Regression Project
STAT 104 - Introduction to Quantitative Methods for Economics

We are interested in the general question of what factors impact health care utilization spending among the elderly. Medicare, the federal health insurance program for the elderly, is the fastest growing expense in the federal budget. Knowledge of what factors contribute to health care expenditures will possibly help identify what sort of programs to implement to reduce future expenditures. Our data comes from the 2005 Medical Expenditures Panel Survey. A description of the variables is at the end of this project document. The explanatory variable is totalexp.

→ Load data into R

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/hospvisits.csv")
```

→ Fit the model

```
> fit = lm (totalexp~., data=mydata)
> summary(fit)

Call:
lm(formula = totalexp ~ ., data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-23276  -3126  -1164   1081  56573

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.230e+02  4.811e+03   0.046  0.96305
age          2.560e+01  4.986e+01   0.513  0.60785
marital     -2.066e+01  3.720e+02  -0.056  0.95573
educ         1.802e+01  9.361e+01   0.193  0.84737
income       2.762e-03  1.374e-02   0.201  0.84068
srhealth     1.118e+03  3.080e+02   3.631  0.00030 ***
mntl_hlth   -3.439e+02  3.264e+02  -1.054  0.29238
phy_lim      1.113e+03  6.738e+02   1.652  0.09900 .
bmi         -6.974e+01  5.778e+01  -1.207  0.22781
chd          8.734e+02  8.892e+02   0.982  0.32626
high_chol   -7.524e+01  5.943e+02  -0.127  0.89928
diabetes     2.363e+03  7.525e+02   3.140  0.00176 **
dr_visits    2.055e+02  2.453e+01   8.379 2.48e-16 ***
msa          4.715e+02  7.313e+02   0.645  0.51923
race_grp    -1.903e+02  3.093e+02  -0.615  0.53848
smoker       6.872e+02  9.531e+02   0.721  0.47107
male        -1.410e+02  6.113e+02  -0.231  0.81762
high_bp     -8.772e+02  6.417e+02  -1.367  0.17207
hosp_vis     1.154e+04  4.866e+02  23.723  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8010 on 780 degrees of freedom
Multiple R-squared:  0.5461,    Adjusted R-squared:  0.5356
F-statistic: 52.13 on 18 and 780 DF,  p-value: < 2.2e-16
```

Karan A. Bhandarkar

Regression Project
STAT 104 - Introduction to Quantitative Methods for Economics

This is an extremely poor model with very high Residual Standard error of 8010.

→ Let's start with checking for Variation Inflation Factor's (VIF) in the model.

Install car package in R.

```
> install.packages("car")
> library(car)
> vif(fit)
```

```
      age   marital      educ    income  srhealth mntl_hlth   phy_lim       bmi       chd high_chol  diabetes dr_visits       msa
 1.216510  1.121520  1.513334  1.275801  1.638846  1.327420  1.323874  1.178747  1.168973  1.093484  1.155539  1.143919  1.041297
 race_grp    smoker      male   high_bp  hosp_vis
 1.263387  1.086478  1.132160  1.136872  1.137975
```

There's nothing unusual in the VIFs so multicollinearity is not a problem.

→ Check for Heteroskedasticity

```
> ncvTest(fit)

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 682.5554    Df = 1     p = 1.858761e-150
```

The p-value is less than 0.05 so the data is heteroskedastic

Check to see if there is any interaction variable:
High_bp is related to age so it can make interaction variable.

```
>
fit1=lm(totalexp~age+marital+educ+income+srhealth+mntl_hlth+phy_lim+bmi+chd+high_ch
ol+diabetes+dr_visits+msa+race_grp+smoker+male+high_bp+hosp_vis+age*high_bp,data=my
data)
```

→ Check again for Heteroskedasticity

> ncvTest(fit1)

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 682.9457    Df = 1     p = 1.528791e-150
```

The p value is less than 0.05 so the data is heteroskedastic. We can remove outliers which has p value less than 1.8.

Regression Project
STAT 104 - Introduction to Quantitative Methods for Economics

```
> newdata=subset(mydata,abs(rstudent(fit1))<1.8)
> fit2=update(fit1,.~.,data=newdata)
```

→ Check again for Heteroskedasticity

> ncvTest(fit2)

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 310.1272    Df = 1     p = 2.048863e-69
```

The p value is less than 0.05 so the data is heteroskedastic. We can take log of Y variable.

```
> fit3=lm(log(totalexp)~.,data=newdata)
```

→ Check again for Heteroskedasticity

> ncvTest(fit3)

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.4165513    Df = 1     p = 0.5186629
```

The p value is greater than 0.05 so the data is homoscedastic.
The new model is created after removing heteroskedasticity and multicollinearity.

→ Let's test the model for normality

```
> summary(fit3)

Call:
lm(formula = log(totalexp) ~ ., data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1663 -0.4290  0.0304  0.5381  1.8876

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.656e+00  4.891e-01  13.608  < 2e-16 ***
age          1.019e-02  5.068e-03   2.011 0.044721 *
marital     -2.316e-02  3.770e-02  -0.614 0.539111
educ         1.899e-04  9.484e-03   0.020 0.984027
income       1.362e-06  1.376e-06   0.990 0.322437
srhealth     1.149e-01  3.166e-02   3.629 0.000304 ***
mntl_hlth   -3.876e-02  3.395e-02  -1.142 0.253925
phy_lim      1.937e-01  6.838e-02   2.833 0.004743 **
bmi         -6.652e-03  6.056e-03  -1.098 0.272361
chd          1.499e-01  9.111e-02   1.645 0.100455
```

```
high_chol    2.907e-01  6.075e-02   4.785 2.07e-06 ***
diabetes     3.266e-01  7.745e-02   4.217 2.79e-05 ***
dr_visits    3.555e-02  2.878e-03  12.353  < 2e-16 ***
msa         -3.560e-02  7.447e-02  -0.478 0.632754
race_grp    -8.450e-02  3.141e-02  -2.690 0.007314 **
smoker      -8.800e-02  9.591e-02  -0.917 0.359200
male         1.922e-02  6.199e-02   0.310 0.756557
high_bp      8.769e-02  6.504e-02   1.348 0.177992
hosp_vis     9.920e-01  6.319e-02  15.698  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7865 on 732 degrees of freedom
Multiple R-squared:  0.5132,    Adjusted R-squared:  0.5013
F-statistic: 42.88 on 18 and 732 DF,  p-value: < 2.2e-16
```

→ Lets do the backward stepwise regression to remove unwanted x variables:

```
> fit4=step(fit3)
> summary(fit4)

Call:
lm(formula = log(totalexp) ~ age + srhealth + phy_lim + chd +
    high_chol + diabetes + dr_visits + race_grp + hosp_vis, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2450 -0.4269  0.0285  0.5352  1.9746

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.481704   0.360672  17.971  < 2e-16 ***
age          0.010080   0.004713   2.139 0.032792 *
srhealth     0.098962   0.028282   3.499 0.000495 ***
phy_lim      0.170418   0.066692   2.555 0.010808 *
chd          0.184649   0.088881   2.077 0.038101 *
high_chol    0.285412   0.059486   4.798 1.94e-06 ***
diabetes     0.321931   0.075606   4.258 2.33e-05 ***
dr_visits    0.035581   0.002837  12.543  < 2e-16 ***
race_grp    -0.092060   0.029045  -3.170 0.001590 **
hosp_vis     0.990304   0.062267  15.904  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7855 on 741 degrees of freedom
Multiple R-squared:  0.5086,    Adjusted R-squared:  0.5026
F-statistic:  85.2 on 9 and 741 DF,  p-value: < 2.2e-16
```

Regression Project
STAT 104 - Introduction to Quantitative Methods for Economics

→ Check again for Heteroskedasticity

```
> ncvTest(fit4)

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.2901777    Df = 1     p = 0.5901067
```

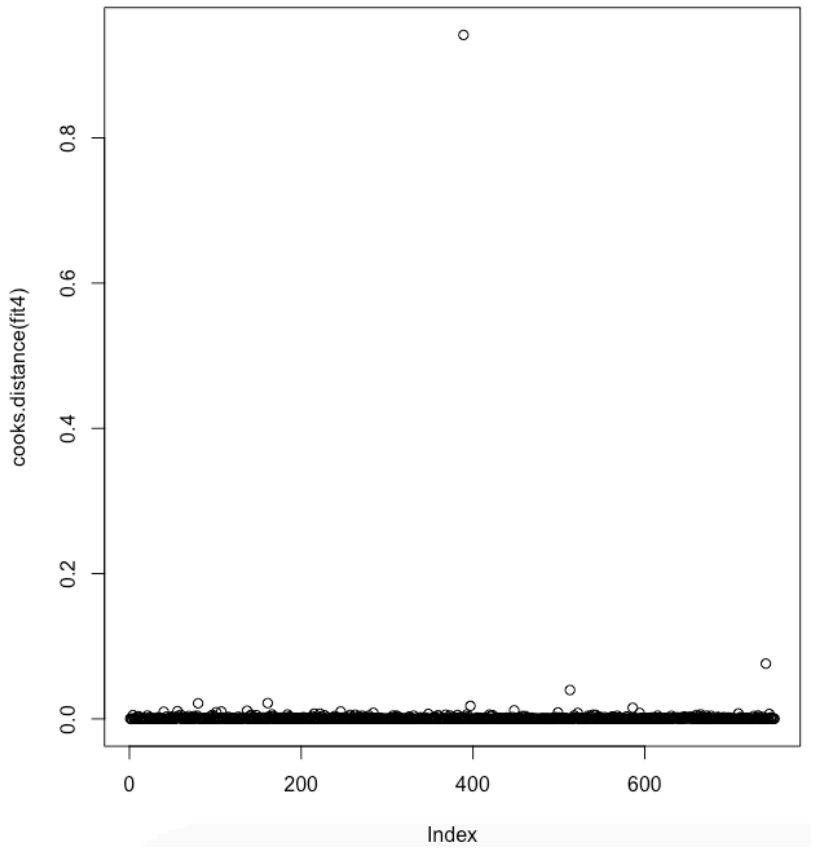
→ Test for normality

```
> shapiro.test(residuals(fit4))

        Shapiro-Wilk normality test

data:  residuals(fit4)
W = 0.97776, p-value = 2.852e-09
```

The residuals are not normal.

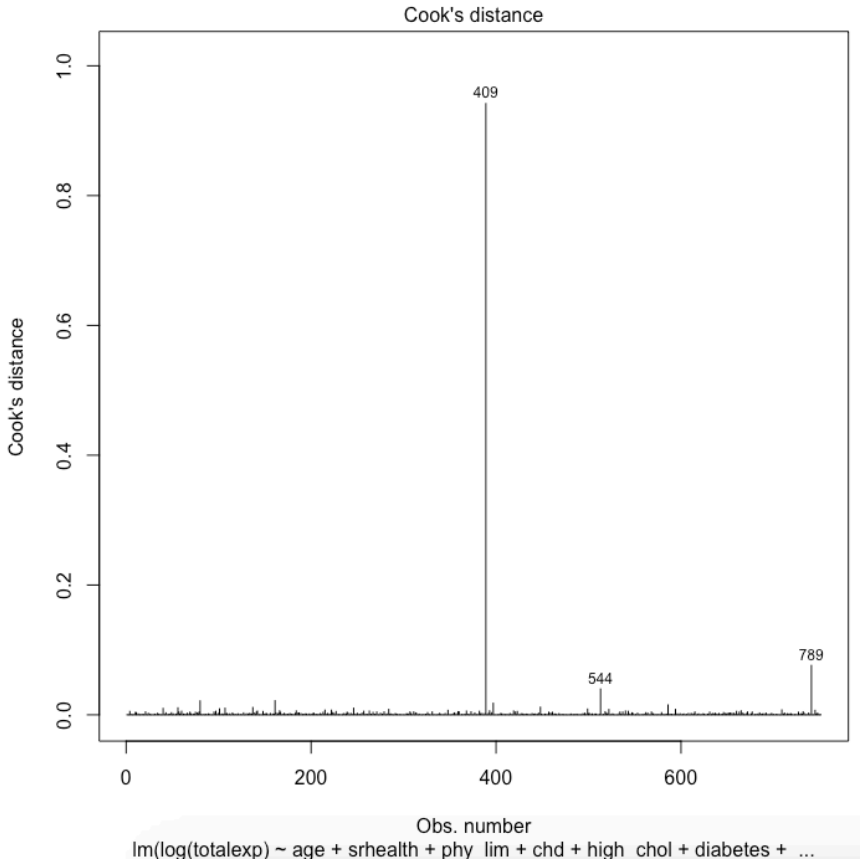
→ Lets look at Cook's distance values larger than usual as cases we want to examine more

```
> plot(cooks.distance(fit4))
```



Regression Project
STAT 104 - Introduction to Quantitative Methods for Economics

```
> plot(fit4, which = 4)
```



→ Remove Outliers

```
> cooksnewdata=newdata[-c(409,544,789),]
> fit5=update(fit4,.~.,data=cooksnewdata)
```

→ Check again for Heteroskedasticity

```
> ncvTest(fit5)

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.2444142    Df = 1     p = 0.621036
```

→ Test for normality

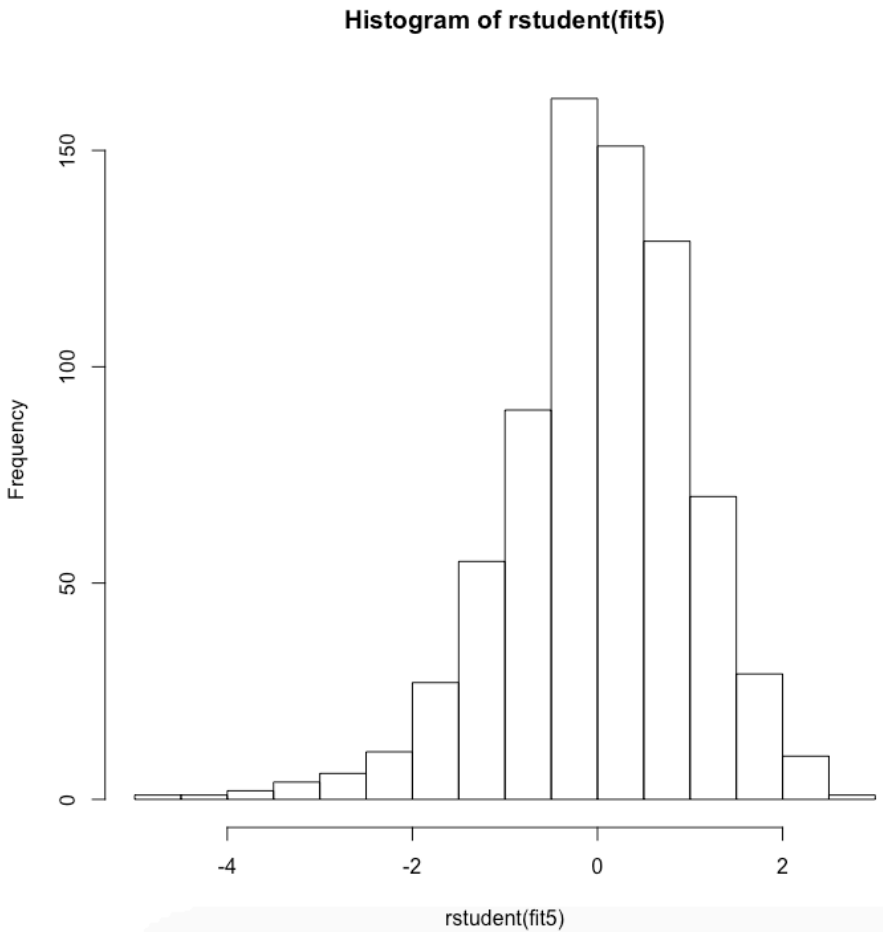
```
> shapiro.test(residuals(fit5))
```

Regression Project
STAT 104 - Introduction to Quantitative Methods for Economics

```
        Shapiro-Wilk normality test

data:  residuals(fit5)
W = 0.97787, p-value = 3.19e-09
```

→ Plot the Histogram

```
> hist(rstudent(fit5))
```



Conclusion: The data does not seem to be normal and there are still outliers.

Regression Project
STAT 104 - Introduction to Quantitative Methods for Economics

→ Best fit model

```
> summary(fit5)

Call:
lm(formula = log(totalexp) ~ age + srhealth + phy_lim + chd +
    high_chol + diabetes + dr_visits + race_grp + hosp_vis, data = cooksnewdata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2451 -0.4270  0.0293  0.5341  1.9750

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.493427   0.361332  17.971  < 2e-16 ***
age          0.009852   0.004726   2.085 0.037432 *
srhealth     0.099884   0.028317   3.527 0.000446 ***
phy_lim      0.167302   0.066898   2.501 0.012605 *
chd          0.187985   0.089296   2.105 0.035611 *
high_chol    0.287826   0.059574   4.831 1.65e-06 ***
diabetes     0.321516   0.075668   4.249 2.42e-05 ***
dr_visits    0.035643   0.002839  12.553  < 2e-16 ***
race_grp    -0.091632   0.029064  -3.153 0.001683 **
hosp_vis     0.990624   0.062343  15.890  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7859 on 739 degrees of freedom
Multiple R-squared:  0.5093,    Adjusted R-squared:  0.5033
F-statistic: 85.22 on 9 and 739 DF,  p-value: < 2.2e-16
```

Karan A. Bhandarkar