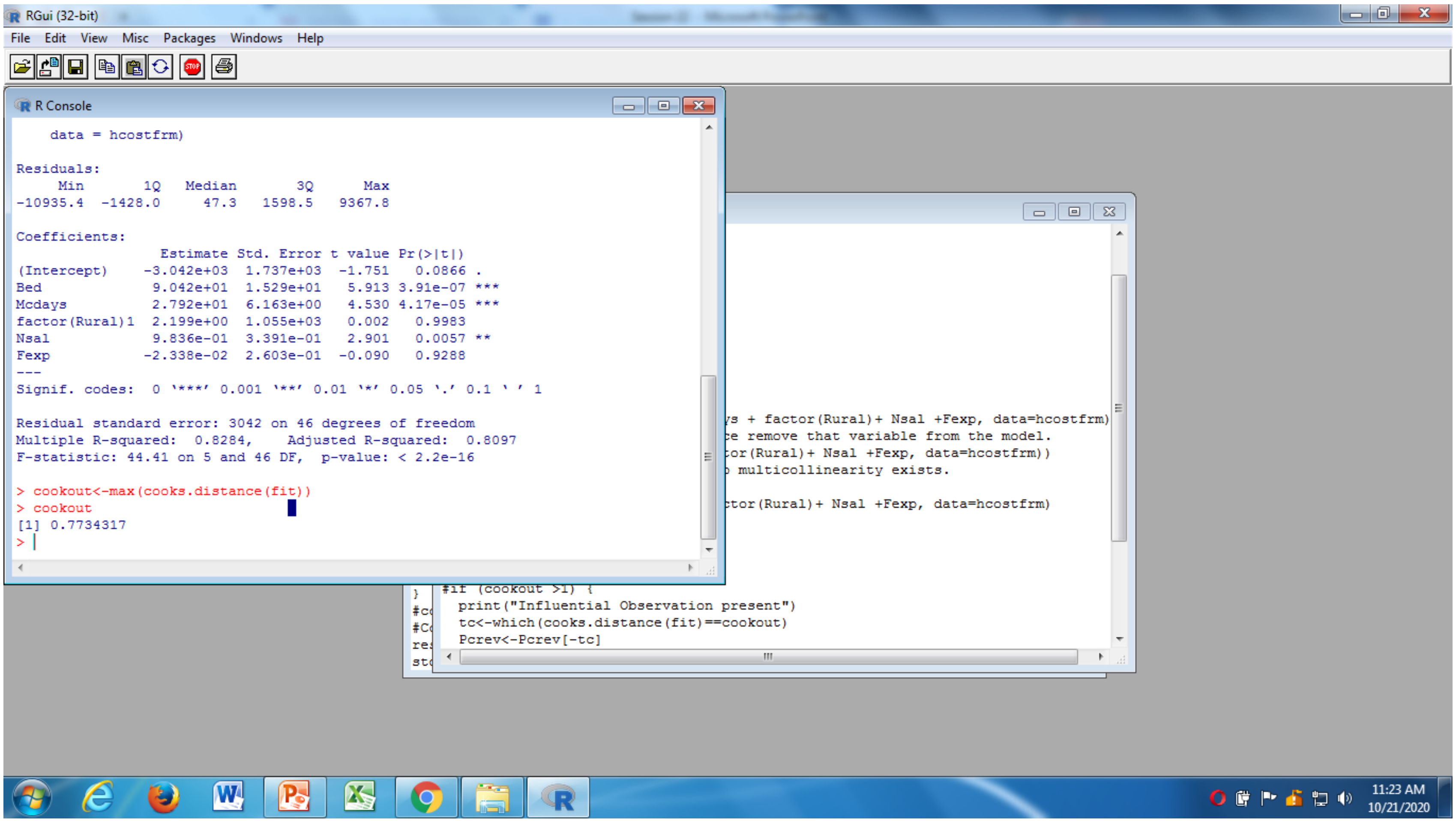
Multiple Linear Regression Modelling

```
# Obtain VIFs
library(car)
vif(lm(Pcrev ~ Bed + Mcdays + Tdays +
factor(Rural)+ Nsal +Fexp, data=hcostfrm))
#Tdays has VIF >4. Remove and run the
regression.
vif(lm(Pcrev ~ Bed + Mcdays + factor(Rural)+
Nsal +Fexp, data=hcostfrm))
```

R-zone

```
fit<-Pcrev ~ Bed + Mcdays + factor(Rural)+ Nsal +Fexp, data=hcostfrm)
summary(fit)
```

Don't look into this output and don't make any interpretation of variables significance etc. at this stage as the model has been fitted only to get the residuals of the fitted model.



Cook’s distance: R-zone

```
cookout<-max(cooks.distance(fit))
cookout
[1] 0.7734317.
```

Interpretation: Cook's distance is not greater than 1. Now proceed to standardized residuals

If in any model, the cook’s distance is more than 1, we can use the following code to remove that leverage (highX) which might have arrived from a particular observation from the data.(which can be located by printing ‘cooks.distance(fit)’.

```
if (cookout >1) {
 print("Influential Observation present")
 tc<-which(cooks.distance(fit)==cookout)
 Pcrev<-Pcrev[-tc]
 Bed<-Bed[-tc]
 Mcdays<-Mcdays[-tc]
Rural<-Rural[-tc]
}
# After getting the cookout value, rerun the model and see the cookout is less than 1.
```

Outliers/Influential observations detection: R-zone

```
res<-residuals(fit) # residuals
stdres<-res/(sd(res))
pout<-max(abs(stdres))
pout

[1] 3.785179

if (pout> 2.5) {
  print("Outlier present")
  t<-which(abs(stdres)==pout)
  Pcrev<-Pcrev[-t]
  Bed<-Bed[-t]
  Mcdays<-Mcdays[-t]
  Rural<-Rural[-t]
  Nsal<-Nsal[-t]
 Fexp<-Fexp[-t]
}
pout

[1] "Outlier present"
```

The above code will tell the program to detect any standardized residual (pout here) exists, and if it exists and is more than 2.5 in absolute value, remove that observation in the data which is causing the pout value here specifically 3.785179

It can be seen that 31st observation is giving us a standardized residual value -3.7851 (Max pout) and hence the R code which we have used might have removed the 31st row (observation) from the data giving us the data dimension as 51 now.

```
> print(stdres)
           1            2            3            4            5            6 
-1.416812884  0.045095066  0.065216746  0.030221230 -0.747984683 -0.306058011 
           7            8            9           10           11           12 
 0.545542943 -0.391090380 -1.054834065 -0.974595560 -0.048737486  1.236712376 
          13           14           15           16           17           18 
 0.054973407 -0.146733117 -0.852627600 -0.291973278 -0.140288192  0.910495774 
          19           20           21           22           23           24 
 0.669722843  0.363265955 -0.628154273 -0.500616390  0.208302056  1.287406912 
          25           26           27           28           29           30 
 0.319151277  1.221200230 -0.364652205  0.002535273 -0.749409367 -1.090168205 
          31           32           33           34           35           36 
-3.785178899  0.576580707 -0.492161208 -0.070890496 -0.424553105  0.665219586 
          37           38           39           40           41           42 
 3.242540936  0.339670968  0.367466144  0.988463601 -0.351746884  0.481411272 
          43           44           45           46           47           48 
-0.167211042  0.707441950  0.434262587 -1.563889882  1.175047865 -0.889039976 
          49           50           51           52 
-0.779469159  1.444671264  0.118579933  0.727677444 
> 
```

Continue this process till you get pout value <2.5.(Remember to remove max 5%(or 10%)of the data gets deleted).

- fit1<-lm(Pcrev ~ Bed + Mcdays + factor(Rural)+ Nsal +Fexp, data=hcost1)
- summary(fit1)

Remember I am not interested to see the output as I am focusing on residuals. Get residuals and accordingly pout for new model (here fit1)

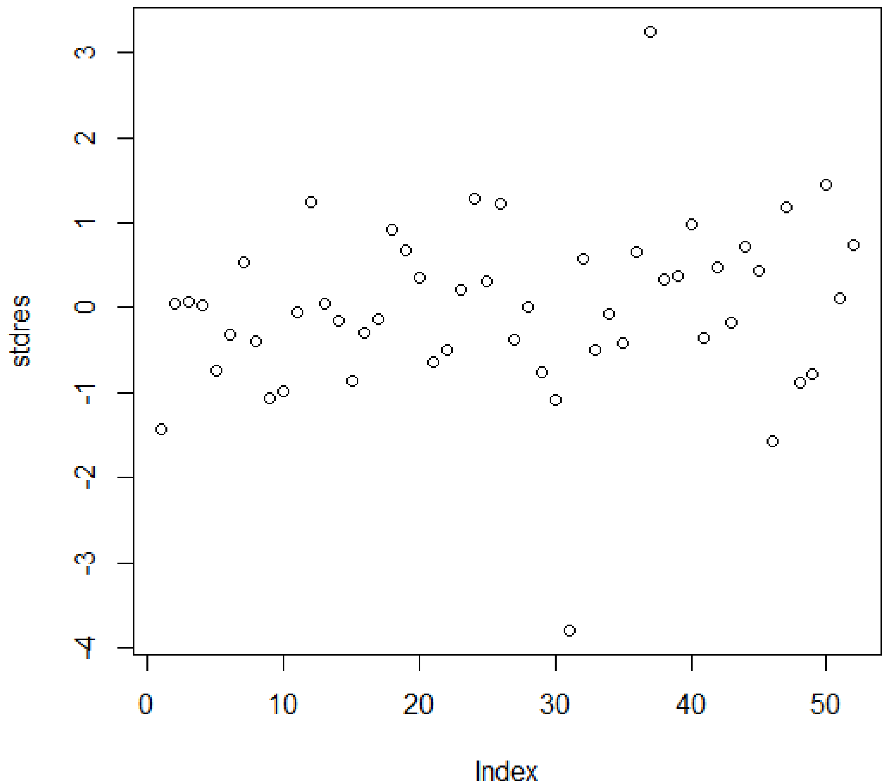
- res1<-residuals(fit1) # residuals
- stdres1<-res1/(sd(res1))
- pout1<-max(abs(stdres1))
- pout1
- [1] 3.600189

Again the model gives a standardized absolute value >2.5. Remove that observation which is resulting into this residual from the data.

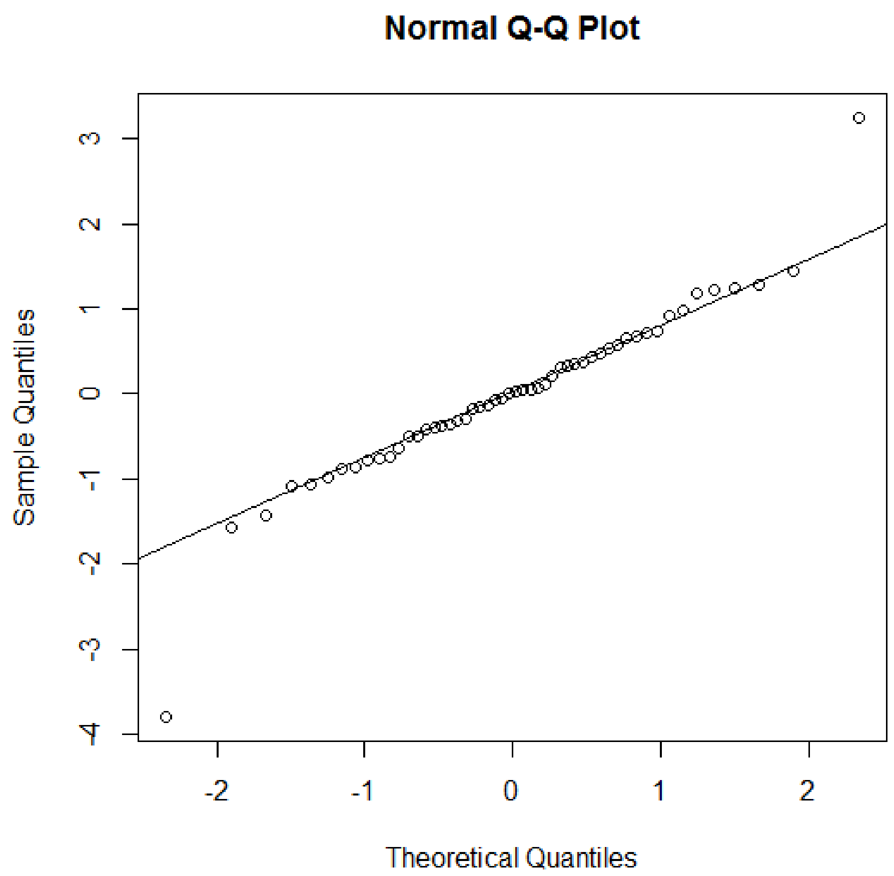
Continuing outlier detection

- fit2<-lm(Pcrev ~ Bed + Mcdays + factor(Rural)+ Nsal +Fexp, data=hcost2)
- summary(fit2)
- res2<-residuals(fit2) # residuals
- stdres2<-res1/(sd(res2))
- pout2<-max(abs(stdres2))
- Pout2
- [1] 2.075211 Now the model gives a standardized residual absolute value <2.5. As I have deleted a total 3 data observations which is causing high influential observations and as the size of the data is small here and there are no more outliers, I am stopping the outlier detection procedure and now focusing on validating the regression assumptions.

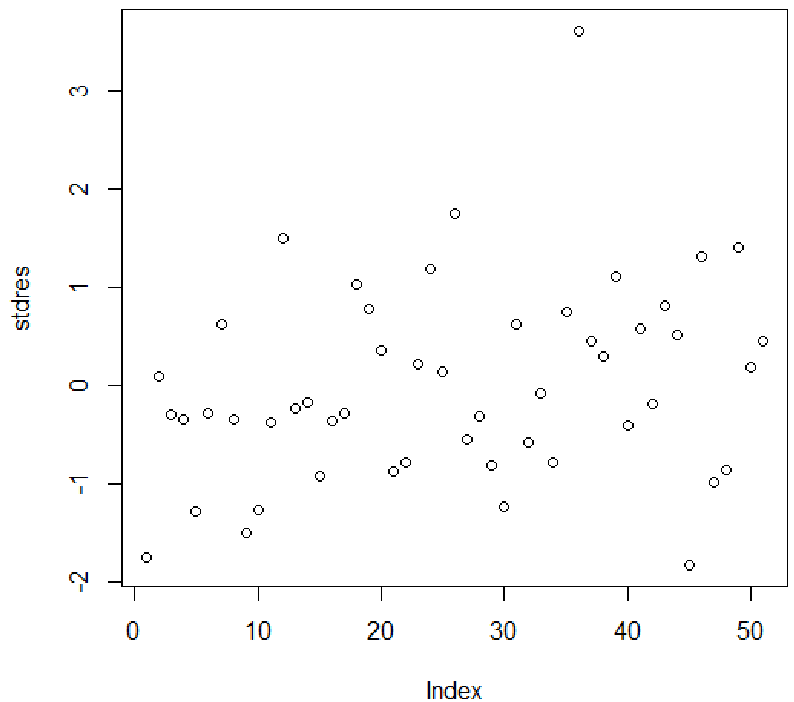
Residual versus fitted plot before outlier deletion

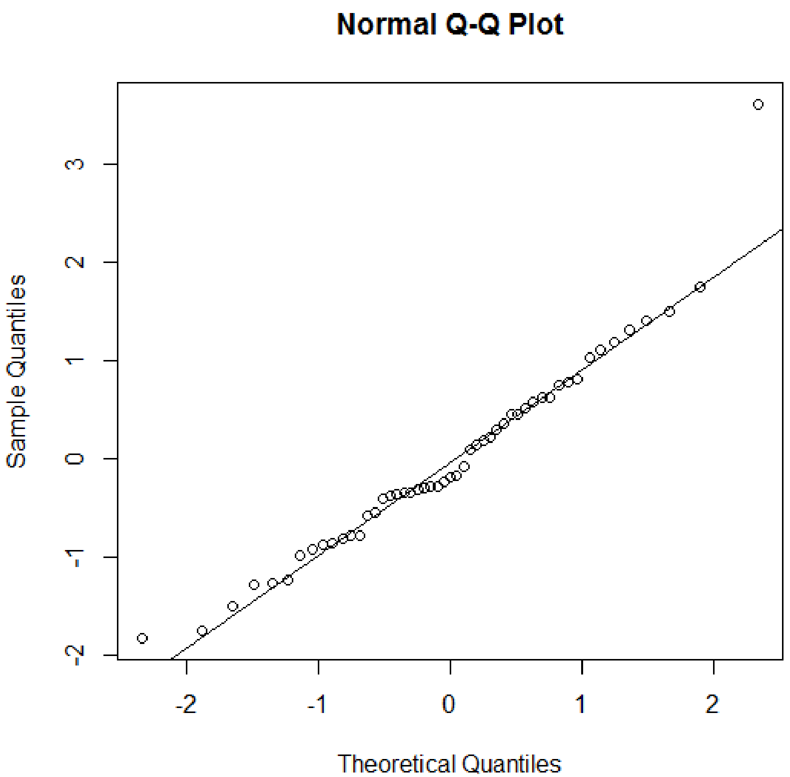


Before outlier deletion normality plot

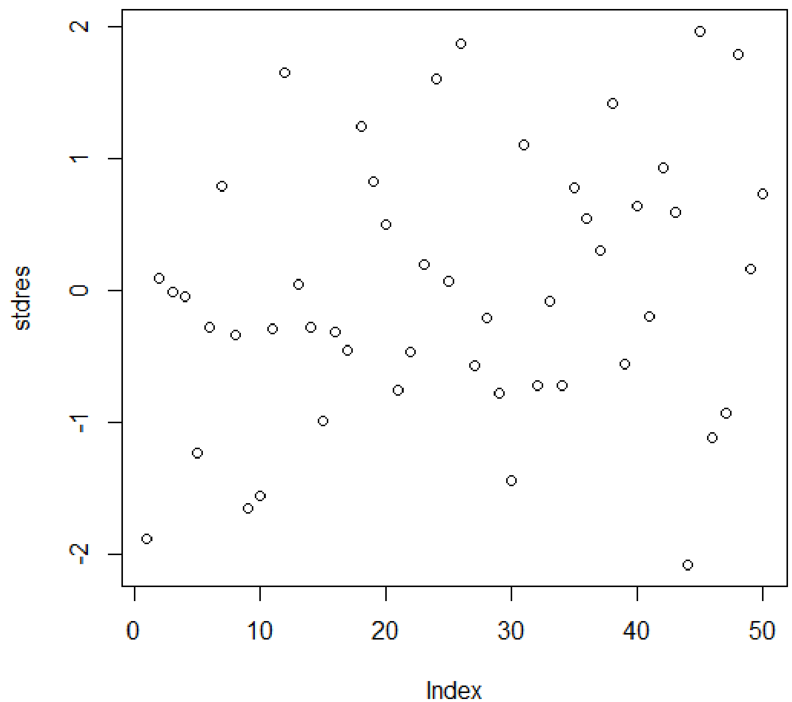


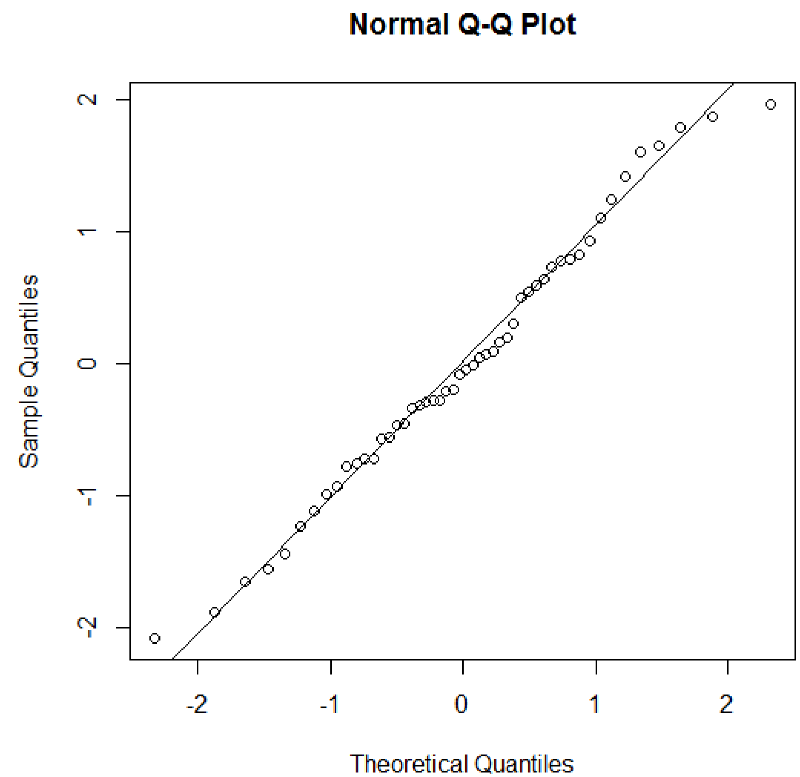
After deleting one outlier





After deleting two outliers





Output after deleting 3 outliers. The overall model is significant. Here we can see that Factor (Rural) and Fexp are insignificant predictors. Remove Rural and rerun the regression

R Console

```
Call:
lm(formula = Pcrev ~ Bed + Mcdays + factor(Rural) + Nsal + Fexp,
    data = hcostfrm)

Residuals:
    Min      1Q  Median      3Q     Max
-4325.4 -1416.4  -123.2  1479.9  4095.7

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2026.9603  1272.3482  -1.593  0.11830
Bed               86.3254    11.1142   7.767 8.73e-10 ***
Mcdays            27.4868     4.4570   6.167 1.92e-07 ***
factor(Rural)1  -414.6699   784.2988  -0.529  0.59966
Nsal               0.8114     0.2475   3.278  0.00205 **
Fexp               0.1208     0.1898   0.636  0.52781
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2200 on 44 degrees of freedom
Multiple R-squared:  0.9001,    Adjusted R-squared:  0.8888
F-statistic:  79.3 on 5 and 44 DF,  p-value: < 2.2e-16
```

Output after removing Rural. Now the variable Fexp is not significant. Remove Fexp and rerun the regression

```
Call:
lm(formula = Pcrev ~ Bed + Mcdays + Nsal + Fexp, data = hcostfrm)

Residuals:
    Min      1Q  Median      3Q     Max 
-4391.0 -1314.2    15.5  1476.6  4086.5 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2484.7547   924.7890  -2.687 0.010071 *  
Bed            87.9867    10.5750   8.320 1.18e-10 ***
Mcdays         26.8907     4.2774   6.287 1.17e-07 ***
Nsal            0.8595     0.2284   3.763 0.000484 ***
Fexp            0.1035     0.1855   0.558 0.579733    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2182 on 45 degrees of freedom
Multiple R-squared:  0.8995,    Adjusted R-squared:  0.8905 
F-statistic: 100.7 on 4 and 45 DF,  p-value: < 2.2e-16
```

Output after removing Fexp. Now all remaining predictors are significant. The removed predictors can be interpreted as in the presence of other predictors in the model, Rural and Fexp is not adding any more information.

```
> fit<-lm(Pcrev ~ Bed + Mcdays + Nsal, data=hcostfrm)
> summary(fit)

Call:
lm(formula = Pcrev ~ Bed + Mcdays + Nsal, data = hcostfrm)

Residuals:
    Min      1Q  Median      3Q     Max 
-4393.7 -1276.3   -96.2  1524.7  3932.5 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2476.2565   917.7141  -2.698 0.009713 ** 
Bed            89.4853    10.1513   8.815 1.92e-11 ***
Mcdays         26.9274     4.2448   6.344 8.83e-08 ***
Nsal            0.8958     0.2173   4.123 0.000155 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2165 on 46 degrees of freedom
Multiple R-squared:  0.8988,    Adjusted R-squared:  0.8922 
F-statistic: 136.2 on 3 and 46 DF,  p-value: < 2.2e-16
```

Estimated regression model

Pcrev(y^)

= - 2476.2565+89.4853Bed+26.9274(Mcdays)+

0.8978(Nsal).

- Interpretation : The coefficient of Bed (here 89.4853) can be interpreted as for every per unit increase in number of Bed in health care home, the Annual Total Patient Care Revenue increases on an average by 89.4853($100) keeping all other predictors constant.

Goodness of model

- Adjusted R-square =0.8922 which is more than 0.8. Hence the model can be used for point prediction.
- Also the difference between multiple R-square and adjusted R-square is small which again justify our conclusion.