HOMEWORK -3

QUESTION 1:

The $R^2$is the co-efficient of multiple determination. It's a ratio of regression sum of squares to the total sum of squares around the mean, it does not have any particular unit associated with it. Its is the percentage change in the Y that is denoted by the predictor variable in the model.

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Thus you can conclude that (SSR/SSTO) % of the variability in Y can be explained by the linear relationship between X and Y. In a Regression model there is no implication that Y necessarily depends on X in a causal or explanatory sense. The coefficient of determination does not say anything about the causal relationship between the explanatory and response variable.

QUESTION 2:

$R^2$is defined as measure the proportionate reduction in of total variation in in Y associated with the use of the set of predictor variables. Its value is defined by

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Adjusted Coefficient of Multiple Determination is

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

Here the degrees of freedom are considered in this co-efficient.

Yes for the econometric model considered the value of R^2 is higher for the second model. This is due to the fact that adding more variables to a model can only explain more errors. Adjusted R square considers into account the number of variables. If explicitly creates a conflict between two terms when an additional variable is added to the model: adding a variable will decrease SSE. Adding a variable to general linear regression model (GLRM) will sufficiently increase the p variable value, which subsequently decreases the adjusted R squared value. Also adding a variable decreases SSE, there by increasing adjusted R square. Thus the value can either go up or down based on the situation. Where as in terms of R square, it can only go up when a new variable is added to the system. Given a GLRM model, always $R^2 > R_a^2$.

QUESTION 3:

SUMMARY STATISTICS:

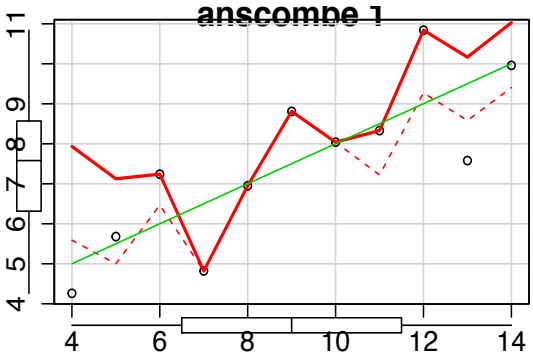
```
> summary(anscombe)
       x1             x2             x3             x4           y1              y2              y3              y4        
 Min.   : 4.0   Min.   : 4.0   Min.   : 4.0   Min.   : 8   Min.   : 4.260   Min.   :3.100   Min.   : 5.39   Min.   : 5.250  
 1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 8   1st Qu.: 6.315   1st Qu.:6.695   1st Qu.: 6.25   1st Qu.: 6.170  
 Median : 9.0   Median : 9.0   Median : 9.0   Median : 8   Median : 7.580   Median :8.140   Median : 7.11   Median : 7.040  
 Mean   : 9.0   Mean   : 9.0   Mean   : 9.0   Mean   : 9   Mean   : 7.501   Mean   :7.501   Mean   : 7.50   Mean   : 7.501  
 3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.: 8   3rd Qu.: 8.570   3rd Qu.:8.950   3rd Qu.: 7.98   3rd Qu.: 8.190  
 Max.   :14.0   Max.   :14.0   Max.   :14.0   Max.   :19   Max.   :10.840   Max.   :9.260   Max.   :12.74   Max.   :12.500  
```

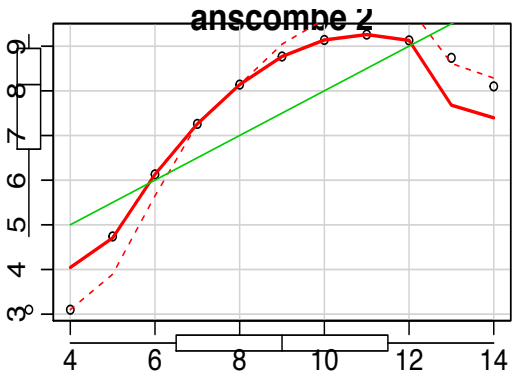
MODEL STATISTICS:

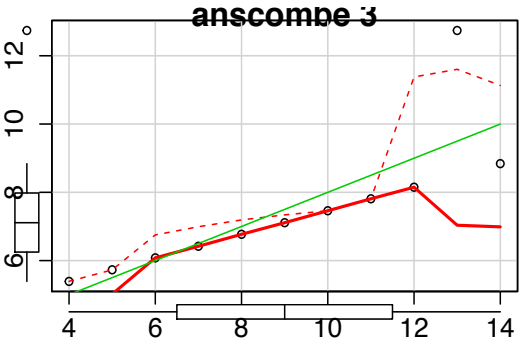
	b_0	b_1	$p-value$	R^2	R_a^2	MSE	SSE
1	3.0000909	0.5000909	0.00217 **	0.6665	0.6295	1.5292	13.763
2	3.000909	0.500000	0.002179 **	0.6662	0.6292	1.5307	13.776
3	3.0024545	0.4997273	0.00218 **	0.6663	0.6292	1.5285	13.756
4	3.0017273	0.4999091	0.00216 **	0.6667	0.6297	1.5269	13.742

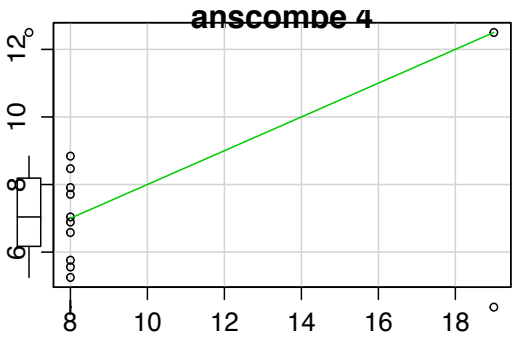
From the summary statistics of the data set, all the predictor variables range is more or less same. The observations fall median fall close to 7.5 – 8.0.

From the table, examining the p-value for b1 is significantly less than alpha=0.05 and imply that the predictor variable is statistically significant in the model. There is a linear association between x and y. Also R^2 is pretty high implies proportionate reduction in the total variation of Y associated with the use of predictor variable. R_a^2 does not mean anything in this SLR context. Also looking ta MSE they all fall in the same range 1.5. Looking at the scatter plots below we can arrive at following conclusions.

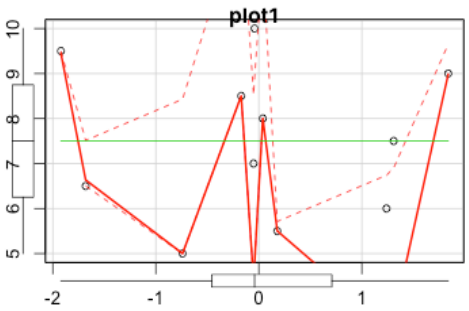


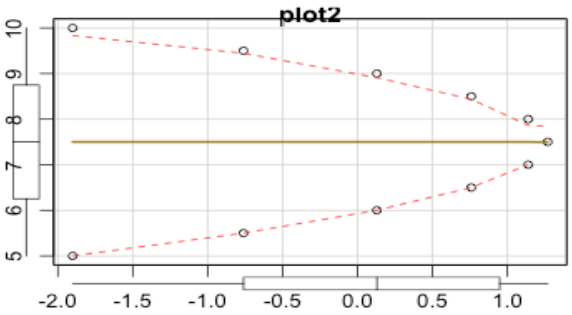


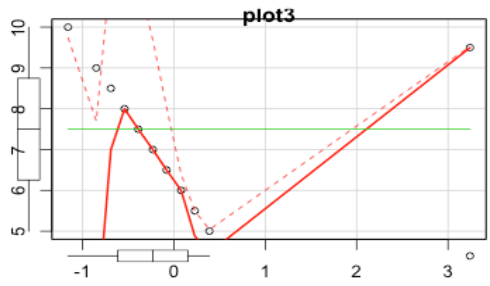


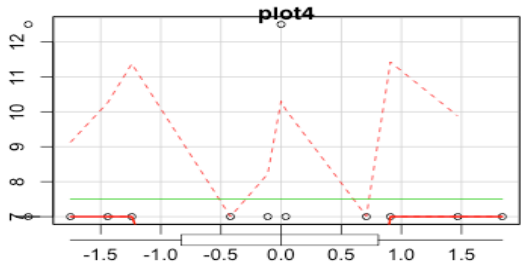


Residual plots:









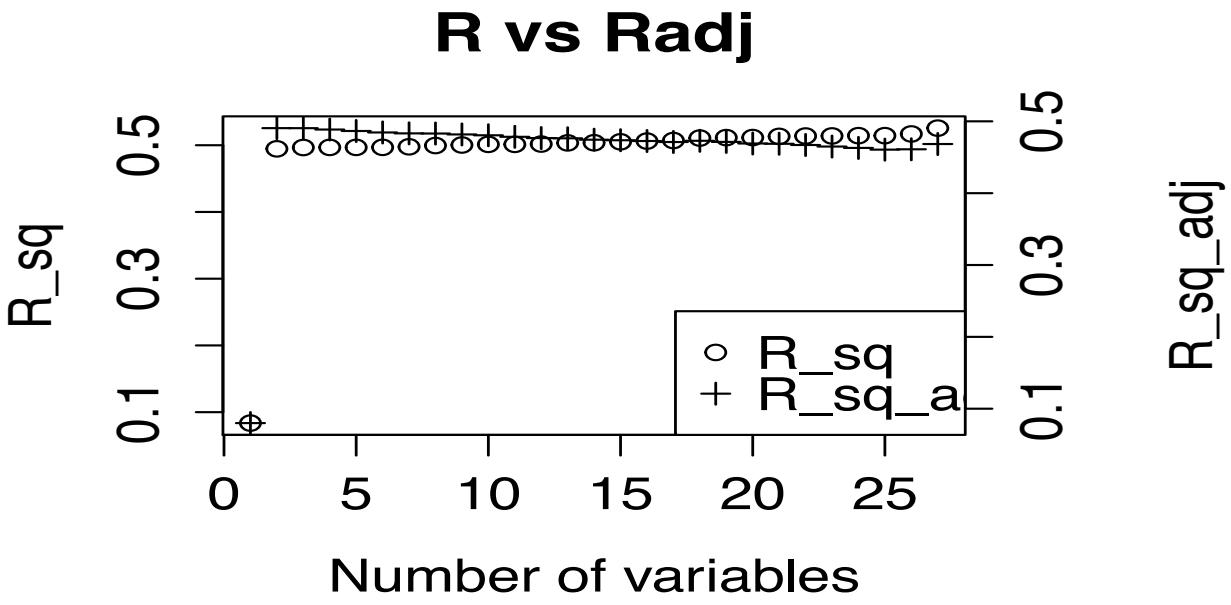
The plot1 shows that the fitted regression line passes through most of the values. The residual plot for ascombe 1 seems to be homoscedastic, and error terms are randomly distributed.

Plot for ascombe 2, shows that fitted regression line does not pass through most of the points, and looking at the residual plot, its heteroskedasticity.

Plot of ascombe 3 shows that the fitted regression line does not pass through most of the points, and there is a linear association between X and Y. The residual plot shows that the error terms change with fitted values. And they do not seem to be random.

Plot of ascombe 4 shows that the fitted regression line not passing through any of the points. There is no linear relationship between x and y. Also the residual plot shows error terms are evenly distributed.

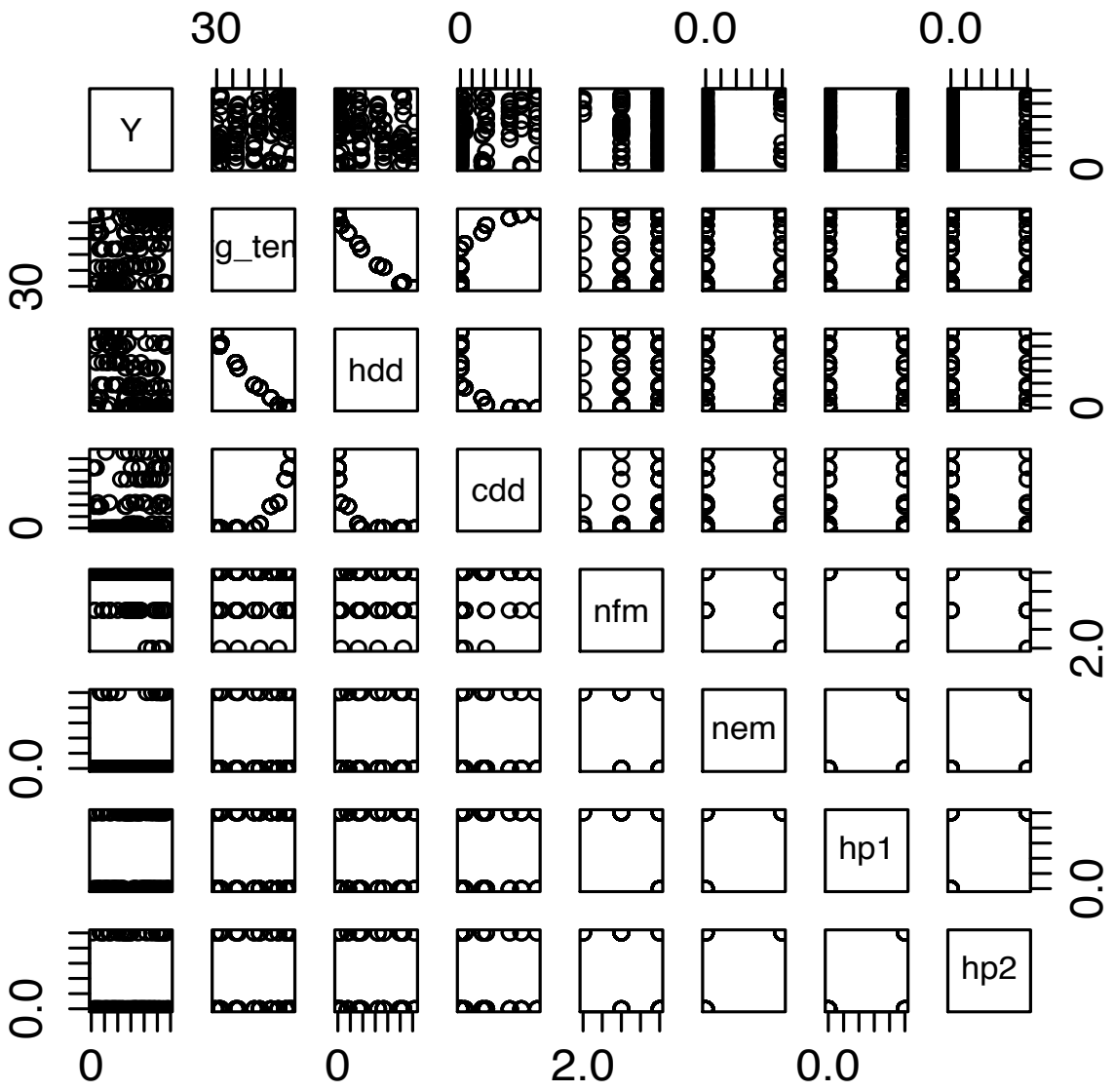
QUESTION 4: CODE FOR THIS IS IN THE APPENDIX.



The Rsquare here increase as number of variables, but r square adjusted decreases as the number of variables.

QUESTION 5:

From the model we overall F-test and observe the F-statistic value is 4.47. And the p-value (0.0002048) which is less than alpha level. Indicates reject null hypothesis, and conclude that there is relationship between predictors and the monthly electric bill.



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 323.22157  163.53655   1.976  0.05058 .
avg_temp     -2.07725    2.42389  -0.857  0.39330
hdd          -0.08611    0.07702  -1.118  0.26598
cdd           0.13234    0.09744   1.358  0.17718
nfm         -31.01015    9.35131  -3.316  0.00123 **
nem           0.21701   13.11322   0.017  0.98683
hp1         -15.46534    8.69137  -1.779  0.07791 .
hp2          -6.44783   11.63823  -0.554  0.58068
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.85 on 111 degrees of freedom
Multiple R-squared:  0.2199,    Adjusted R-squared:  0.1707
F-statistic:  4.47 on 7 and 111 DF,  p-value: 0.0002048
```

Interpret R2 for the model. Here the r square is less 0.2199. The 22% percentage change in y variable that can be explained by the use of predictor variables. Due to low R square the precise predictability of the model is low.

The p-value of the variable new electric meter is 0.98 which is very high compared to alpha =0.05. We accept null hypothesis, and conclude that new electric meter is not statistically significant from zero. We can remove this variable from the model, as it does not contribute to the variation in Y. The r squared value remains the same after removing new electric meter from the model.

Also the p-value of is high for new heat pump 2 is 0.58, which is greater than alpha=0.05 significance level. R square does not change much after removing this variable from the model. Hence its contribution is less.

P-value of new heat pump 1 is closer to to significance level alpha, may contribute to the prediction of Y.

The number of family members has a low p-value of 0.00123 ** which is less than significance level of 0.05. Hence it is statistically significant from zero. Removing this variable from the model reduces the R-squared value from 0.2199 to 0.1426. Hence its contribution to Y is

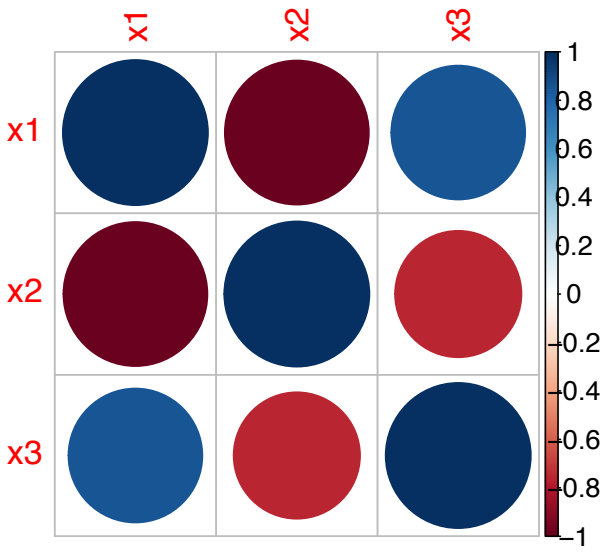
significant, its important to keep this value.

If the heating degree days was removed from the model the R-squared value reduces = 0.2111 from 0.2199. Which is not that a significant change. Also the p-value is 0.26598, which is greater than significance level. Conclude null hypothesis, that it is not statistically significant from zero. Its contribution to Y is less, when compared to number of family members.

1.correlation matrix between the following three variables:

The average temperature is highly co-related with heating degree days(HDD) -0.98, and average temperature is also highly correlated with cooling degree days(CDD) = 0.85. The matrix suggests that they need to be in the model. And the contribution of each HDD and CDD one is dependent on the time of the year. Also the HDD and CDD are strongly correlated to each other, which may affect the model in general.

```
> cor(dt_cor)
           x1         x2         x3
x1  1.0000000 -0.9846580  0.8515935
x2 -0.9846580  1.0000000 -0.7588285
x3  0.8515935 -0.7588285  1.0000000
```



2. The model **lm_electric <- lm(Y~.-avg_temp, data=df_e_bill)** after dropping the average temperature seems to make sense now. The contribution of the HDD becomes significantly, since its p-value is 0.04 less than significance level. Also the F-stat of the model has increased from the previous value, implying that the model is a better in predicting the electric bill amount.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 186.81394   37.49151   4.983  2.3e-06 ***
hdd          -0.02068    0.01012  -2.043  0.04340 *
cdd           0.05590    0.03919   1.426  0.15657
nfm         -30.78416    9.33651  -3.297  0.00131 **
nem           0.14591   13.09740   0.011  0.99113
hp1         -15.46931    8.68106  -1.782  0.07747 .
hp2          -6.23814   11.62185  -0.537  0.59250
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.81 on 112 degrees of freedom
Multiple R-squared:  0.2147,    Adjusted R-squared:  0.1727
F-statistic: 5.105 on 6 and 112 DF,  p-value: 0.0001143
```

VARIABLES IN THE MODEL	B1	B2	B3
AVERAGE TEMPERATURE	0.6628	-	-
HEATING DEGREE DAYS	0.25944	-0.01618	-
COOLING DEGREE DAYS	-1.78216	-0.07771	0.08875

Here we notice that the average temperature regression co-efficient changes sign. This indicates multicollinearity problem.

Based on the location of the collection of the data, which is Indianapolis, where the heating days and cooling days are more or less the same, they drive the electric bill. Where as in colder places the heating days are more significant than cooling days. Hence considering geographical location on modeling of the electric bill variable, would be helpful.

Since its highly correlated with HDD and CDD. After removing average temperature from the regression model, we noticed that removing CDD drops the R-square to 0.2005. Where as dropping HDD drops R-square significantly to 0.1855, implying that HDD impacts the response variable more than CDD. Also the p-value of the HDD = 0.04340 * is significantly less than CDD = 0.15657. Hence HDD is better variable to

QUESTION 6

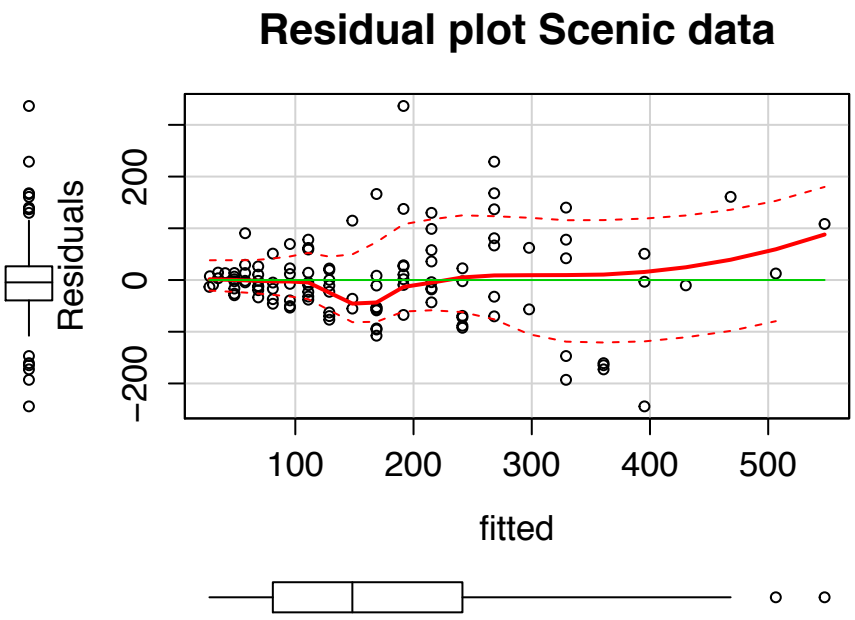
```
Call:
lm(formula = nn ~ afs + I(afs^2), data = df_sd)

Residuals:
    Min      1Q  Median      3Q     Max
-244.32  -39.42   -4.55   26.48  336.48

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.54823   51.41432   0.653  0.51544
afs         -1.66613    2.43463  -0.684  0.49519
I(afs^2)     0.10116    0.02723   3.716  0.00032 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82.31 on 110 degrees of freedom
Multiple R-squared:  0.6569,    Adjusted R-squared:  0.6507
F-statistic: 105.3 on 2 and 110 DF,  p-value: < 2.2e-16
```

Fit the second-order regression model. Plot the residuals against the fitted values. How well does the second-order model appear to fit the data?



The second order model significantly fits the data, well. This is observed in the p-value of co-efficient of the squared term. It is significantly less than alpha, supporting the alternate hypothesis that $\beta_2 \neq 0$.And it is significantly different from 0, hence this predictor variable affects the response variable. The error terms distribution in the residual plot indicates they are heteroscedastic, and also presence of outliers.

The R squared value for the second order regression model is 0.6569396.
The R squared value for the first order regression model is 0.6138809.

These values differ, also comparing the adjusted R-squared values for the 2 cases, show second order model is 0.6507022, first order model is 0.6104023. There is an increase in adjusted r-square, and r-square indicating that the addition of the quadratic term has improved effect on the prediction of the response variable.

The co-efficient of partial determination $R^2_{y2|1}$ measures the marginal contribution of the new variable added to the linear regression model.

$$R^2_{y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, \quad X_2)$$

Using R these are calculated to be 0.1115168, is marginal contribution of adding the quadratic term to the model.

We perform the Ramsey reset test to check whether the quadratic term can be dropped from the system. Here

$$H_0 : \alpha_1 = \ldots = \alpha_k = 0$$

$$H_a : not\ all\ \alpha_k = 0$$

R-code to run the test is given by

```
resettest(lm_sd_2, power=2, type="fitted")
```

R output:

```
> resettest(lm_sd_2, power=2, type="fitted")

        RESET test

data:  lm_sd_2
RESET = 0.45606, df1 = 1, df2 = 109, p-value = 0.5009

>
```

The p-value is significantly higher indicating that quadratic term needs to be included.

QUESTION 7

The co-efficient of the predictor variables of the multiple linear regression model are are represented by a matrix X_{nxp}. The X matrix contains column of 1 ‘s as well as column of n observations for each p-1 X variables in the model.
The least squares estimators are given by

$$b = (X'X)^{-1}\ (X'X)Y$$

If the columns are linearly independent, then the inverse of the matric exists. And the matrix has a complete rank. If the columns are co-related, they are linearly dependent, the the matrix is does not have a full rank, hence no solution exists, hence it is a singular matrix. Multicollinearity exists between predictors of the regression model.

R code for question 4

```
#############################################################################
#reading data from data source
d <- read.csv('extraColumnsOfRandomData.csv', header = TRUE)
x_num <- c()
R_sq <- c()
```

```
R_sq_adj <- c()
# looping all the elements to calculate the r_sq and rsq_adj for each
# model
for(i in 2:(ncol(d))){
  df <- d[, c(1:i)]
  lm_mod<- lm(df$BODYFAT~., data=df)
  x_num <- append(x_num, ncol(df)-1)
  R_sq <- append(R_sq , summary(lm_mod)$r.squared) #0.6569396
  R_sq_adj<- append(R_sq_adj, summary(lm_mod)$adj.r.squared)
}
#data frame with x, r_sq, r_sq_adj
df1 <- data.frame(x_num, R_sq, R_sq_adj)
#plotting code
opar <- par(mar=c(5,4,4,5)+0.1)
plot(df1$x_num,df1$R_sq, xlab = "Number of variables",ylab = "R_sq",
     main = "Scatter Plot")
par(new=T)
plot(df1$x_num,df1$R_sq_adj,pch=3, axes = F,ylab="",xlab="")
axis(side=4)
mtext(side=4,line=3.8,"R_sq_adj")
#legend("bottomright", legend=c("R_sq","R_sq_adj"), pch=c(1,3))opar <-
par(mar=c(5,4,4,5)+0.1)
plot(df1$x_num,df1$R_sq, xlab = "Number of variables",ylab = "R_sq",main =
"R vs Radj")
par(new=T)
plot(df1$x_num,df1$R_sq_adj,pch=3, axes = F,ylab="",xlab="")
axis(side=4)
mtext(side=4,line=3.8,"R_sq_adj")
legend("bottomright", legend=c("R_sq","R_sq_adj"), pch=c(1,3))
##############################################################################
```