Predicting Medical Expenses Using Multiple Regression

Project Summary

The goal is simple: we would like to be able to use historical data on individuals to predict medical expenses. This is certainly the kind of thing that insurance companies would be concerned about in setting rates, or governmental bodies would like to know in order to assess the economic impact of changes to health insurance programs like Medicare and Medicaid.

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 b.m.i (body mass index), and the number of children of the policy holder. Sex, whether or not 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.

Data Processing

#1

```
> dataframe <- read.csv("Desktop/insurance.csv")
```

Frequency distribution of each of the categorical variable (sex, smoker, and region):

```
> table(dataframe$sex)
female   male
   662    676

> table(dataframe$smoker)
  no  yes
1064  274

> table(dataframe$region)
northeast northwest southeast southwest
      324       325       364       325
```

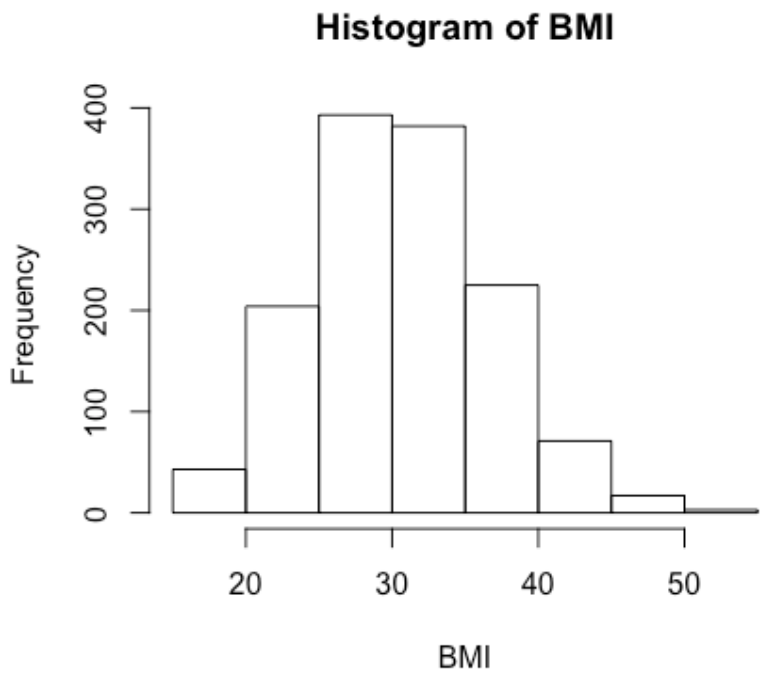
Basic statistics of age, bmi, children, and expenses variable:

```
> summary(dataframe$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

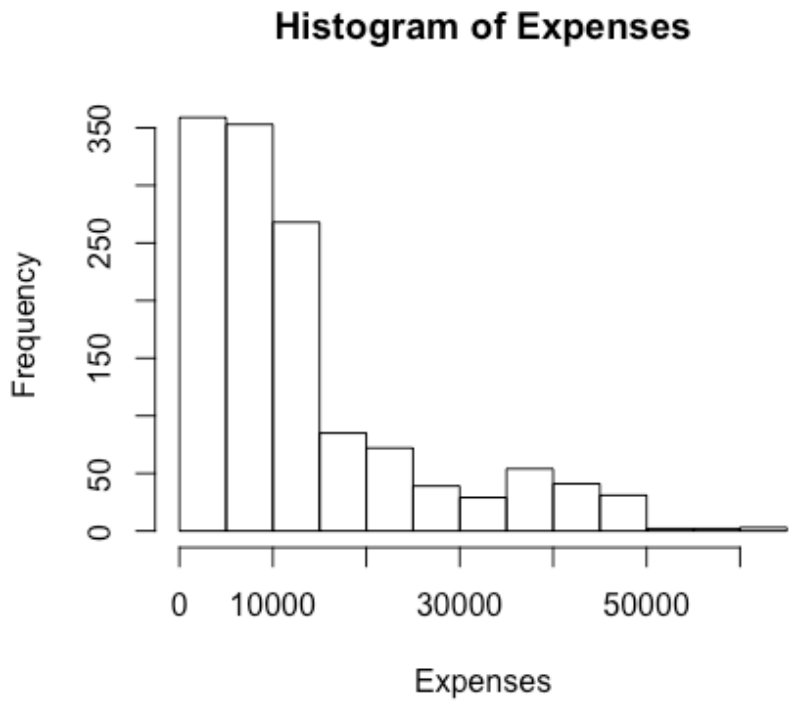
```
  18.00   27.00   39.00   39.21   51.00   64.00
> summary(dataframe$bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   26.30   30.40   30.67   34.70   53.10
> summary(dataframe$children)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   1.095   2.000   5.000
> summary(dataframe$expenses)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1122    4740    9382   13270   16640   63770
```

Histogram of BMI and expenses:

```
> hist(dataframe$bmi,main="Histogram of BMI",xlab="BMI")
```




```
> hist(dataframe$expenses,main="Histogram of Expenses", xlab="Expenses")
```



Looking at the above two histograms, we see that predictor BMI appears to be normally distributed, while response expenses does not and its data tends to skew right.

The log of expenses, however, does appear to be more normally distributed, as shown in its histogram below:

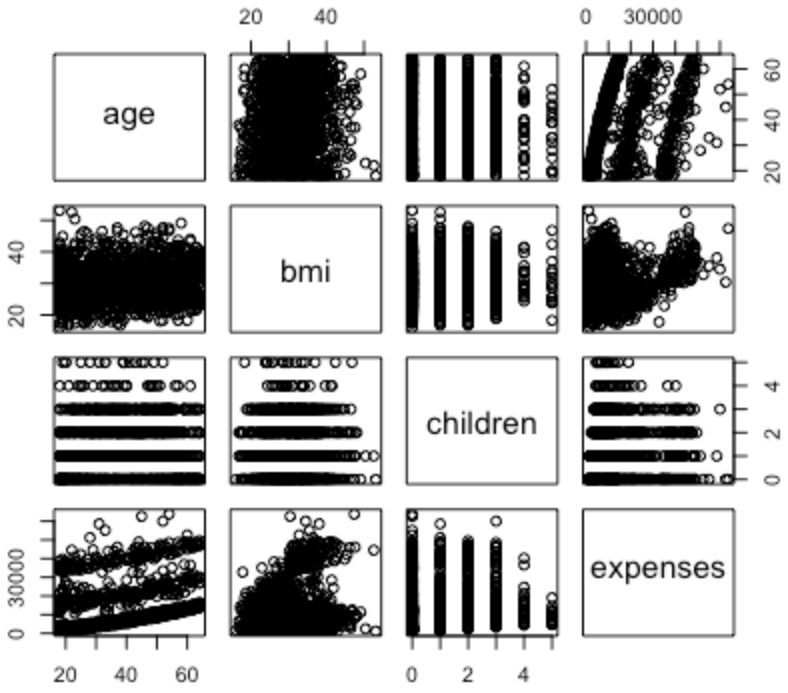
```
> logexp<-log(dataframe$expenses)
> hist(logexp,main="Histogram of Logged Expenses",xlab="Logged Expenses")
```


Histogram of Logged Expenses
Frequency
0
50
150
250
350
7
8
9
10
11
Logged Expenses

#2.

Scatter plot matrix of age, bmi, children, and expenses:

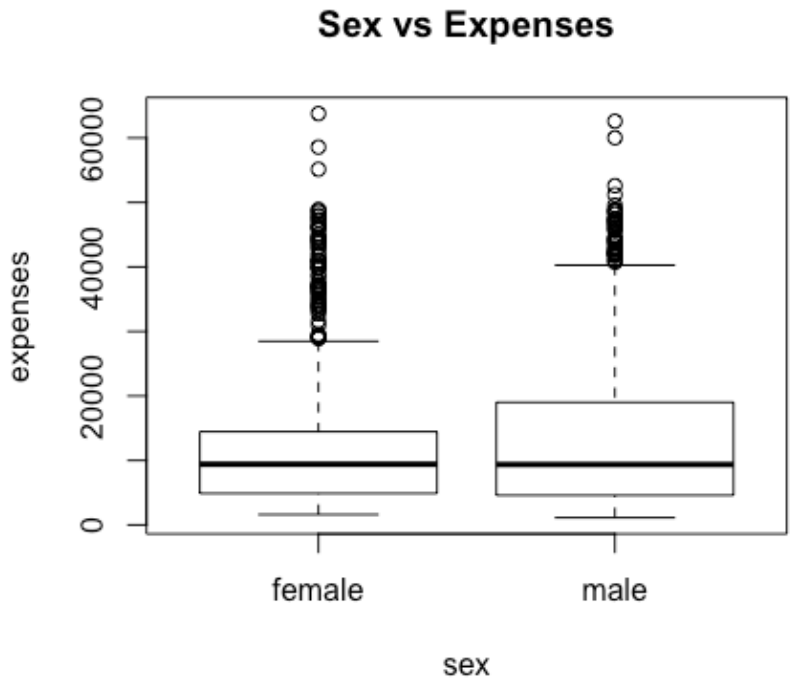
```
> age<-c(dataframe$age)
> bmi<-c(dataframe$bmi)
> children<-c(dataframe$children)
> expenses<-c(dataframe$expenses)
> pairs(data.frame(age,bmi,children,expenses))
```



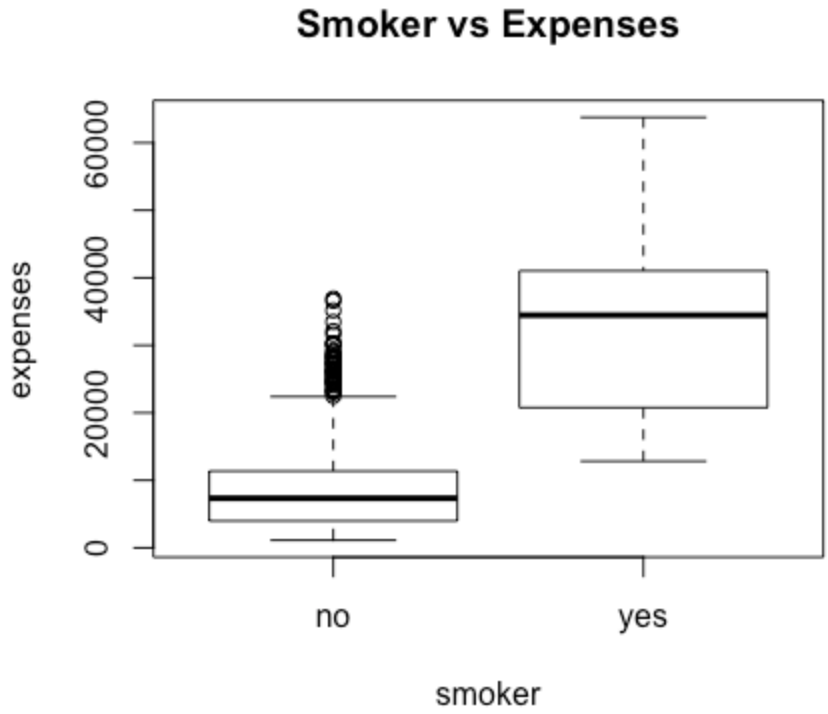
There is a slight curvature in the scatterplot of expenses against age.

This indicates a quadratic relationship between the two variables. A quadratic regression will give a better result rather than a linear regression, which will be elaborated further in question 4.

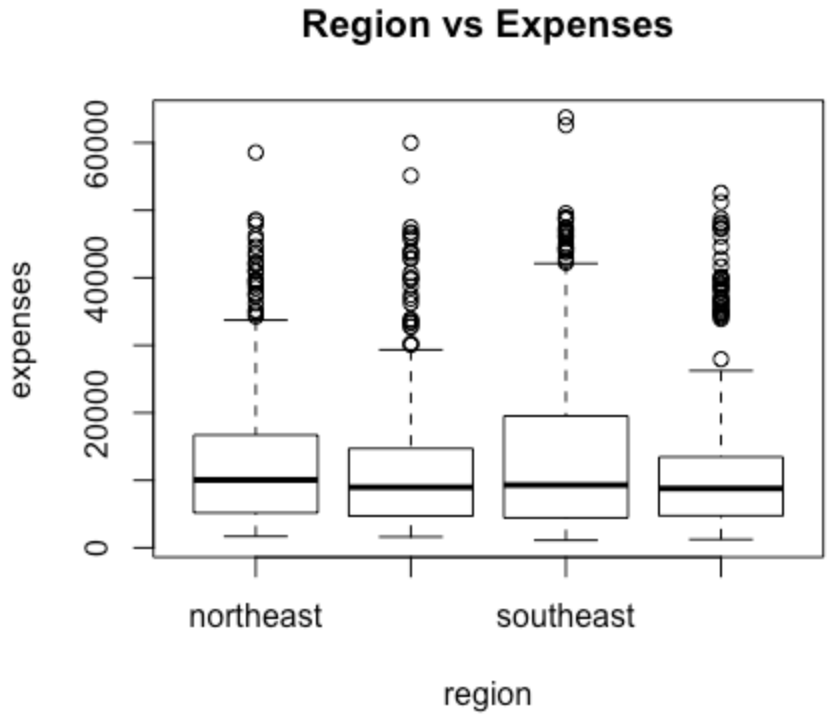
```
> plot(dataframe$sex,expenses,main="Sex vs Expenses",xlab="sex",ylab="expenses")
```



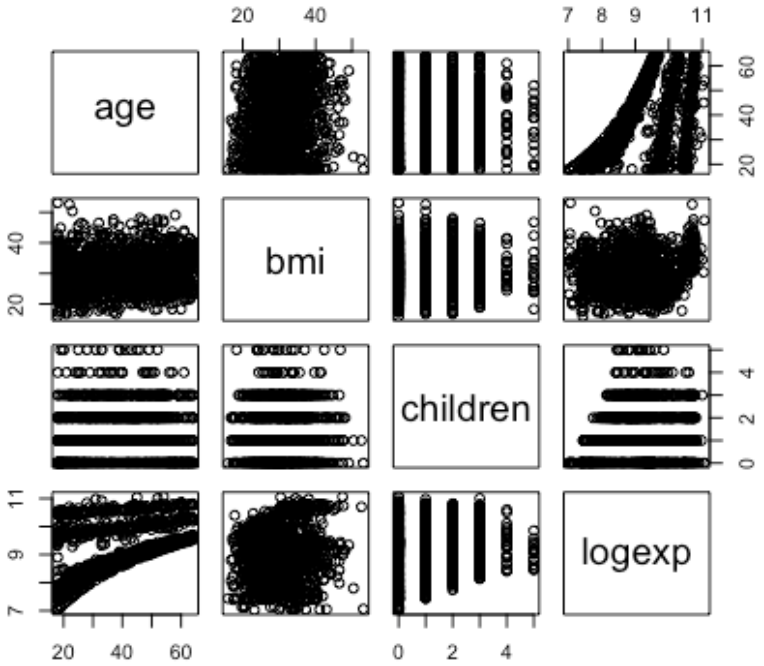

```
> plot(dataframe$smoker,expenses,main="Smoker vs Expenses",
xlab="smoker",ylab="expenses")
```



```
> plot(dataframe$region,expenses,main="Region vs Expenses",xlab="region",ylab="expenses")
```

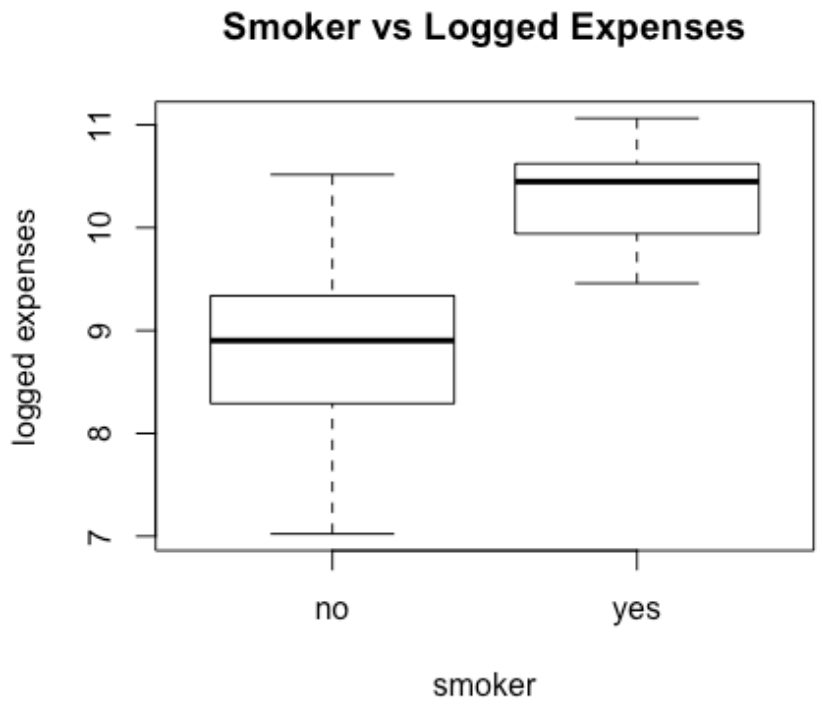



```
> pairs(data.frame(age,bmi,children,logexp))
```

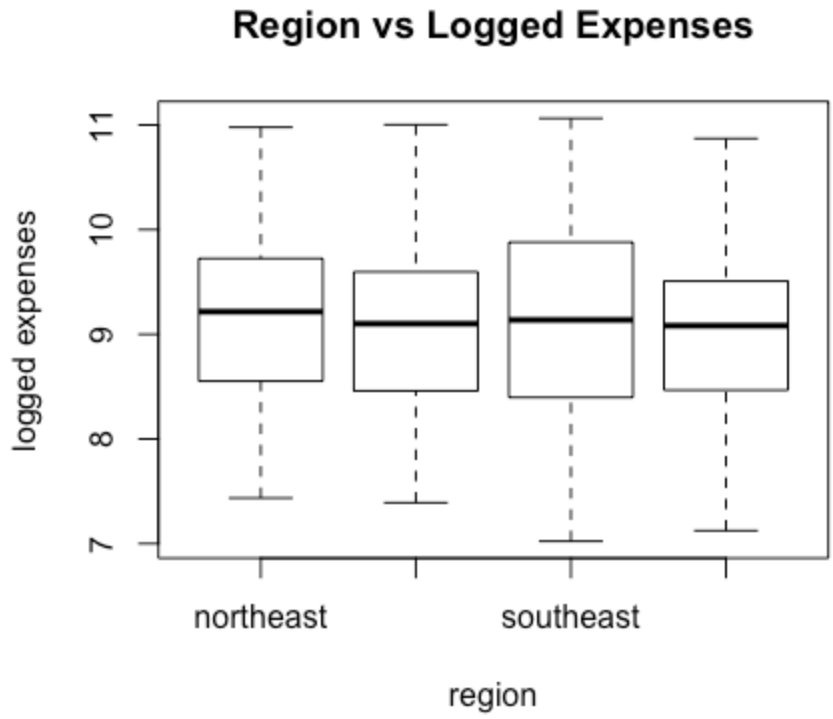


```
> plot(dataframe$sex,logexp,main="Sex vs Logged Expenses",xlab="sex",ylab="logged expenses")
```


```
> plot(dataframe$smoker,logexp,main="Smoker vs Logged Expenses",xlab="smoker",ylab="logged expenses")
```




```
> plot(dataframe$region,logexp,main="Region vs Logged
Expenses",xlab="region",ylab="logged expenses")
```



As can be seen from the graphs with original response variable expenses, higher values tend to be more spread out than lower values. After the logarithmic response variable transformation, the values are much more normally distributed.

#3

Regression of the expenses on the full set of variables

```
> model=lm(expenses~age+sex+bmi+children+smoker+region,data=dataframe)
> summary(model)

Call:
lm(formula = expenses ~ age + sex + bmi + children + smoker + region, data = dataframe)

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 ***
age                256.8       11.9   21.586      < 2e-16 ***
sexmale           -131.3      332.9   -0.395     0.693255
bmi                339.3       28.6   11.864      < 2e-16 ***
children           475.7      137.8    3.452     0.000574 ***
smokeryes         3847.5      413.1   57.723      < 2e-16 ***
regionnorthwest   -352.8      476.3   -0.741     0.458976
regionsoutheast  -1035.6      478.7   -2.163     0.030685 *
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
```

Interpretation of the model

MLR function: $\widehat{Y}$= -11941.6 + 256.8 age -131.3 sexmale + 339.3 bmi +475.7 children + 3847.5 smokeryes - 352.8 regionnorthwest - 1035.6 regionsoutheast - 959.3 regionsouthwest

The model also indicates standard errors measuring the precision of the regression parameter estimators, individual t-statistics for the regression parameter estimators, and the p-value that

goes along with it. At least at the level of 0.05, only the sexmale (p-value = 0.693255) appears to be an insignificant predictor.

The residual standard error of 6062 points out that the actual future value of Y is likely to be far from the predicted value - the model prediction is inaccurate. This results in the lower coefficient of determination R^2 compared to log expenses model (as will be elaborated below) since more variation will remain in Y, SSE will be close to SSTO.

The big F-statistic (500.9) value with a relatively small p-value (<0.05) shows that we have enough evidence to reject the null hypothesis.

Number of predictors

Since there are categorical variables in the models, the sex variable has values male and female, the smoker variable has values yes and no, and the region variable has values northwest, northeast, southwest, and southeast. In the function, sexmale, smokeryes, regionnorthwest, regionsoutheast, and regionsouthwest act as dummy indicator variables which refers to the omitted indicator variable in each category. If the categorical variable exists, the model returns 1; otherwise, it would return 0. For example, if the person is from southeast, the function will return value 1 for southeast and 0 for other region variables; if the person is from northeast, the function will return 0 for all region variables.

Therefore, the model contains 3 indicators for numerical variables + dummy variables with (2-1) for sex category + (2-1) for smoker category + (4-1) for region category = 8 predictors.

Regression of the logged expenses on the variables

```
> logmodel=lm(logexp~age+sex+bmi+children+smoker+region,data=dataframe)
> summary(logmodel)

Call:
lm(formula = logexp ~ age + sex + bmi + children + smoker + region, data = dataframe)

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 ***
age             0.0345816   0.0008721   39.654   < 2e-16 ***
```

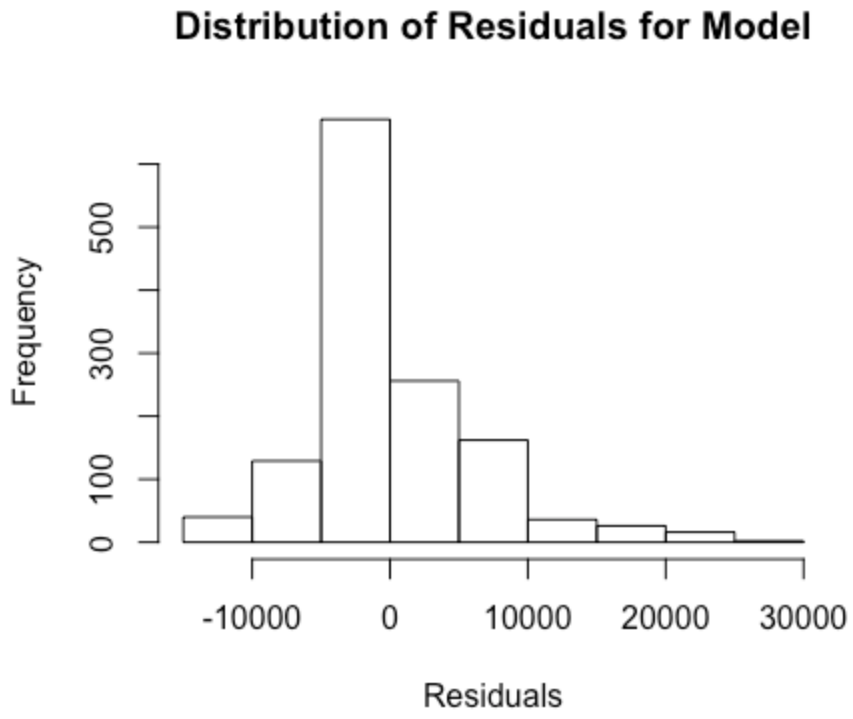
sexmale	-0.0754109	0.0244017	-3.090	0.002040 **
bmi	0.0133658	0.0020960	6.377	2.49e-10 ***
children	0.1018651	0.0100997	10.086	< 2e-16 ***
smokeryes	1.5542783	0.0302800	51.330	< 2e-16 ***
regionnorthwest	-0.0637805	0.0349064	-1.827	0.067896 .
regionsoutheast	-0.1571654	0.0350837	-4.480	8.12e-06 ***
regionsouthwest	-0.1289048	0.0350274	-3.680	0.000242 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

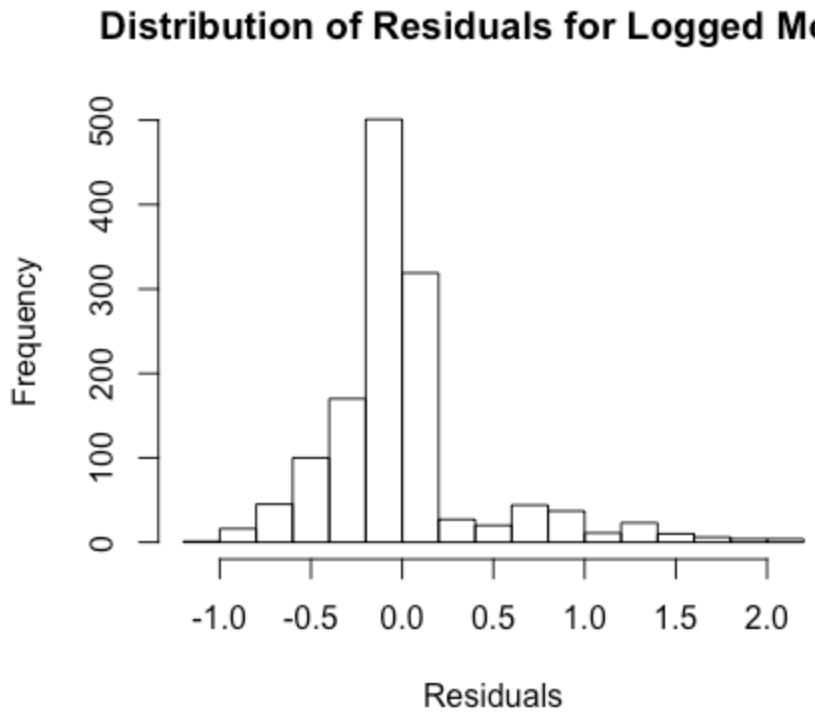
Residual standard error: 0.4443 on 1329 degrees of freedom
Multiple R-squared: 0.7679, Adjusted R-squared: 0.7665
F-statistic: 549.7 on 8 and 1329 DF, p-value: < 2.2e-16

Histograms for distribution of the residuals:

```
> resid1 <- residuals(model)
> resid2 <- residuals(logmodel)
> hist(resid1, main="Distribution of Residuals for Model", xlab="Residuals")
```




```
> hist(resid2, main="Distribution of Residuals for Logged Model", xlab="Residuals")
```



The regression with the log expenses is significantly better.

At a level of 0.05 or below, the new p-value indicates that every predictor appears to be significant in this model.

The residual standard error of 0.4443 (close to 0) comparing to 6062 in the previous model shows that the prediction is notably more precised. As only little variation remain in Y, SSE is relatively small, and R^2 as well as adjusted R^2 will be higher, which is 76.79% compared to 75.09%.

The distribution of the residuals is distributed randomly in the logmodel as it is more symmetric about its mean 0 and not as strongly skewed as the original model.

As a result, p-value is significantly small. The null hypothesis is rejected.

#4
Regression with the square of age

```
> sqage<-age*age
> sqagemodel=lm(expenses~age+sex+bmi+children+smoker+region+sqage,data=dataframe)
```

```
> summary(sqagemodel)

Call:
lm(formula = expenses ~ age + sex + bmi + children + smoker +
    region + sqage, data = dataframe)

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 ***
age                -54.423      80.989    -0.672      0.501716
sexmale           -138.451     331.189    -0.418      0.675983
bmi                335.291      28.467    11.778       < 2e-16 ***
children           642.121     143.613     4.471      8.44e-06 ***
smokeryes        23858.690     410.976    58.054       < 2e-16 ***
regionnorthwest   -367.632     473.771     -0.776     0.437905
regionsoutheast -1031.998     476.164     -2.167     0.030388 *
regionsouthwest   -956.787     475.398     -2.013     0.044358 *
sqage                3.925       1.010      3.885     0.000107 ***
---
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
```

Since there is a curvature in the scatterplot of expenses against age, a quadratic definitely gives a better model.

At the level of 0.05, the p-value of 0.000107 indicates that square of age is a significant predictor. The multiple R-squared increases as a result and the p-value still remains significantly small. The model prediction is also more precised as the residual standard error decreases.

Regression with the interaction term between the bmi and smoker variables

```
> smoker<-ifelse(dataframe$smoker=="yes",1,0)
> bmismoker<-bmi*smoker
```

```
> intmodel=lm(expenses~age+sex+bmi+children+smoker+region+sqage+bmismoker,data=dataframe)
> summary(intmodel)

Call:
lm(formula = expenses ~ age + sex + bmi + children + smoker +
    region + sqage + bmismoker, data = dataframe)

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 *
age             -3.299e+01   6.459e+01   -0.511    0.60957
sexmale         -5.054e+02   2.644e+02   -1.911    0.05617 .
bmi              2.005e+01   2.541e+01    0.789    0.43037
children         6.750e+02   1.145e+02    5.894    4.77e-09 ***
smokeryes       -2.035e+04   1.635e+03  -12.445     < 2e-16 ***
regionnorthwest -5.987e+02   3.779e+02   -1.584    0.11336
regionsoutheast -1.206e+03   3.798e+02   -3.176    0.00153 **
regionsouthwest -1.226e+03   3.792e+02   -3.234    0.00125 **
sqage            3.740e+00    8.057e-01    4.642    3.79e-06 ***
bmismoker        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
```

Both the quadratic term and interaction term result in better prediction.

With p-values much lower than 0.05 level, both the square of age and the interaction term bmi*smoker are significant predictors. By adding the bmismoker variable, the model’s R-squared as well as adjusted R-squared increase notably (about 10%) comparing to the original model. The p-value also remains small. The residual standard error decreases significantly, indicating more accurate prediction.