STAT 223: Project 2

Nrepesh Joshi

Professor Kevin Hastings

Applied Analytics

4/21/20

Medical expenses using multiple regression

Use historical data on individuals to predict medical expenses.

=============================

BRIEF DESCRIPTION OF THE DATA

=============================

The data set has six potential predictor variables for a total of 1338 individuals. Three variables are numerical and three are categorical. The numerical variables are the age of the policy holder, his or her bmi (body mass index), and the number of children of the policy holder. Sex, whether the policy holder is a smoker, and geographical region are the categorical predictors. The response is calendar year total medical expenses for the covered members of the policy holder's family.

========================

HOW TO USE THE DATA FILE

========================

The data file is in csv format, under the name “insurance.csv”. Each of the 1338 rows after the row containing variable names has the information on a policy holder that is described above, so there are six predictor columns followed by the response column. There are no missing values. The variable names are:

age......integer values.

sex........character values, “female” or “male”. (R may consider this as a factor variable, with codes 1 and 2 if you use read.csv to input the data)

bmi...........numerical values.

children.........integer values.

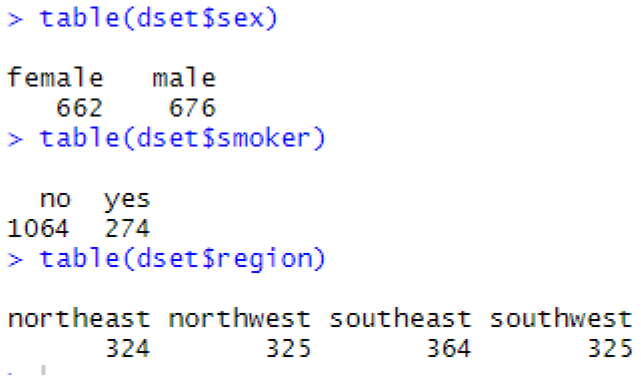
smoker.......character values, “no” or “yes”. (R may consider this as a factor variable, with codes 1 and 2 respectively for “no” and “yes”)

region........character values “northeast”,”northwest”,”southeast”,”southwest”. (As a factor, the codes for these will be alphabetized 1,2,3,4 respectively)

expenses.........numerical values

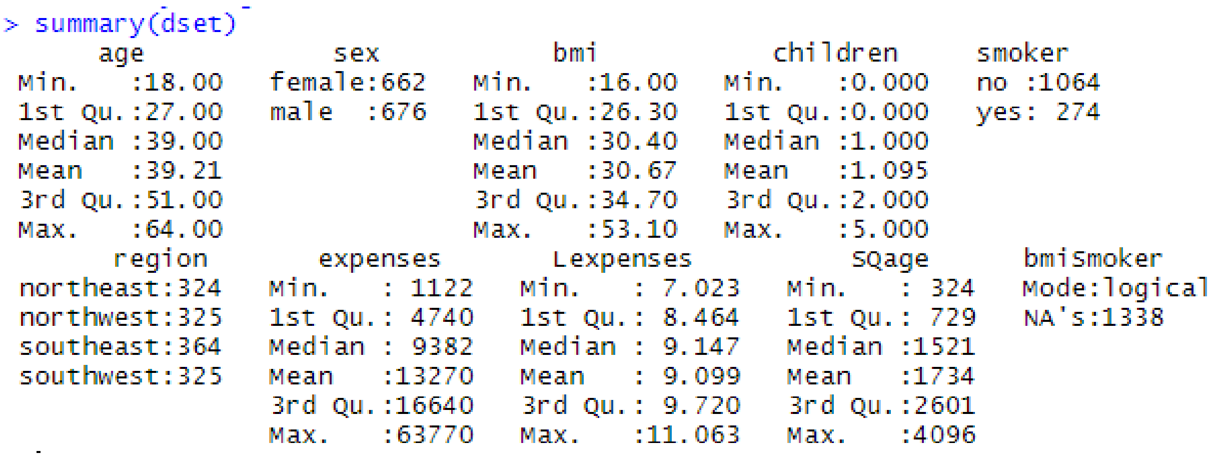
The data file is “insurance.csv”.

1. Understanding the variables. Use R’s table() command to get a frequency distribution of each of the categorical variables and summary() to get basic statistics on the integer and numeric variables. Produce histograms to decide whether the predictor bmi and the response expenses appear to be normally distributed. If you set up a new variable for the log of expenses, does that appear to be more normal?



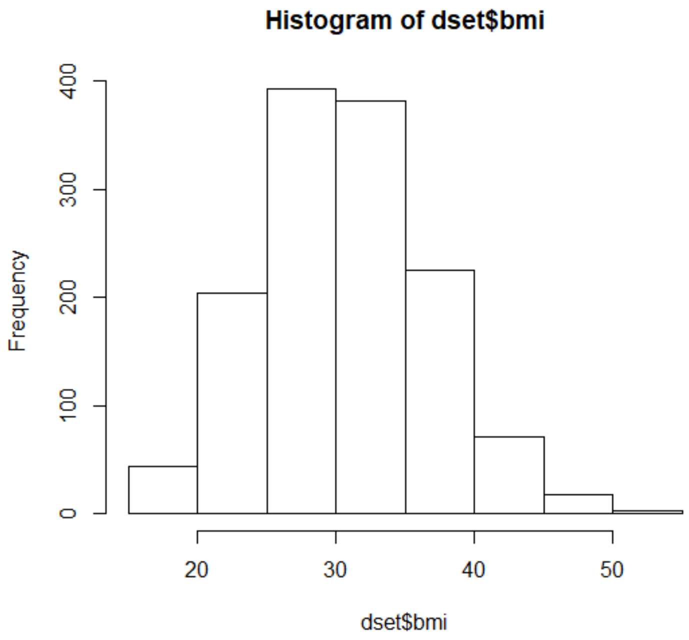
```
> table(dset$sex)

female   male 
   662    676 
> table(dset$smoker)

  no  yes 
1064  274 
> table(dset$region)

northeast northwest southeast southwest 
      324       325       364       325 
```

We notice from above that the dataset had 662 females and 676 males. 1064 were non-smokers and 274 were smokers. We also see about equal distribution in the regions.

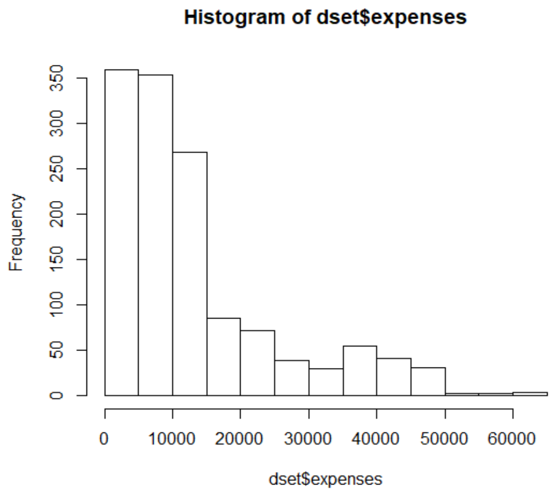


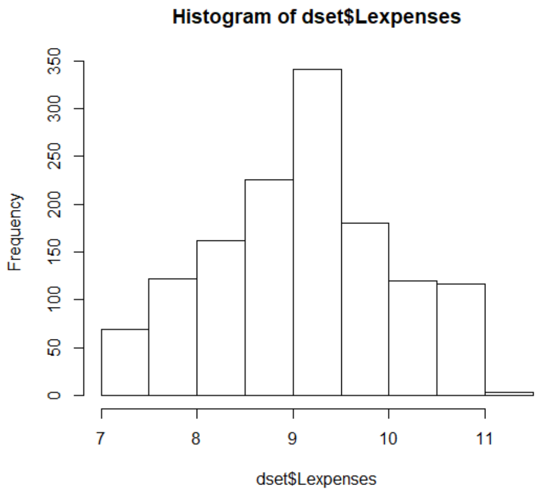
```
> summary(dset)
      age            sex           bmi           children       smoker    
 Min.   :18.00   female:662   Min.   :16.00   Min.   :0.000   no :1064  
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274  
 Median :39.00                Median :30.40   Median :1.000             
 Mean   :39.21                Mean   :30.67   Mean   :1.095             
 3rd Qu.:51.00                3rd Qu.:34.70   3rd Qu.:2.000             
 Max.   :64.00                Max.   :53.10   Max.   :5.000             
       region       expenses        Lexpenses          SQage       bmiSmoker     
 northeast:324   Min.   : 1122   Min.   : 7.023   Min.   : 324   Mode:logical  
 northwest:325   1st Qu.: 4740   1st Qu.: 8.464   1st Qu.: 729   NA's:1338     
 southeast:364   Median : 9382   Median : 9.147   Median :1521                 
 southwest:325   Mean   :13270   Mean   : 9.099   Mean   :1734                 
                 3rd Qu.:16640   3rd Qu.: 9.720   3rd Qu.:2601                 
                 Max.   :63770   Max.   :11.063   Max.   :4096                 
```

The summary function is very useful as it tells us min/max and quartiles information of integer and numeric values in the datset.



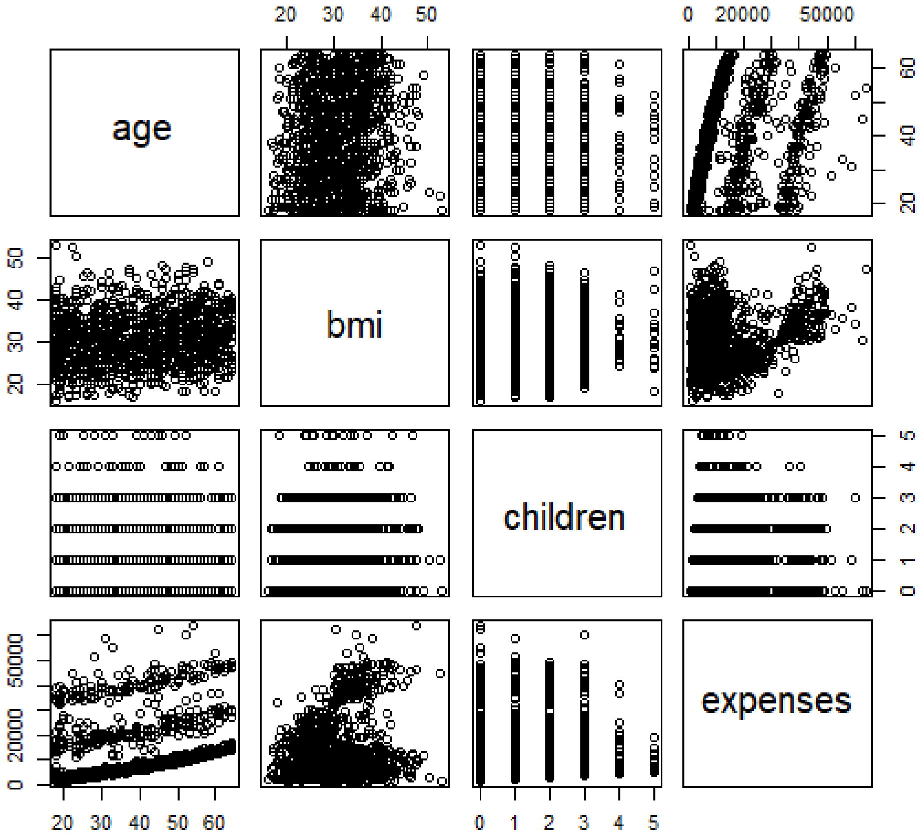
The histogram plot of predictor bmi appears normal.



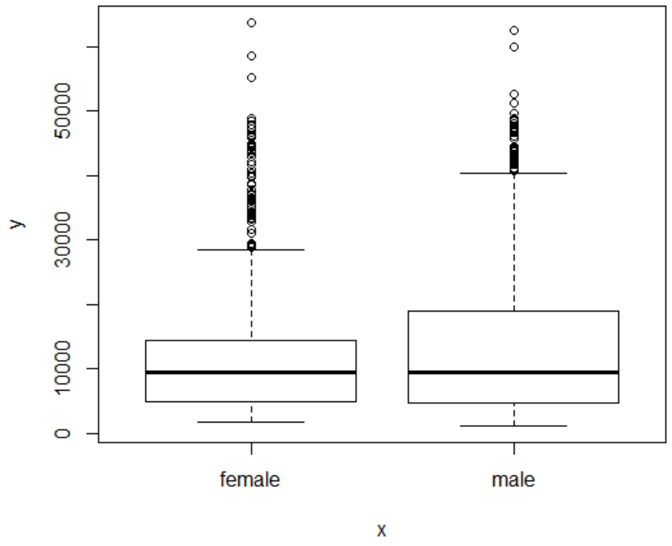


The histogram plot of response expenses alone did not appear to be normally distributed but when we set up a new variable by taking the log of expenses, the data appears to be more normal which we will further use for our analysis.

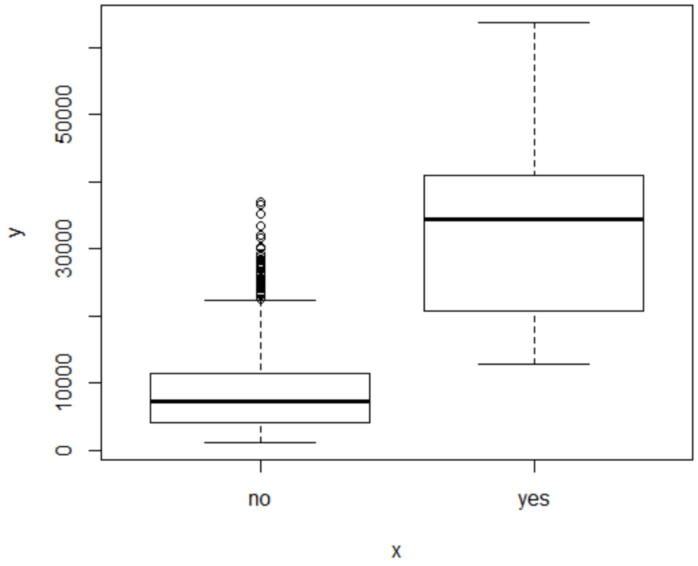
2. Understanding the relationships. Use the pairs(insurance[c(“age”,”bmi”,”children”,”expenses”)]) command in R to produce a scatterplot matrix of the integer and numeric variables and note any features that stand out. What do you get when you apply the plot(varname,expenses) command to the categorical variables? Try both again with the logged expenses.



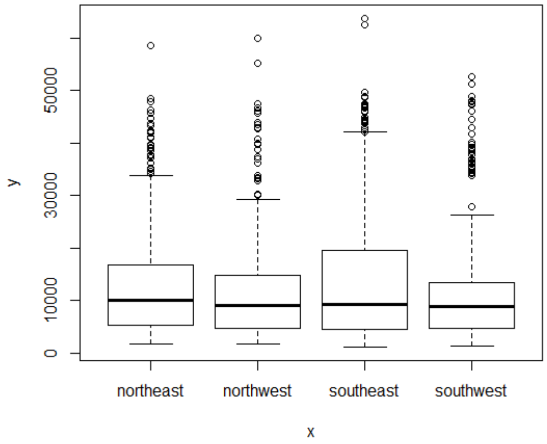
The pair plot above is for all categorical variables and expenses. Most of the plots do not give us a pattern or correlation to work with. However, the plot expenses vs age shows a linear correlation with each other. This positive relationship entails that the older the policyholder, the more they spend. One other thing that stands out about this scatterplot is that there’s slight curvature. In addition to the curves, it seems as though the data is being divided into groups. This particular influence on the shape of the graph might be a result of the categorical variables in our insurance dataset.



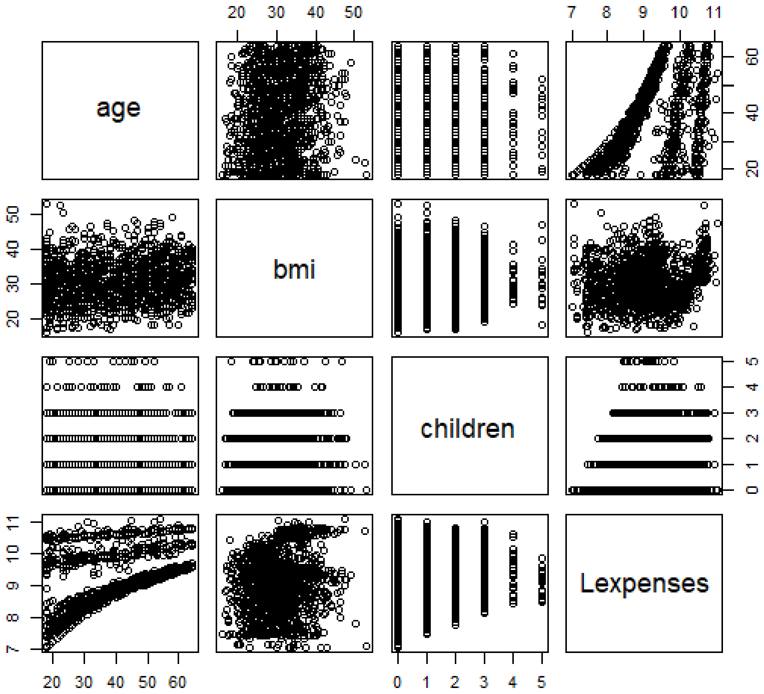
The middle values for both female and male policyholders are about $10,000. In general, both genders seem to have the same spread, except for the fact the males have a higher 75th percentile value. The distributions of expenses for both female and male policyholder are heavily skewed to the right.



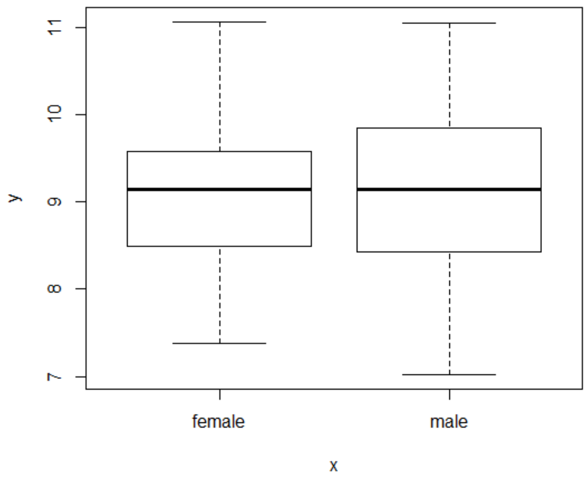
As for the box plot of the smoker variable, and the expense, we observe that the median for smokers is well above than that of non-smokers. Even the minimum value of the smokers is higher than the third quartile of non-smokers. In general, we would conclude that smokers have a much wider spread than non-smokers. We also notice that the distribution of the policyholder’s expenses for non-smokers is affected by large outliers.



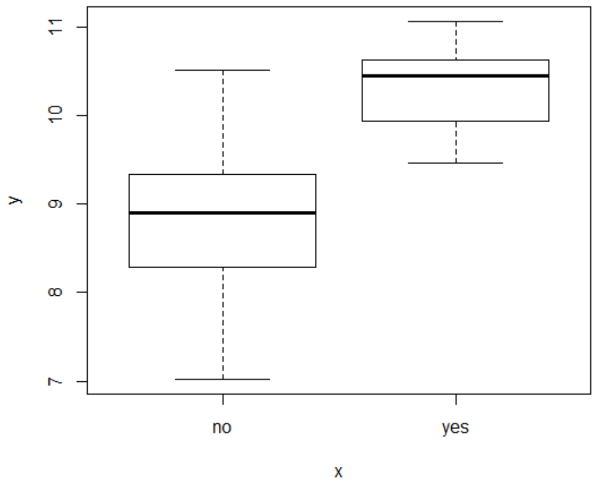
Unlike the previous two variables, it appears that all four regions have about the same spread. The distributions of expenses for all four regions, however, are once again influenced by data points that lie beyond the “whiskers” of the box plots. One way to fix this problem is to transform our response variable using a log function.



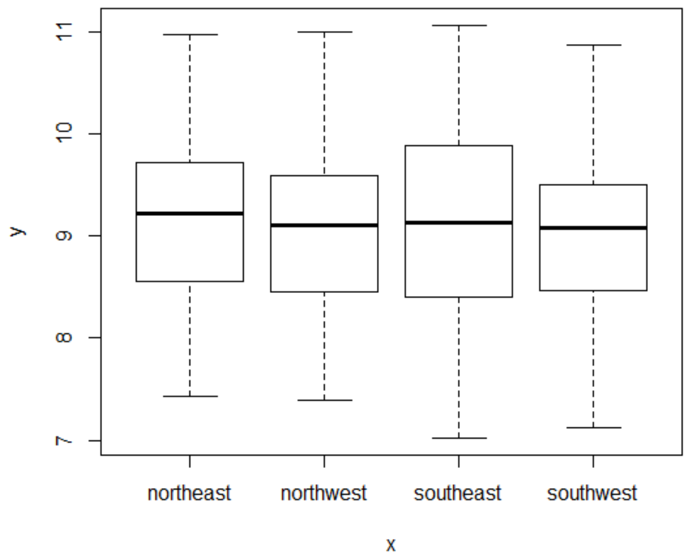
We use the pairs command again. The pair plot above is for all categorical variables and log of the expenses. Most of the plots do not give us a pattern or correlation to work with. However, the plot log expenses vs age shows a stronger linear correlation with each other.



The box and whisker plot tells us how the data is distributed among the sex categorical variables. The distribution now is now at a more suitable data range.

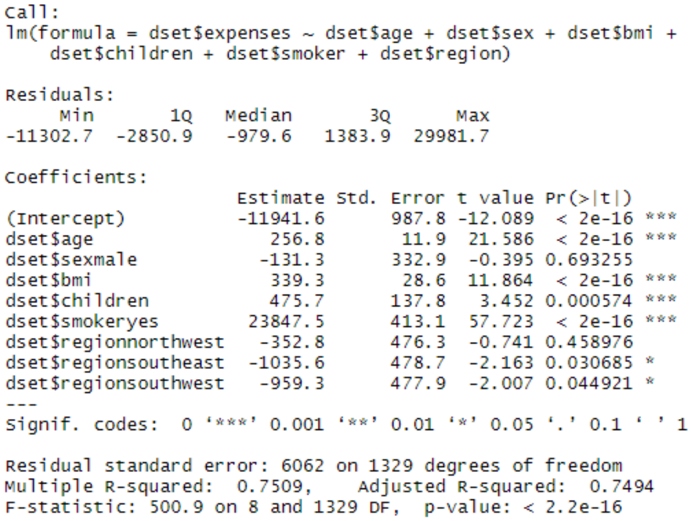


The box and whisker plot tells us how the data is distributed among the smoker categorical variables. ‘Yes’ if the subject smokes and ‘no’ if not. The distribution now is now at a more suitable data range.



The box and whisker plot tells us how the data is distributed among the region categorical variables. The distribution now is now at a more suitable data range.

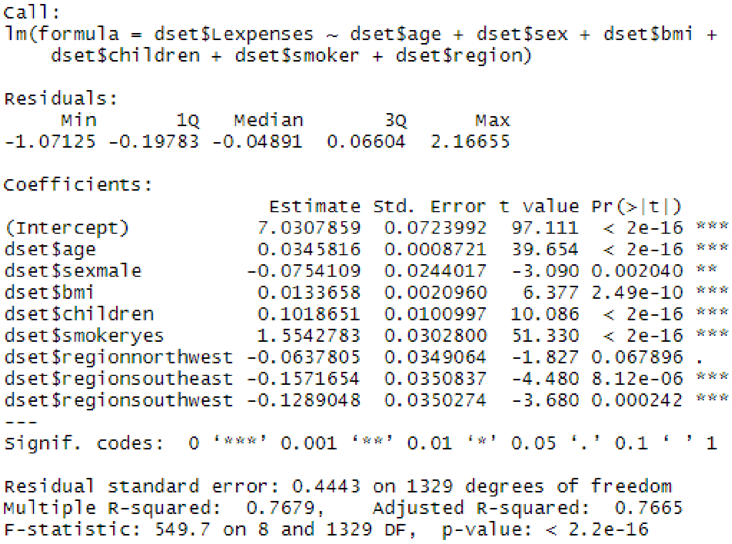
3. The regression. Now run a regression of the expenses on the full set of variables. Interpret the results. Why do you see more than 6 predictors? (Hint: remember the concept of the indicator variable.) Compare your results to a regression of the logged expenses on the variables. Which seems better, in terms of both the significance of the predictors, the R2 value, and the distribution of the residuals?



```
Call:
lm(formula = dset$expenses ~ dset$age + dset$sex + dset$bmi +
    dset$children + dset$smoker + dset$region)

Residuals:
     Min       1Q   Median       3Q      Max
-11302.7  -2850.9   -979.6   1383.9  29981.7

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -11941.6      987.8 -12.089  < 2e-16 ***
dset$age                256.8       11.9  21.586  < 2e-16 ***
dset$sexmale           -131.3      332.9  -0.395 0.693255
dset$bmi                339.3       28.6  11.864  < 2e-16 ***
dset$children           475.7      137.8   3.452 0.000574 ***
dset$smokeryes        23847.5      413.1  57.723  < 2e-16 ***
dset$regionnorthwest   -352.8      476.3  -0.741 0.458976
dset$regionsoutheast  -1035.6      478.7  -2.163 0.030685 *
dset$regionsouthwest   -959.3      477.9  -2.007 0.044921 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,	Adjusted R-squared:  0.7494
F-statistic: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

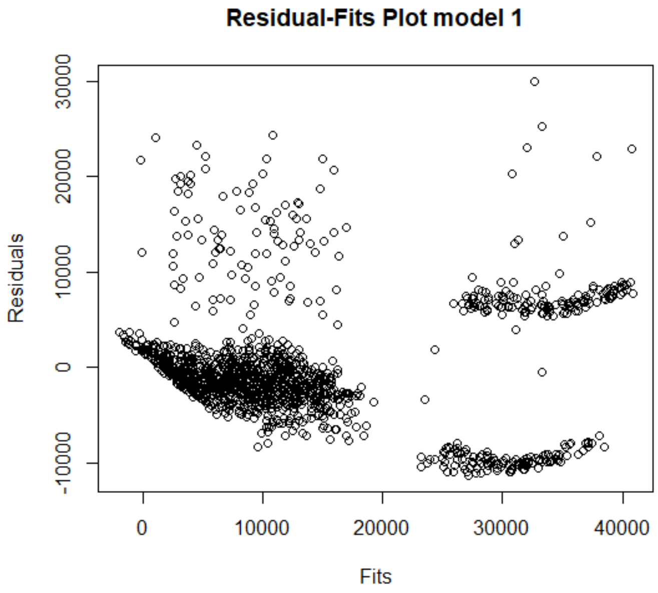
From the model summary we see that the R-squared value is 0.75 which is very close to 1 with the p-value of less than 10^-16 which is highly significant. This means about 75% of the variation can be explained by our model. We should notice however that a few predictors do not have significant p-values for example: sex male.

We see more than 6 predictor variables because of the concept of dummy variables. We know that if a categorical variable has n different values, you only need n-1 dummy variables to indicate which category is active. So, in the multiple regression equation, if the sex value is 0 then it means female and 1 if male. Same thing with the Smoker predictor. For the region category, if all the values are 0 then it will indicate 'Northeast'. It is just a way to account for both variables to influence the regression model.

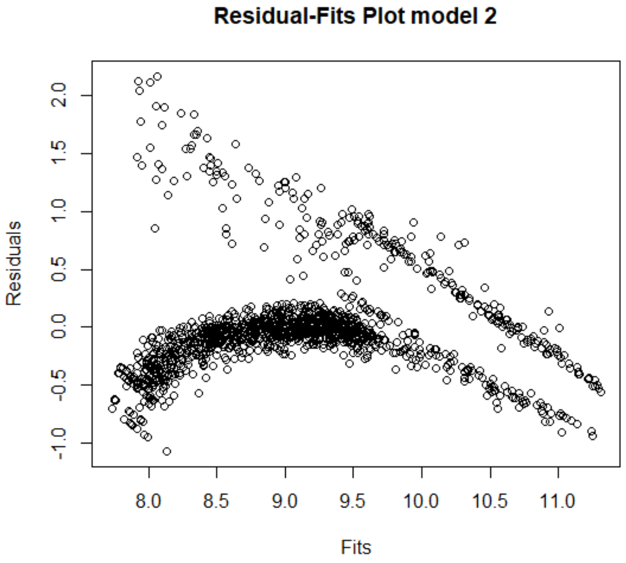


```
Call:
lm(formula = dset$Lexpenses ~ dset$age + dset$sex + dset$bmi + 
    dset$children + dset$smoker + dset$region)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.07125 -0.19783 -0.04891  0.06604  2.16655 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          7.0307859  0.0723992  97.111  < 2e-16 ***
dset$age             0.0345816  0.0008721  39.654  < 2e-16 ***
dset$sexmale        -0.0754109  0.0244017  -3.090 0.002040 ** 
dset$bmi             0.0133658  0.0020960   6.377 2.49e-10 ***
dset$children        0.1018651  0.0100997  10.086  < 2e-16 ***
dset$smokeryes       1.5542783  0.0302800  51.330  < 2e-16 ***
dset$regionnorthwest -0.0637805  0.0349064  -1.827 0.067896 .  
dset$regionsoutheast -0.1571654  0.0350837  -4.480 8.12e-06 ***
dset$regionsouthwest -0.1289048  0.0350274  -3.680 0.000242 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4443 on 1329 degrees of freedom
Multiple R-squared:  0.7679,	Adjusted R-squared:  0.7665 
F-statistic: 549.7 on 8 and 1329 DF,  p-value: < 2.2e-16
```

The model above uses the logged expenses on the variables. This model does better for a various reason. Firstly, we notice that all the p-values for the predictors are less than 0.05 which says that the model is highly significant considering all predictors. The R-squared value is 0.77 which is better than our previous model and is close to 1 so we can say that the model is doing a good job with the linear regression. This means about 77% of the variation can be explained by our model.



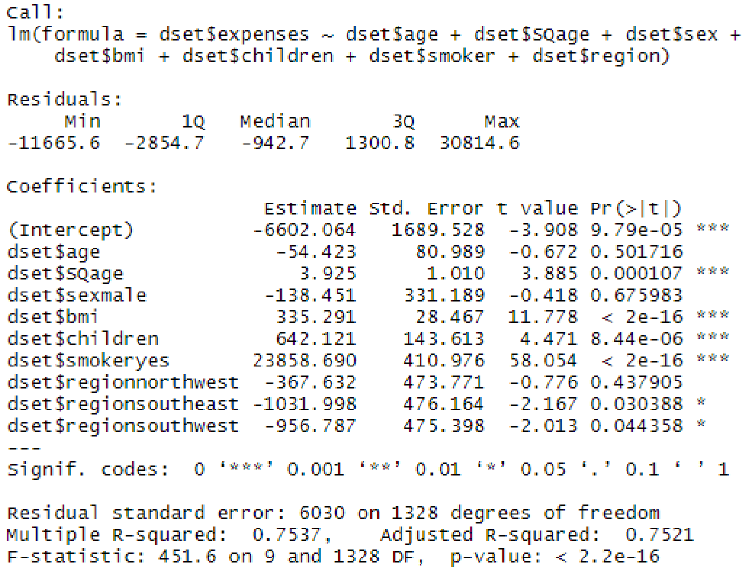
We see here that when residuals are plotted with the fitted y values in the expenses plot, the linearity assumption is not met because we are seeing a pattern of a group of points and two lines that are formed. We also notice that the variation is not constant as it varies from the start of the graph and to the end of the graph.



We see here that when residuals are plotted with the fitted y values in the log expenses plot, the linearity assumption is not met because we are seeing a curved pattern of a group of points and a linear straight line. This means that one of our predictors could be improved by squaring each values of the predictor (Question 4). We also notice that the variation is not constant as it varies from the start of the graph and to the end of the graph.

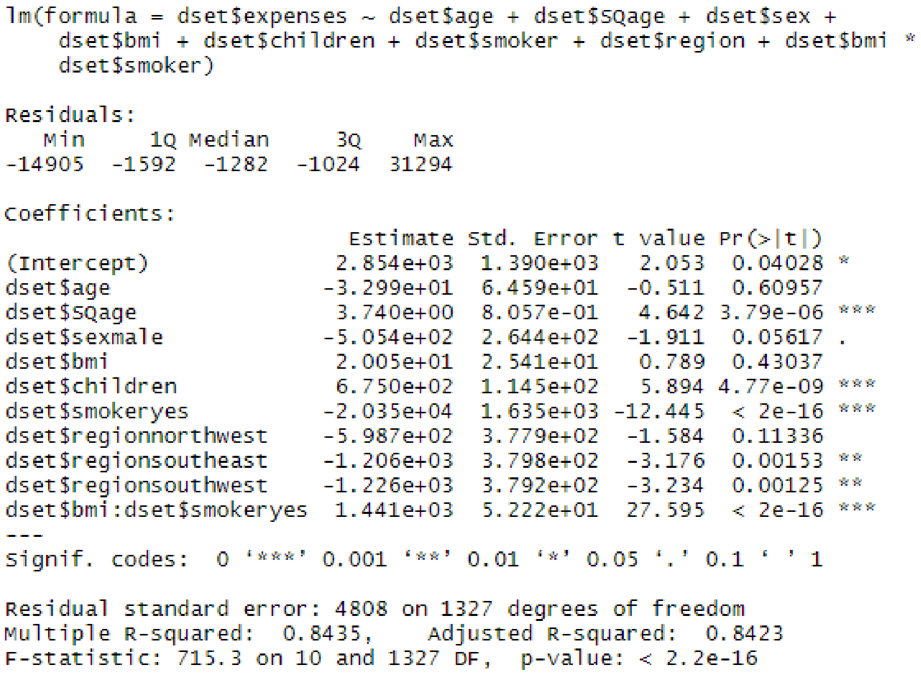
4. Improving the model. Returning to the original response variable expenses (not the logs), you may have noticed a slight curvature in the scatterplot of expenses against age. Create a new variable that is the square of the age and rerun the regression with it included. If the quadratic term is significant, keep it in the model and then try adding an interaction term between the bmi and smoker variables. (In R this would be a term bmi*smoker in the model equation). Do either of these changes or both seem to result in better prediction?

After we create a new variable of square of age we rerun the regression.



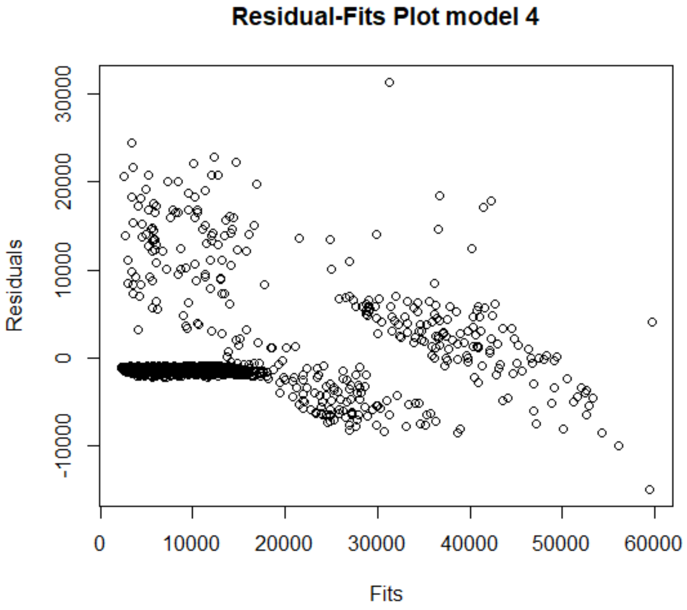
```
Call:
lm(formula = dset$expenses ~ dset$age + dset$SQage + dset$sex +
    dset$bmi + dset$children + dset$smoker + dset$region)

Residuals:
     Min       1Q   Median       3Q      Max
-11665.6  -2854.7   -942.7   1300.8  30814.6

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -6602.064   1689.528  -3.908 9.79e-05 ***
dset$age               -54.423     80.989  -0.672 0.501716
dset$SQage               3.925      1.010   3.885 0.000107 ***
dset$sexmale          -138.451    331.189  -0.418 0.675983
dset$bmi               335.291     28.467  11.778  < 2e-16 ***
dset$children          642.121    143.613   4.471 8.44e-06 ***
dset$smokeryes       23858.690    410.976  58.054  < 2e-16 ***
dset$regionnorthwest  -367.632    473.771  -0.776 0.437905
dset$regionsoutheast -1031.998    476.164  -2.167 0.030388 *
dset$regionsouthwest  -956.787    475.398  -2.013 0.044358 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6030 on 1328 degrees of freedom
Multiple R-squared:  0.7537,    Adjusted R-squared:  0.7521
F-statistic: 451.6 on 9 and 1328 DF,  p-value: < 2.2e-16
```

We see here that the p-value is not significant for just the age predictor as it has a value of 0.502. However, we notice that the square age gives us a p-value of 0.0001 which is highly significant. The R-square is doing just as well like the previous models with 75% of the variations explained. Due to this we decide to keep the new age-squared variable in our model.



```
lm(formula = dset$expenses ~ dset$age + dset$SQage + dset$sex + 
    dset$bmi + dset$children + dset$smoker + dset$region + dset$bmi * 
    dset$smoker)

Residuals:
   Min     1Q Median     3Q    Max 
-14905  -1592  -1282  -1024  31294 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              2.854e+03  1.390e+03   2.053  0.04028 *  
dset$age                -3.299e+01  6.459e+01  -0.511  0.60957    
dset$SQage               3.740e+00  8.057e-01   4.642 3.79e-06 ***
dset$sexmale            -5.054e+02  2.644e+02  -1.911  0.05617 .  
dset$bmi                 2.005e+01  2.541e+01   0.789  0.43037    
dset$children            6.750e+02  1.145e+02   5.894 4.77e-09 ***
dset$smokeryes          -2.035e+04  1.635e+03 -12.445  < 2e-16 ***
dset$regionnorthwest    -5.987e+02  3.779e+02  -1.584  0.11336    
dset$regionsoutheast    -1.206e+03  3.798e+02  -3.176  0.00153 ** 
dset$regionsouthwest    -1.226e+03  3.792e+02  -3.234  0.00125 ** 
dset$bmi:dset$smokeryes  1.441e+03  5.222e+01  27.595  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4808 on 1327 degrees of freedom
Multiple R-squared:  0.8435,	Adjusted R-squared:  0.8423 
F-statistic: 715.3 on 10 and 1327 DF,  p-value: < 2.2e-16
```

After we add our interaction term, we notice the following: The age-squared predictor is still highly significant, and our interaction term is also highly significant with a p-value of 10^-16. This means that we cannot exclude both bmi and smoker from our model. The R-squared value is very good as we can explain up to 84% of the variation from the model.



When we plot a residual-fits for this model we notice that the pattern has disappeared which tells us that the linearity assumptions are met. The points are scattered, and the variation is fairly equal.

Code:

```
#----------------------------------------------------
# Importing data

dset <- read.csv("insurance.csv", header = TRUE);
head(dset)

#----------------------------------------------------
# Part 1

# Frequency distribution of each of the categorial variables

table(dset$sex)
table(dset$smoker)
table(dset$region)

# Summary to get basic statistics
summary(dset)

# Histogram to check normal distribution
hist(dset$bmi)
hist(dset$expenses)

# Log of expences as a new variable
dset$Lexpenses <- log(dset$expenses); View(dset)
hist(dset$Lexpenses)

#----------------------------------------------------
# Part 2

# scatter plot matric of numeric variables
pairs(dset[c("age","bmi","children","expenses")])

# Plot of categorical variables with expences
```

```
plot(dset$sex,dset$expenses)
plot(dset$smoker,dset$expenses)
plot(dset$region,dset$expenses)


# Repeat for log expences
pairs(dset[c("age","bmi","children","Lexpenses")])

plot(dset$sex,dset$Lexpenses)
plot(dset$smoker,dset$Lexpenses)
plot(dset$region,dset$Lexpenses)

#-----------------------------------------------------
# Part 3

#Multiple regression

model1 <- lm(dset$expenses~dset$age+dset$sex+dset$bmi+dset$children+dset$smoker+dset$region);summary(model1)
model2 <- lm(dset$Lexpenses~dset$age+dset$sex+dset$bmi+dset$children+dset$smoker+dset$region);summary(model2)
anova(model1)
anova(model2)

fits1 <- fitted(model1)
resids1 <- residuals(model1)
plot(fits1, resids1, main = "Residual-Fits Plot model 1", xlab = "Fits", ylab = "Residuals")

fits2 <- fitted(model2)
resids2 <- residuals(model2)
plot(fits2, resids2, main = "Residual-Fits Plot model 2", xlab = "Fits", ylab = "Residuals")
```

```
#----------------------------------------------------
# Part 4

dset$SQage <- dset$age^2; View(dset)

model3 <-
lm(dset$expenses~dset$age+dset$SQage+dset$sex+dset$bmi+dset$children+dset$smoker+dset$region);summary(model3)
summary(model3)
model4 <-
lm(dset$expenses~dset$age+dset$SQage+dset$sex+dset$bmi+dset$children+dset$smoker+dset$region+dset$bmi*dset$smoker);summary(model4)
summary(model4)

fits3 <- fitted(model3)
resids3 <- residuals(model3)
plot(fits3, resids3, main = "Residual-Fits Plot model 3", xlab = "Fits", ylab = "Residuals")


fits4 <- fitted(model4)
resids4 <- residuals(model4)
plot(fits4, resids4, main = "Residual-Fits Plot model 4", xlab = "Fits", ylab = "Residuals")
```