Problem 1

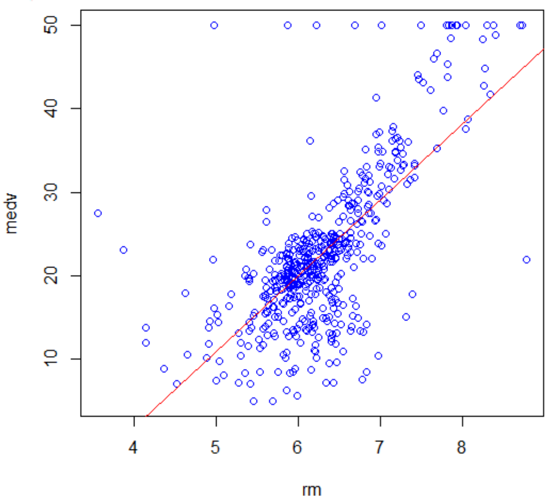
(1) Fit a linear model for the Boston dataset in MASS library using median value of owner-occupied homes (*medv*) as response and average number of rooms per dwelling (*rm*) as the predictor (use the basic syntax *lm(y~x,data=dataname)*. What are the coefficients? What does it suggest about the fitness? Show the scatter plot as well as the linear model fit in one figure.

Solution:

Below is the snapshot of the code using R. The coefficients of the linear model is *Intercept* (β_0) = -34.670621 and *rm* (β_1) = *9.102109*. From the p-value of the coefficients β_0 and β_1, which is less than .05 (assuming we want 95% confidence in our prediction model), we can say that both the coefficients are significant and there exists a relationship between *medv* and *rm* variables in the dataset.

```
> lm.fit<-lm(medv~rm,Boston)
> summary(lm.fit)

Call:
lm(formula = medv ~ rm, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-23.346  -2.547   0.090   2.986  39.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -34.671      2.650  -13.08   <2e-16 ***
rm             9.102      0.419   21.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.616 on 504 degrees of freedom
Multiple R-squared:  0.4835,    Adjusted R-squared:  0.4825
F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16

> coef(lm.fit)
(Intercept)          rm
 -34.670621    9.102109
```

```
> plot(rm,medv,col = "blue")
> abline(lm.fit,col="red")
```



The R^2 value is .4835 which suggests that the *rm* variable accounts for approximately 48.35% variation in *medv*. From the coefficients we can see a positive relationship between *medv* and *rm* since the slope or β1 is positive. For one unit increase in *rm* the *medv* (in $1000) increases approximately 9 times. From a business point of view this relation tells us that the higher the average number of rooms the higher will be the median value of the homes.

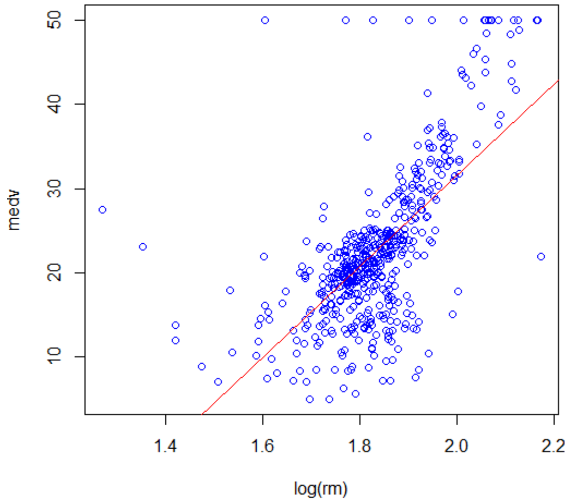
(2) Fit a linear model using the same input and output in Question (1), but replace the predictor with log(rm) (i.e., use the basic syntax lm(y~log(x),data=dataname)). What are the coefficients? Is this model a better fit compared to the one in Question (1)? Justify your answer. Show the data and the linear model fit in one figure.

Solution:

Below is the snapshot of the code using R. The coefficients of the linear model is *Intercept* (β_0) = -76.488 and *rm* (β_1) = *54.055*. From the p-value of the coefficients β_0 and β_1, which is less than .05 (assuming we want 95% confidence in our prediction model), we can say that both the coefficients are significant and there exists a relationship between *medv* and log(*rm*) variables in the dataset.

```
> lm.fitlogrm<-lm(medv~log(rm),data=Boston)
> summary(lm.fitlogrm)

Call:
lm(formula = medv ~ log(rm), data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.487  -2.875  -0.104   2.837  39.816 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -76.488      5.028  -15.21   <2e-16 ***
log(rm)       54.055      2.739   19.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.915 on 504 degrees of freedom
Multiple R-squared:  0.4358,	Adjusted R-squared:  0.4347 
F-statistic: 389.3 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
> plot(log(rm),medv,col="blue")
> abline(lm.fitlogrm,col="red")
```



The R^2 value is 0.4358 which suggests that the log(*rm*) variable accounts for approximately 43.58% variation in *medv*. However, the R^2 value is less that of the first case (*medv* ~ *rm*). This suggests that the model *medv*~log*(rm)* is not as good a fit as the model *medv~rm* as log(*rm*) accounts for less variation as compared to the *rm* (0.4835) variable.

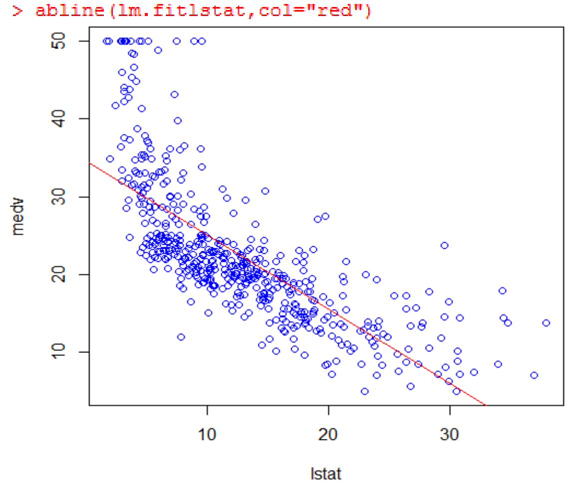
(3) Fit a linear model using the same output (medv) in Question (1), but regress it against the lstat variable. What are the coefficients? How does this model fit compared to the one in Question (1)?

Solution:

Below is the snapshot of the code using R. The coefficients of the linear model is *Intercept* (β_0) = 34.55384 and *lstat* (β_1) = *-0.95005*. From the p-value of the coefficients β_0 and β_1, which is less than .05 (assuming we want 95% confidence in our prediction model), we can say that both the coefficients are significant and there exists a relationship between *medv* and *lstat* variables in the dataset.

```
> lm.fitlstat<-lm(medv~lstat,data=Boston)
> summary(lm.fitlstat)

Call:
lm(formula = medv ~ lstat, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.168  -3.990  -1.318   2.034  24.500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,	Adjusted R-squared:  0.5432 
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
> plot(lstat,medv,col="blue")
> abline(lm.fitlstat,col="red")
```



The R^2 value is 0.5441 which suggests that the *lstat* variable accounts for approximately 54.41% variation in *medv*. The R^2 value is higher than that of *medv~rm*. From the coefficients we can see a negative relationship between *medv* and *lstat* since the slope or β1 is negative. For one unit increase in *lstat* the medv (in $1000) decreases approximately 0.9 times. From a business point of view this relation tells us that as the lower status population percentage increases in a neighborhood, the median value of the homes decreases.

Problem 2

Fit a regression model of medv on lstat and lstat^2 (syntax *lm(y~x+I(x^2), data=dataname)*). Provide a summary of the model. Suppose that we have another linear model which simply fits medv with predictor lstat (used in Question 1(3)), which model has better fitness? Justify your answer.

Solution:

Below is the regression analysis for a quadratic model. When compared to the model *medv~lstat*, the quadratic model has a better R2 (0.6407 compared to 0.5441). This shows that the quadratic term with *lstat* and *lstat^2* capture more variance in *medv* than a simple linear model with only *lstat* as predictor variable. To further test the quadratic model against simple linear model, we can compare the two models using **anova()** function.

```
> lm.fit2<-lm(medv~lstat+I(lstat^2))
> summary(lm.fit2)

Call:
lm(formula = medv ~ lstat + I(lstat^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2834  -3.8313  -0.5295   2.3095  25.4148 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 42.862007   0.872084   49.15   <2e-16 ***
lstat       -2.332821   0.123803  -18.84   <2e-16 ***
I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.524 on 503 degrees of freedom
Multiple R-squared:  0.6407,    Adjusted R-squared:  0.6393 
F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16
```

```
> anova(lm.fitlstat,lm.fit2)
Analysis of Variance Table

Model 1: medv ~ lstat
Model 2: medv ~ lstat + I(lstat^2)
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)    
1    504 19472                                 
2    503 15347  1    4125.1 135.2 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova() function performs a hypothesis test by comparing the two models. It tests the below hypothesis:

H_0: The two models are the same fit
H_a: The quadratic model is a better fit

The resulting F-statistic with a p-value nearly zero proves that H_0 can be rejected and H_a is true. This proves that the quadratic model provides a superior fit when compared to the simple linear model.

Problem 3

(1) Except lstat and rm, there are other predictors in the Boston dataset. You can check the whole dataset using syntax *?Boston* and *summary(Boston)*. Fit a multiple linear regression model of medv on all the predictors (syntax: *lm(y~., data=dataname)*). What are the coefficients? What does it suggest about fitness?

Solution:

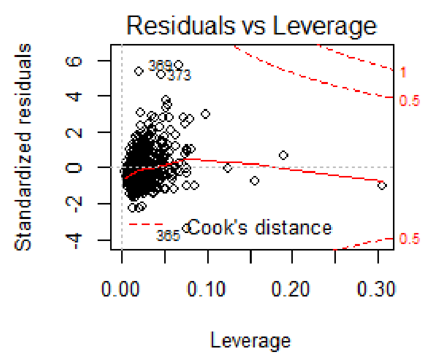
We fit the model with all the predictor variables included in the linear model equation.

```
> lm.fitall<-lm(medv~.,data=Boston)
> summary(lm.fitall)

Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.595  -2.730  -0.518   1.777  26.199 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 ** 
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288    
chas         2.687e+00  8.616e-01   3.118 0.001925 ** 
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229    
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 ** 
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,	Adjusted R-squared:  0.7338 
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

The model has high R^2 of 0.7406 which accounts for 74.06% of variance in *medv* and the model is a good fit as the R^2 value is near 1.

Additionally, the F-statistic is 108.1 and the corresponding p-value is negligible. This test proves that at least one input variable has a significant effect on the output variable *medv*.

The plot also has high leverage points which has potentially affected the model fit.



(2) Do you think this model is some sort of cumbersome? Improve this model by reducing the inputs based on the summary of the model in Question 3(1)). Explain the methodology used for variable selection and provide a summary of the final model.
(Syntax: lm(y~predictor1+predictor2+…+predictorN, data=dataname))

Solution:

Yes the model in (1) is cumbersome as it has a high complexity and involves using all the available variables to create a model. However, by observing the individual p-values of the variables *indus* and *age* we can conclude that these variables are not significantly related to the response or don't contribute significantly towards the response as they have p-values greater than .05. We can use a backward selection to achieve a model with variables that give a better fit.

We can try excluding these variables in our next model to check if the model can be improved.

```
> lm.fitall2<-lm(medv~.-age-indus,data=Boston)
> summary(lm.fitall2)

Call:
lm(formula = medv ~ . - age - indus, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.5984  -2.7386  -0.5046   1.7273  26.2373 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
crim         -0.108413   0.032779  -3.307 0.001010 ** 
zn            0.045845   0.013523   3.390 0.000754 ***
chas          2.718716   0.854240   3.183 0.001551 ** 
nox         -17.376023   3.535243  -4.915 1.21e-06 ***
rm            3.801579   0.406316   9.356  < 2e-16 ***
dis          -1.492711   0.185731  -8.037 6.84e-15 ***
rad           0.299608   0.063402   4.726 3.00e-06 ***
tax          -0.011778   0.003372  -3.493 0.000521 ***
ptratio      -0.946525   0.129066  -7.334 9.24e-13 ***
black         0.009291   0.002674   3.475 0.000557 ***
lstat        -0.522553   0.047424 -11.019  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.736 on 494 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7348 
F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

By eliminating the indus and age variables we have achieved a model with the same R^2 as the initial cumbersome model.

Additionally the F-statistic has increased to 128.2 which tell us that this is a better model than (1). Also all the individual p-values of the variables are less than .05 which tells us that each parameter has a significant relationship with the output variable *medv*.