Sample Answer Lab06: Regression Analysis

Part I: Multiple Regression Analysis Tasks (2.8 points)

Attach the libraries **car** and **effects** to your session. Read information on 506 communities in the Boston area from the internet with the statement:

```
hprice <- foreign::read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/hprice2.dta")
```

We will focus on the variables:

Variable	Description
price	Median home value in $100 in the community (dependent variable)
nox	Nitrogen-oxide (measure of traffic related air pollution)
dist	Weighted distance from 5 major employment centers
stratio	Student-teacher ratio in the community
rooms	Average number of rooms

Your ***response variable*** is the **price** and the remaining variables are your ***exogenous variables***.

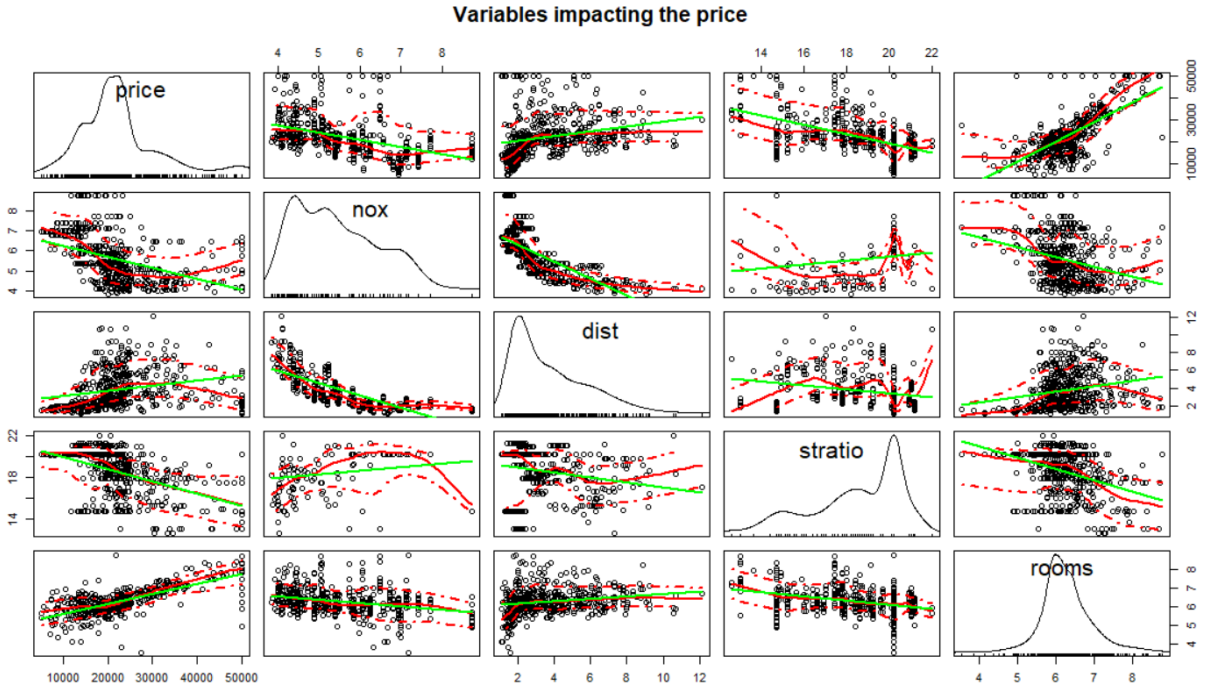
Task 1: Formulate ***explicit hypotheses*** and your ***common-sense*** rationale on how four exogenous variables will influence home price. Place your answer in the table below (0.4 points)

Independent Variable	Direction of Influence	Rationale
nox	a negative relationship is expected between it and house price.	Concentration of nitrogen-oxide, a measure of air pollution, negatively impacts on the house price
dist	distance is expected to be negatively associated with house price	The closer a home is to employment centers the less are the commuting cost. Thus the home value will be higher.
stratio	a negative relationship is expected between the two variables.	A high student-teacher ratio implies the school is underfunded and the education quality lacks, so the houses in these districts should be in lower values
rooms	rooms variable is expected to be positively related with house price.	Houses should be more expensive if they are larger and have more rooms

Task 2: Generate a scatterplot matrix of the five variables **price**, **nox**, **dist**, **stratio** and **rooms**. Make sure that the dependent variable is the first one in the list. ***Thoroughly interpret*** the individual distributions of the variables and their pairwise relationships. (0.5 points)

```
car::scatterplotMatrix(~price+nox+dist+stratio+rooms,
       data=hprice,main="Variables impacting the price",col = "black",
```

```
smooth=list(span =
        0.35,lty.smooth=1,col.smooth="red",col.var="red"),regLine=list(col="gr
        een"))
```



Individual distribution:

House price is slightly positively skewed, with few outliers appearing on the right.

nox is positively skewed, also has one apparent outlier on the right.

dist is highly positively skewed.

rooms is very close to a symmetric distribution.

stratio is negatively skewed, with few outliers appearing on the left.

Pairwise relationships:

nox: negatively associated with house price, however, the relationship appears to be weak.

dist: positively associated with house price.

rooms: an apparent quadratic relationship with house price. House price drops as number of rooms increases when the number is small. However, house price increases as number of rooms increases when total number of rooms is large.

stratio: negatively associated with house price.

Task 3: Run a multiple regression model of the **`price`** onto the four independent variables. Interpret the estimated regression coefficients with regards to your stated hypothesis in Task 1. (0.4 points)

Also interpret the R^2 statistics. (0.3 points)

```
mod1 <- lm(price~nox+dist+stratio+rooms, data=hprice)
summary(mod1)
Call:
```

```
lm(formula = price ~ nox + dist + stratio + rooms, data = hprice)
Residuals:
   Min     1Q Median     3Q    Max
-14310  -3124   -546   2181  38580
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23716.2     5120.6   4.632 4.63e-06 ***
nox          -3044.9      353.7  -8.609  < 2e-16 ***
dist          -965.5      191.5  -5.042 6.45e-07 ***
stratio      -1269.2      127.4  -9.965  < 2e-16 ***
rooms         6808.8      401.4  16.964  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5701 on 501 degrees of freedom
Multiple R-squared:  0.6198,  Adjusted R-squared:  0.6168
F-statistic: 204.2 on 4 and 501 DF,  p-value: < 2.2e-16
```

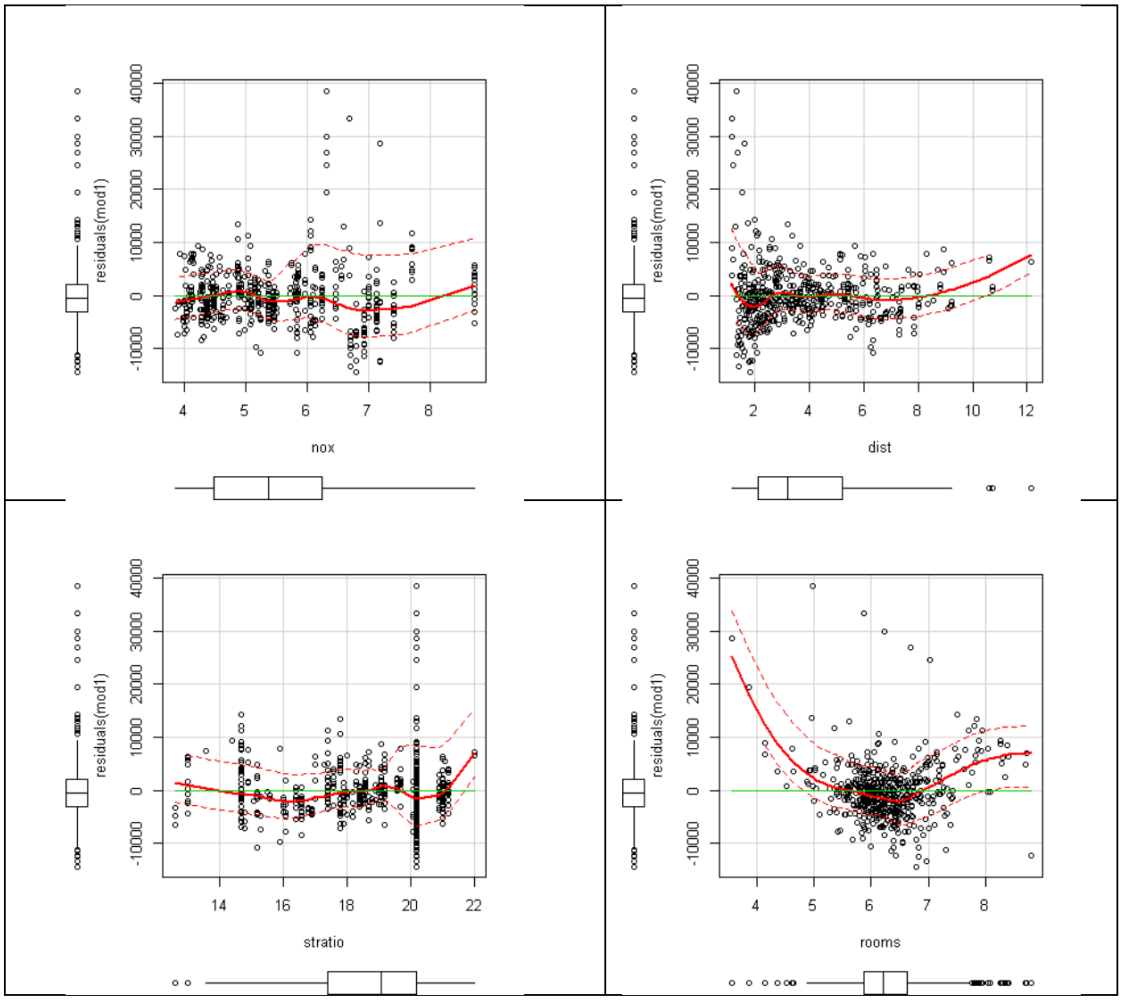
Regression results suggest that all independent variables are relevant (see ***). $R2$ is 0.6198, it means the independent variables combined explain 61.98% of the variance of the dependent variable.

nox, dist, and stratio are negatively associated with house price, which means as these values increase, house price goes down. The estimated coefficients suggest that with one unit increase of each of the three variables, the house price decreases $3044.9, $965.5, and $1269.2 respectively. These negative associations are the same as the stated hypothesis. Rooms variable is positively associated with house price, which means house price increases as the number of room increases. Essentially, one additional room in a house lead to house price increases by $6908.8, this positive relationship is also consistent with the stated hypothesis.

Interestingly, while the bivariate relationship between home value and distance indicated a positive relationship, the multiple regression model corrected this counterintuitive relationship.

Task 4: Generate a scatterplot of the model's regression residuals from Task 3 against each independent variable. Place the independent variables on the x-axis. Are the residuals independent of the exogenous variables? (0.4 points)

```
car::scatterplot(residuals(mod1)~nox, data=hprice, col =
       "black",smooth=list(span = 0.35, lty.smooth=1,col.smooth="red",
       col.var="red"), regLine=list(col="green"))
car::scatterplot(residuals(mod1)~dist, data=hprice,col =
       "black",smooth=list(span = 0.35, lty.smooth=1,col.smooth="red",
       col.var="red"), regLine=list(col="green"))
car::scatterplot(residuals(mod1)~stratio, data=hprice, col =
       "black",smooth=list(span = 0.35, lty.smooth=1,col.smooth="red",
       col.var="red"), regLine=list(col="green"))
car::scatterplot(residuals(mod1)~rooms, data=hprice, col =
       "black",smooth=list(span = 0.35, lty.smooth=1,col.smooth="red",
       col.var="red"), regLine=list(col="green"))
```



Task 5: Is there potentially a quadratic relationship with regards to independent variable **`rooms`**? (0.1 points)

The red curves in the first three plots align with the horizontal green lines; it means that there is no non-linear relationship between these three variables and the regression residuals. However, the fourth scatterplot indicate a quadratic relationship with regards to the independent variable rooms.

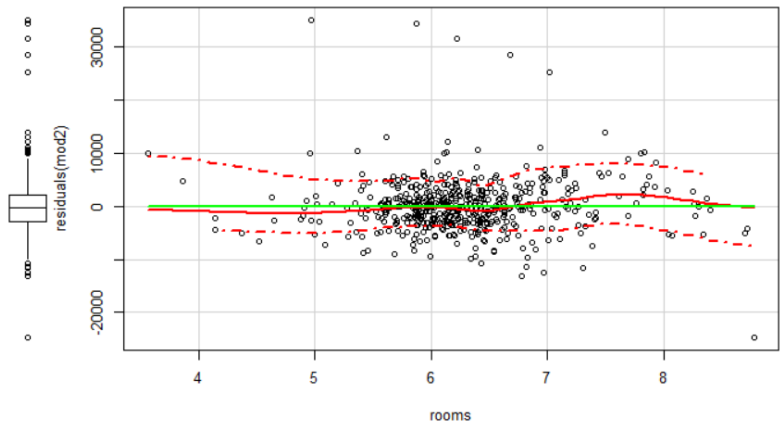
Task 6: Rerun the model in Task 3 by augmenting it with the squared number of rooms as fifth exogenous variable.

Note: the R-formula statement in the **`lm( )`** function needs to wrap the squared number of rooms inside the inhibit function **`I(rooms^2)`**.

Plot the residuals of the augmented model against the number of rooms. Has the potentially non-linear relationship been fixed? (0.2 points).

```
mod2 <- lm(price~nox+dist+stratio+rooms+I(rooms^2), data=hprice)
summary(mod2)
Call:
lm(formula = price ~ nox + dist + stratio + rooms + I(rooms^2),
    data = hprice)
Residuals:
```

```
   Min     1Q Median     3Q    Max
-24609  -2831   -225   2167  34950
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 120386.8    10964.2  10.980  < 2e-16 ***
nox          -3086.5      324.5  -9.511  < 2e-16 ***
dist          -723.5      177.4  -4.078 5.29e-05 ***
stratio      -1082.9      118.4  -9.146  < 2e-16 ***
rooms       -24993.1     3279.8  -7.620 1.28e-13 ***
I(rooms^2)    2477.3      253.9   9.758  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5230 on 500 degrees of freedom
Multiple R-squared:  0.6806,  Adjusted R-squared:  0.6774
F-statistic: 213.1 on 5 and 500 DF,  p-value: < 2.2e-16
```



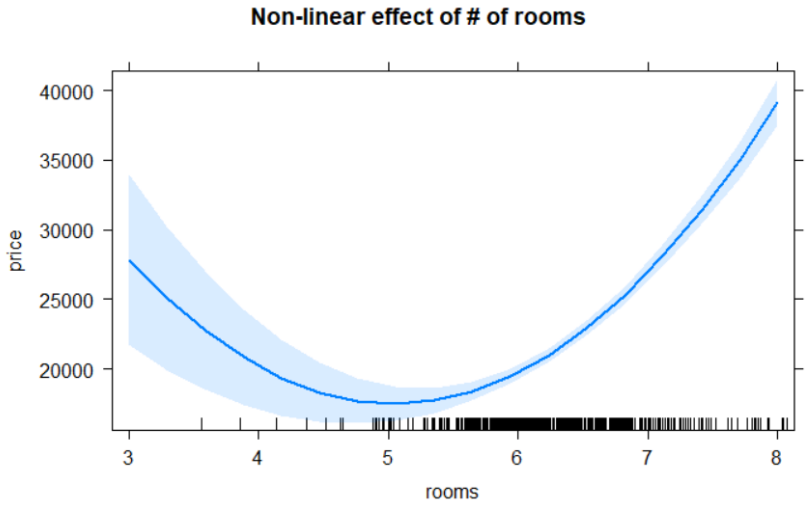
After introducing the quadratic term to the model, the quadratic pattern is no longer apparent, the red curves track the horizontal green line well.

Task 7: Why does it become difficult to interpret the regression parameters associated with the number of rooms in the augmented quadratic model? Hint: look at the signs of the linear and the quadratic terms. (0.1 points)

To enable the interpretation, visualize the quadratic relationship between the number of rooms and the home price with an effects plot. Note that the number of rooms ranges from 3 to 9. (0.2 points)

Interpret your non-linear effects plot. (0.2 points)

```
lm.03.eff <-
      effects::allEffects(mod2,xlevels=list(rooms=3:max(hprice$rooms)))
plot(lm.03.eff, "rooms", main="Non-linear effect of # of rooms")
```



Because there are two terms that related to rooms variable in the augmented model, both terms need to be interpreted simultaneously in order to infer the correct relationship between house price and rooms variable. In addition, the quadratic term measures a non-linear relationship, its interpretation is different from a linear term. The estimated coefficient for this quadratic term does not indicate a unique "slope" that applies throughout the value range of rooms.

The effect plot depicts the quadratic relationship between the rooms variable and house price. Essentially, it indicates that house price decreases as number of rooms increase when there are no more than 5 rooms in a house. However, the rugs indicate that number of houses with no more than 5 rooms only accounts for a small portion. A majority of the houses have more than 5 rooms, and their prices increase as number of rooms increases.

Part II: Partial Regression Effects (1.2 points)

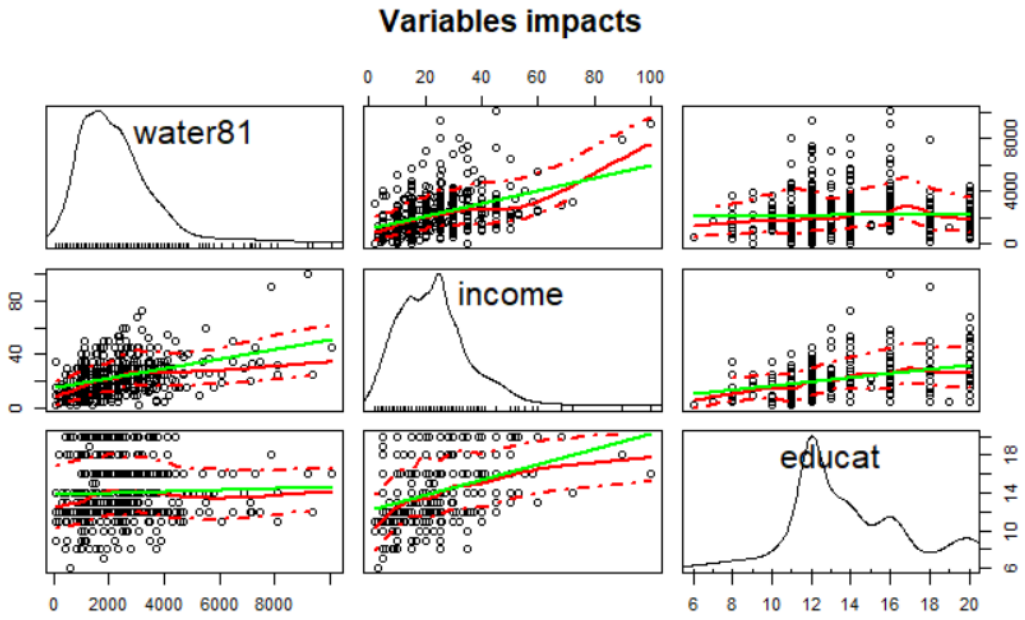
Open the SPSS file **`Concord1.sav`** in your R session.

Task 8: Formulate hypotheses on how the water consumption in 1981 (**`water81`**) is influenced by the household's income (**`income`**) and the education level (**`educat`**) of the household's head. Justify your hypotheses with common sense arguments. If you are not sure, then explain why. (0.4 points)

Variable	Common Sense Arguments
income	Income is expected to be positively associated with water consumption, because high income people usually have larger houses, which consume more water.
educat	Education is related to income, for example, well-educated population are generally high income earners. So similarly, education is expected to be positively associated with water consumption. Alternatively one can make an argument that well educated person are better in not wasting money and water and therefore consume less. Thus the multiple analysis will need to figure the relationship out.

Task 9: Generate a scatterplot matrix of the three variables. Do the bivariate relationships between **`water81~income`** and **`water81~educat`** support your hypotheses from Task 8? (0. 4 points)

```
df <- foreign::read.spss("Concord1.sav")
car::scatterplotMatrix(~water81+income+educat, data=df,main="Variables
      impacts",col = "black",
smooth=list(span =
      0.35,lty.smooth=1,col.smooth="red",col.var="red"),regLine=list(
      col="green"))
```



The scatterplots above suggests that income variable is positively associated with water consumption; education variable also displays a slight positive association with water. Furthermore, education and income are positively correlated.

Task 10: Run the multiple regression model **`water81~income+educat`**. In comparison to the bivariate relationship in Task9 why does the interpretation of the education effect in the multiple regression model change? Consider the correlation between income and education in your argument. (0.4 points)

```
mod<- lm(water81 ~ income + educat, data=df)
summary(mod)
lm(formula = water81 ~ income + educat, data = df)
Residuals:
    Min      1Q  Median      3Q     Max
-2821.2  -874.8  -232.0   594.2  6887.0
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1891.262    280.334   6.746 4.26e-11 ***
income        52.218      4.927  10.599  < 2e-16 ***
educat       -56.976     20.816  -2.737  0.00642 **
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1343 on 493 degrees of freedom
Multiple R-squared:  0.1869,  Adjusted R-squared:  0.1836
F-statistic: 56.66 on 2 and 493 DF,  p-value: < 2.2e-16
```

Regression model results indicate that both income and education variables are relevant (*** and ** stars). The estimated coefficients suggest a positive relationship between water consumption and income, when income increases one-thousand dollars, the water consumption increases 52.218 $ft3$probably because affluent people have bigger houses with larger lots and perhaps swimming polls.

Education variable is negatively associated with water consumption variable; essentially higher educated people tend to consume less water. This may be because of environmental concerns or being savvy in saving money on the water bill. So if the household head has one more year of education, the water consumption decreases 56.976 $ft3$. The association between water consumption and education is different from the slightly positive relationship in the bivariate scatterplot. After controlling for income, the impact of education on the water consumption changes.