Tuan Dang
04/13/2021

Project 2: Predicting Medical Expenses using Multiple Regression

Part 1.

R Code:

```
mydata <- read.csv("insurance.csv")
View(mydata)
```

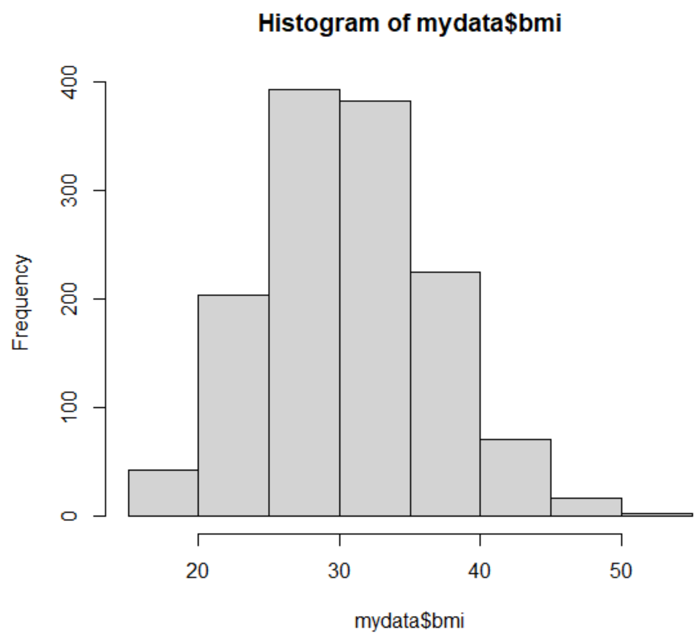
Categorical Variables:

```
table(mydata$sex)
female   male
   662    676

table(mydata$smoker)
  no  yes
1064  274

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

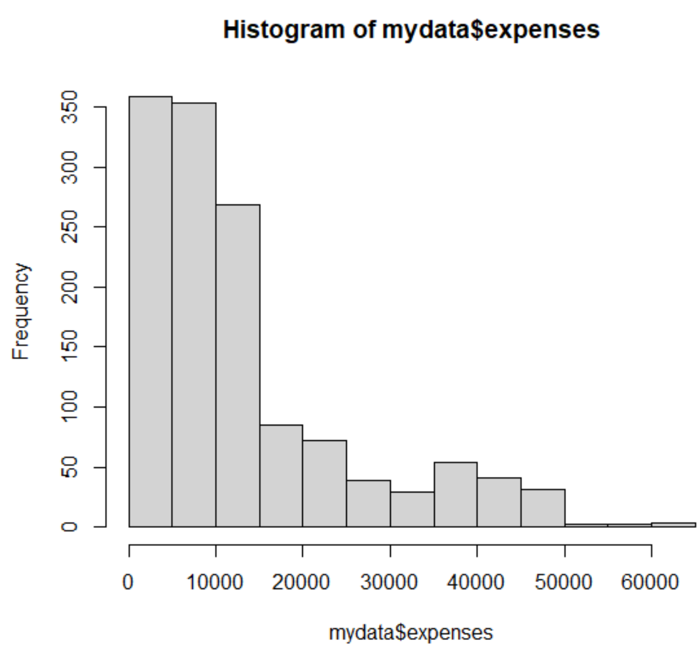
Numerical Variables:

```
summary(mydata$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   27.00   39.00   39.21   51.00   64.00

summary(mydata$bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   26.30   30.40   30.67   34.70   53.10

summary(mydata$children)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   1.095   2.000   5.000
```

Histogram of the BMI predictor variable:
hist(mydata$bmi)

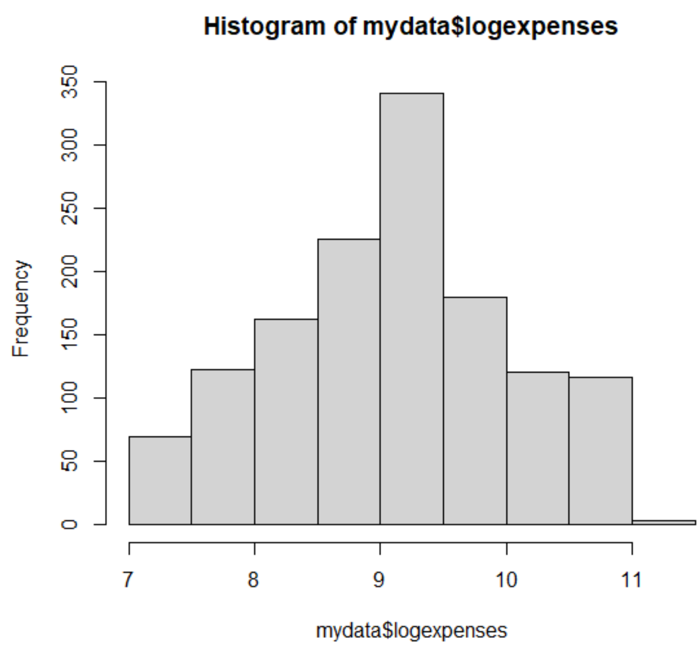


Histogram of the expenses response variable:
hist(mydata$expenses)



The BMI variable seems normal. However, the expenses variable does not, since the histogram is extremely right-skewed.

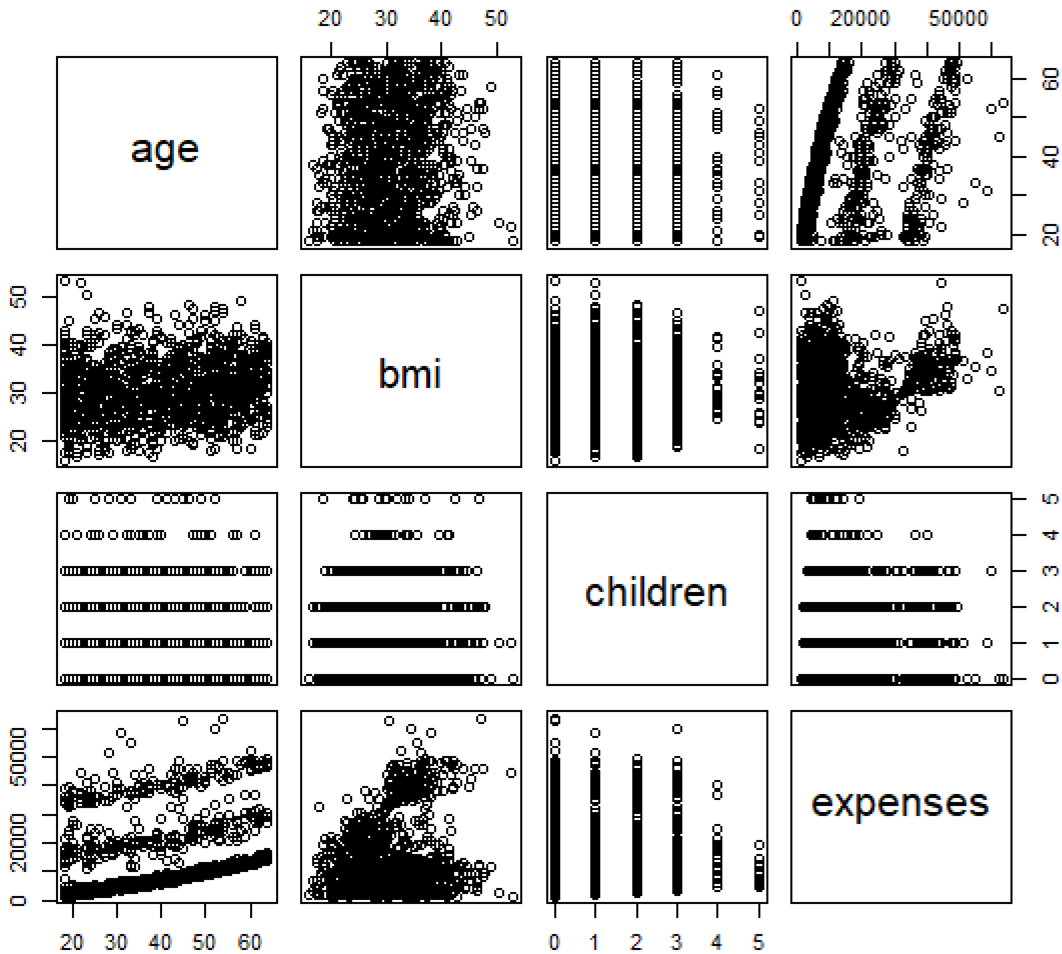
Histogram of the logged expenses response variable:
mydata$logexpenses <- log(mydata$expenses)
hist(mydata$logexpenses)



Creating a new response variable by taking the log of expenses does seem to fix the issue from before. This histogram looks much more normal, with a slight left skewness.

Part 2.

R Code:
pairs(mydata[c("age","bmi","children","expenses")])

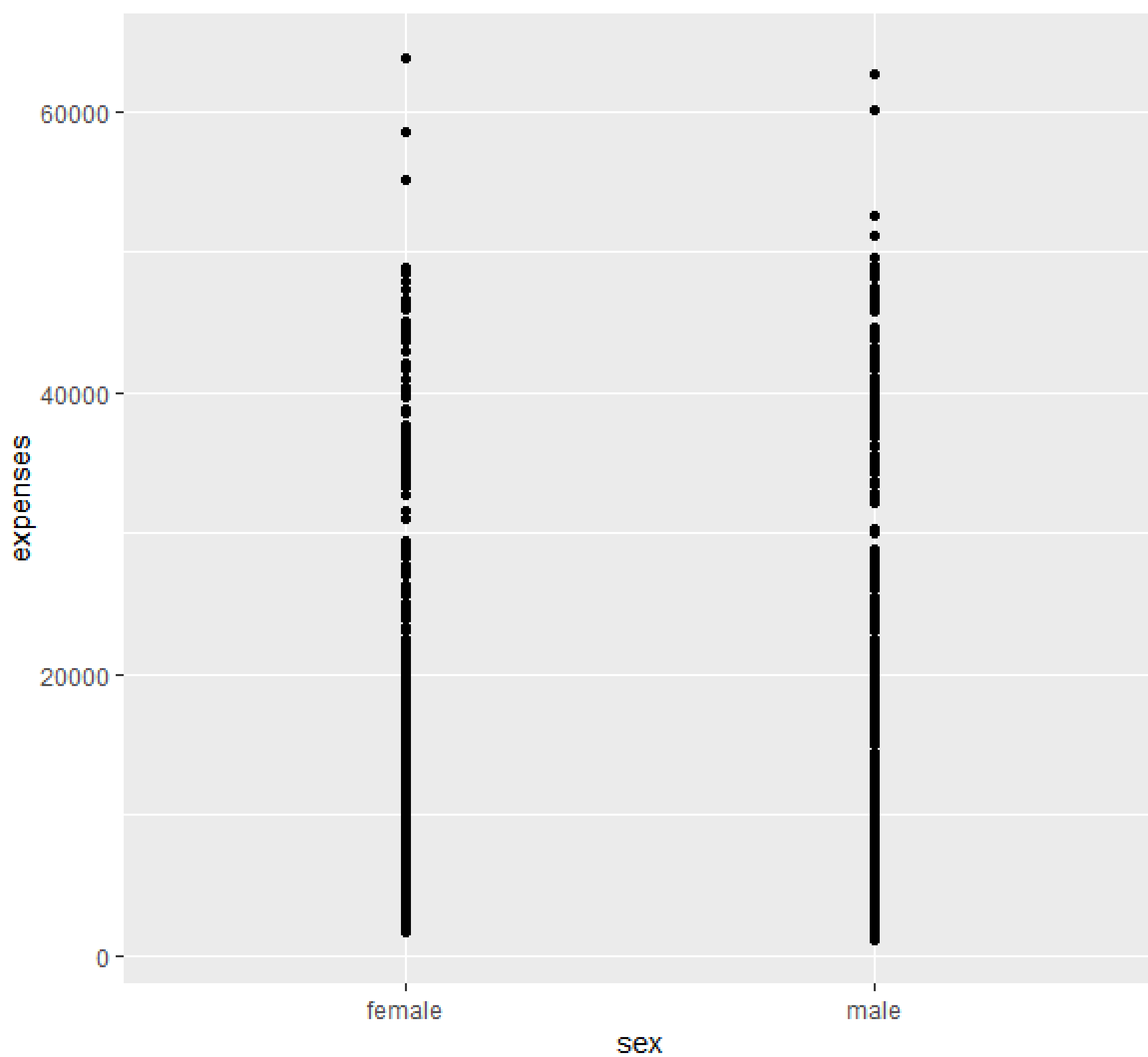


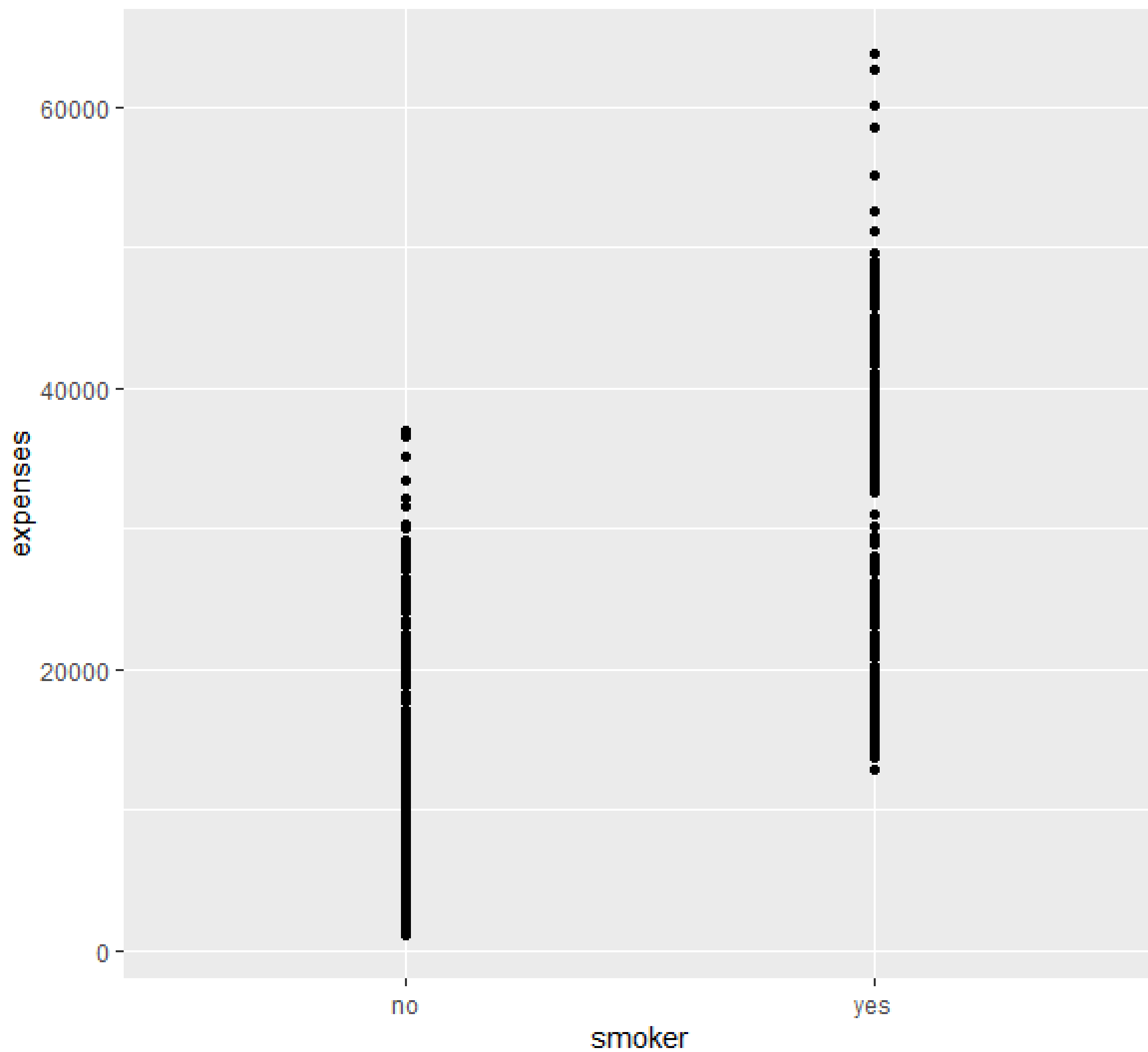
These numerical predictor variables don't seem to have any noticeable relationship between each of them. However, they all have some sort of a correlation with the response variable. Most notably, age has a strong positive correlation with expenses, with the data points being separated into three distinct lines. BMI also sort of has a positive correlation, but the pattern is much less clear. On the other hand, the number of children the participants have surprisingly seems to have a negative correlation with medical expenses.

R Code:

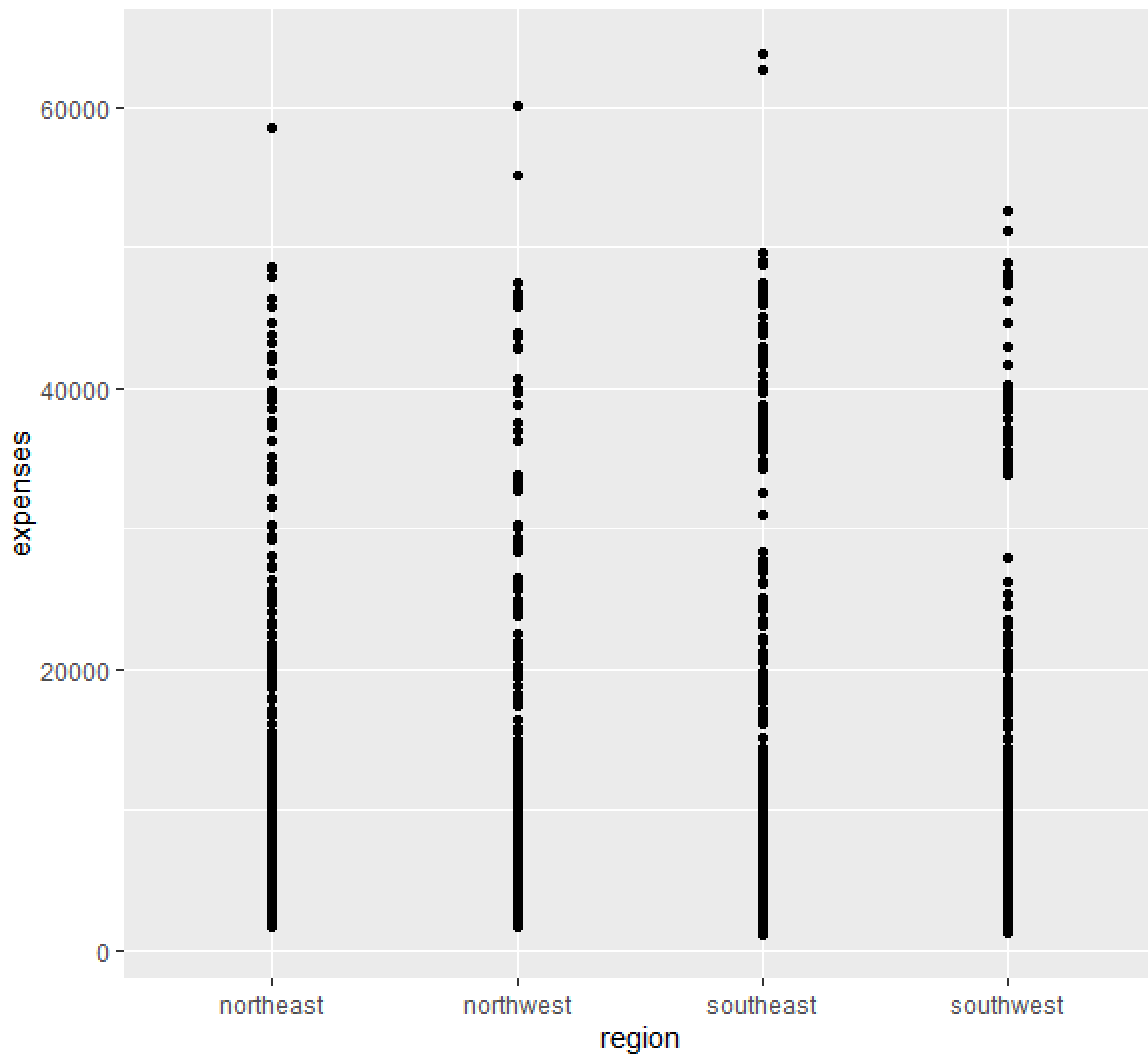
```
library(ggplot2)
library(dplyr)

ggplot(mydata, aes(x=sex,y=expenses))+geom_point()
ggplot(mydata, aes(x=smoker,y=expenses))+geom_point()
ggplot(mydata, aes(x=region,y=expenses))+geom_point()
```

Scatter Plots of Expenses based on Categorical Variables:





60000
40000
20000
0
expenses
no
yes
smoker

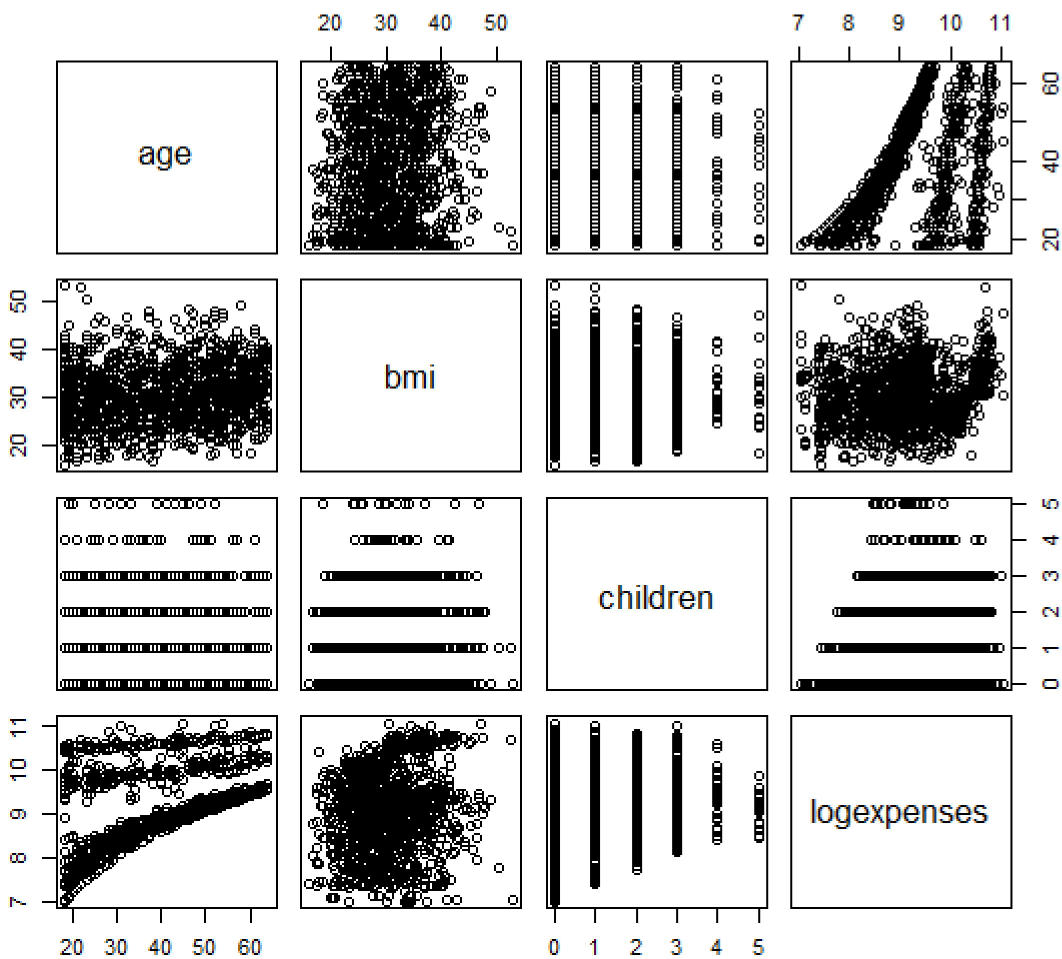


The difference in sex or region does not seem to have an impact on expenses, except for a few outliers. Although, the fact that a participant is a smoker does significantly increase their expenses based on what we can see on the graph.

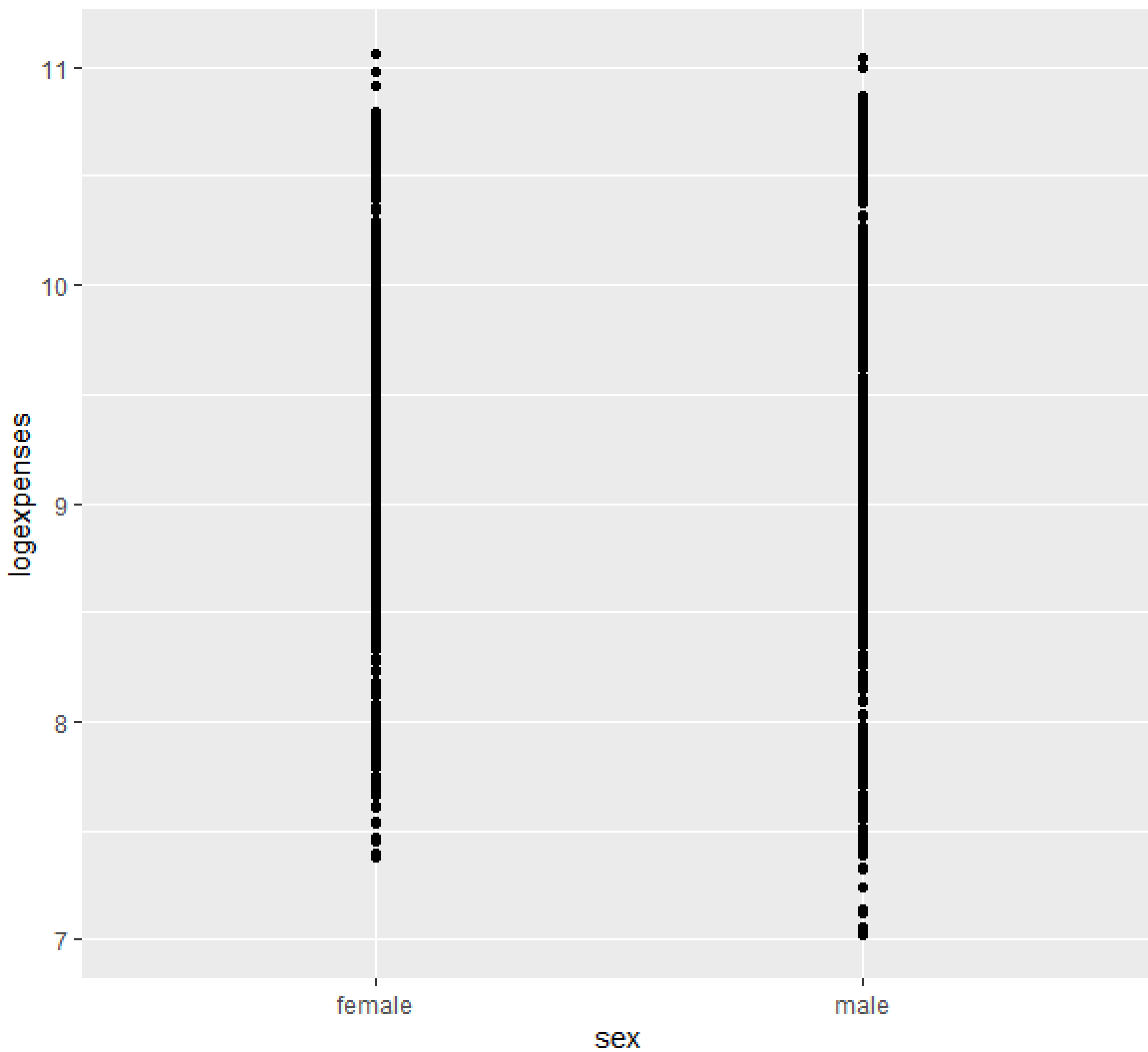
Scatter Plots of Logged Expenses based on Categorical Variables:

R Code:

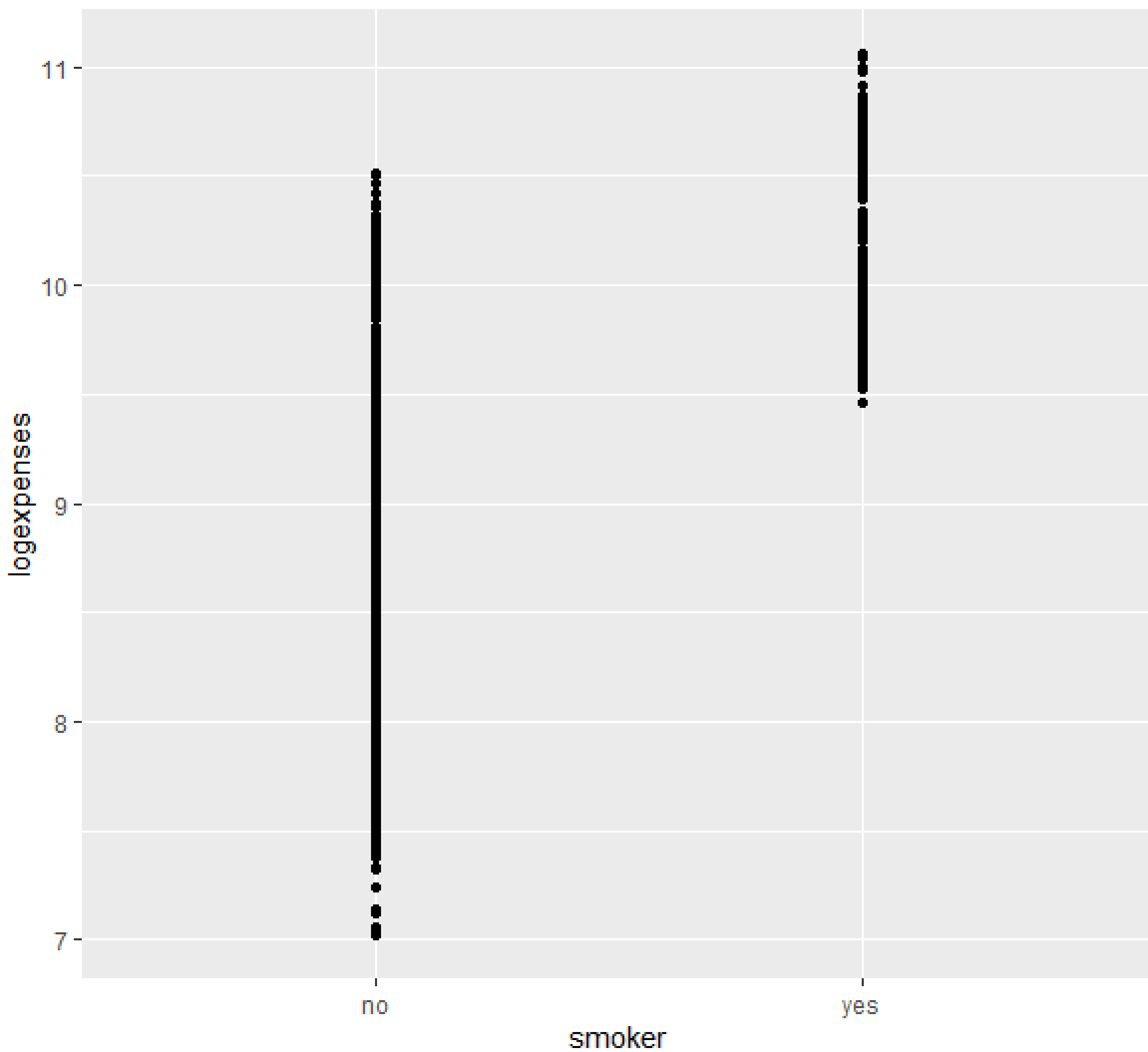
```
pairs(mydata[c("age","bmi","children","logexpenses")])
ggplot(mydata, aes(x=sex,y=logexpenses))+geom_point()
ggplot(mydata, aes(x=smoker,y=logexpenses))+geom_point()
ggplot(mydata, aes(x=region,y=logexpenses))+geom_point()
```



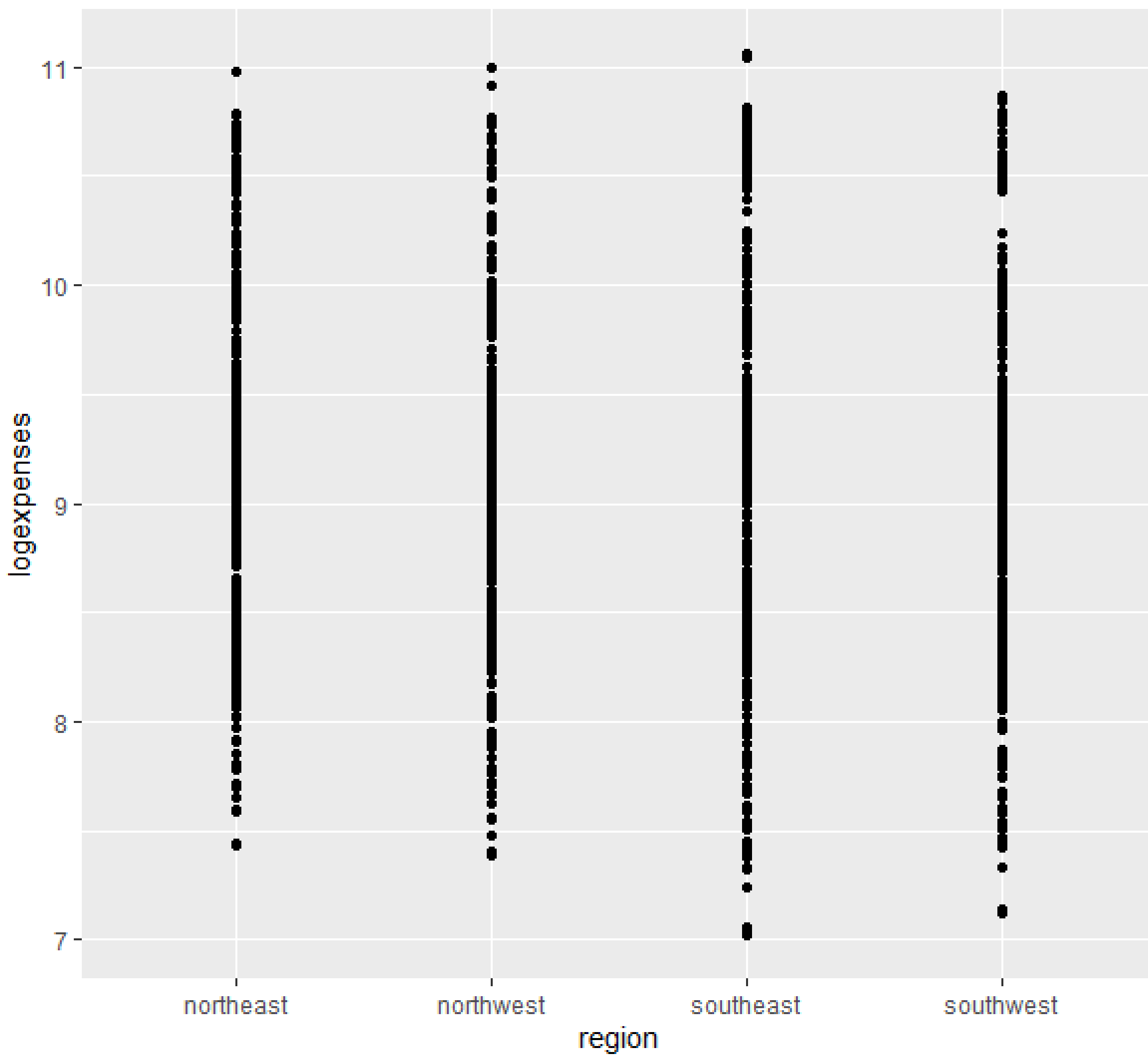
With the expense variable being modified, we can still see a strong positive correlation between age and the response variable, although the two lines on top of the graph seem flat. BMI does not seem to have any correlation with logged expenses, which is a change from before. The number of children also does not seem to have a significant effect, although the variance of logged expenses decreases the more children they have.



11
10
9
8
7
logexpenses
female
male
sex



11
10
9
8
7
logexpenses
no
yes
smoker



After modifying the response variable, sex and region still don't have a huge difference between the inputs. It still looks very clear that being a smoker increases medical expenses.

Part 3.

Model for predicting expenses:

R Code:
model <- lm(expenses ~ age + sex + bmi + children + smoker + region, mydata)

Call:

```
lm(formula = expenses ~ age + sex + bmi + children + smoker +
    region, data = mydata)

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        23847.5      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
```

The model summary shows that age, BMI, children, and smoker are all significant predictors. Sex has a very high p-value, which means it could be removed. Lastly, there is some conflict with the region variable, since only the northwest seems insignificant, which is surprising considering each region did not seem that different in the graphs. The multiple R-squared and adjusted R-squared look very good at around 75% for each of them.

There are more than 6 predictors in the model summary since this is how the model accounts for categorical variables. Sex and smoker only have two inputs each, so only one of the inputs need to be in the formula. With region, we have all of them except for the northeast, since the absence of the other three regions would mean that the data point will be northeast. When it is any other region, a 1 will indicate their presence, which will include their coefficient in the formula.

Model for predicting the log of the expenses:

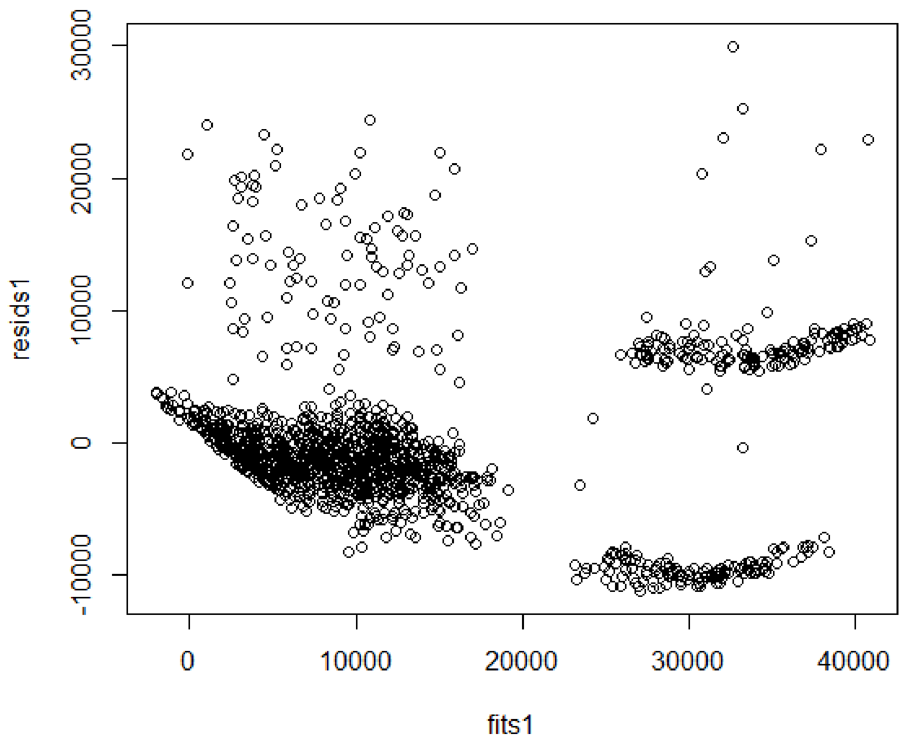
R Code:

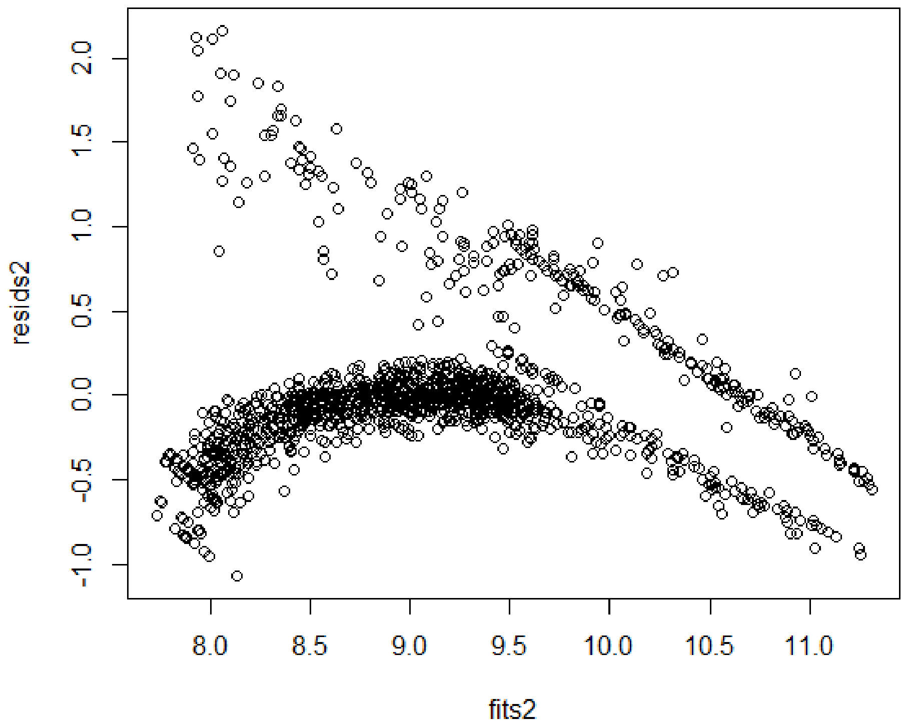
```
model2 <- lm(logexpenses ~ age + sex + bmi + children + smoker + region,
mydata)
```

```
Call:
lm(formula = logexpenses ~ age + sex + bmi + children + smoker +
    region, data = mydata)

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
```

There are two main differences between this model and the previous one, although they are mostly similar. The first is that sex becomes a much more significant predictor. The second difference is that the multiple R-squared and adjusted R-squared are slightly higher by only 1-2%.

These are the two Residual-Fits Plots for the model before and after we modify expenses. The modified model gives residuals with a much lower variance, where the previous model can give residuals with a difference in the tens of thousands.





Part 4.

The regression with the square of the age included:

R Code:
```
mydata$agesq <- (mydata$age)^2
model3 <- lm(expenses ~ age + agesq + sex + bmi + children + smoker + region,
mydata)
summary(model3)
```

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

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
agesq               3.925      1.010   3.885 0.000107 ***
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 *
---
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
```

With the square of the age added, age becomes a much less significant predictor, while its square is very significant. Other than this, the new variable does not change the other variables much, and the R-squared value stays roughly the same, so the prediction shouldn't be any better.

The regression with the square of the age and the interaction term between bmi and smoker included:

R Code:

```
model4 <- lm(expenses ~ age + agesq + sex + bmi + children + smoker + region +
bmi*smoker, mydata)
summary(model4)
```

```
Call:
lm(formula = expenses ~ age + agesq + sex + bmi + children +
```

```
    smoker + region + bmi * smoker, data = mydata)

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
agesq            3.740e+00  8.057e-01   4.642 3.79e-06 ***
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 **
bmi: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
```

With the addition of the interaction term between BMI and smoker, BMI becomes insignificant while smoker stays the same. The interaction term itself is very significant. Overall, this new variable actually increased the R-squared value from 75% to 84%, which means this new model makes better predictions than the previous one. Furthermore, since R-squared is now above 0.8, we can consider this a good model for making predictions.