Understand the factors on which the pricing of health insurance depends as a healthcare consulting company

Data Description

The data consists of a data frame with 1338 observations on the following 7 variables:

1. price: Response variable ($)
2. age: Quantitative variable
3. sex: Qualitative variable
4. bmi: Quantitative variable
5. children: Quantitative variable
6. smoker: Qualitative variable
7. region: Qualitative variable

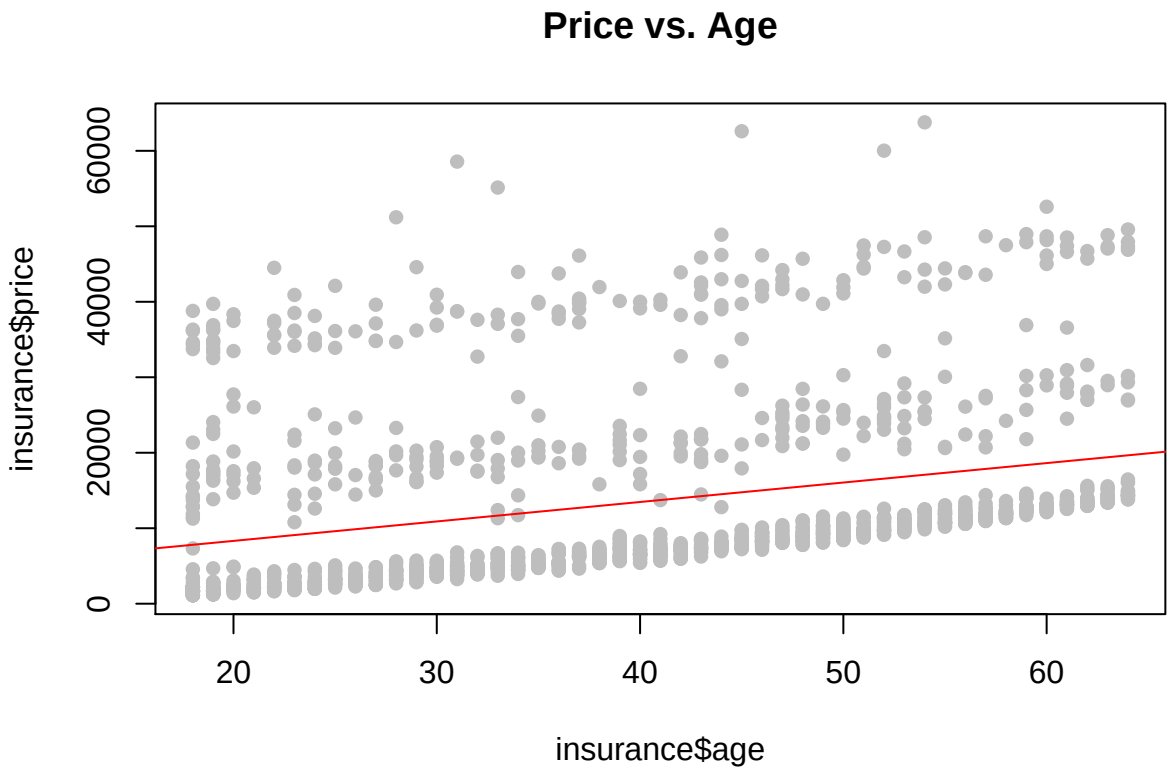
Read data

```
insurance <- read.csv("insurance.csv", head = TRUE)
```

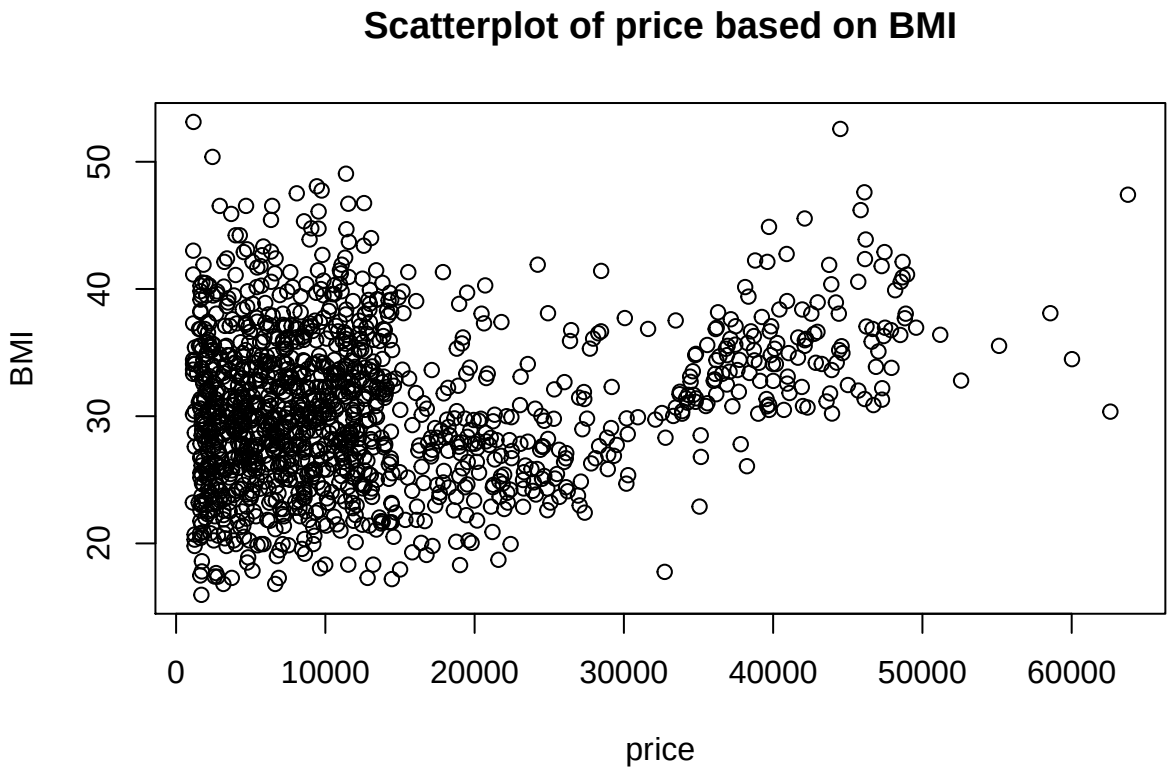
#Exploratory Data Analysis

Scatterplots of the response, *price*, against three quantitative predictors *age*, *bmi*, and *children*

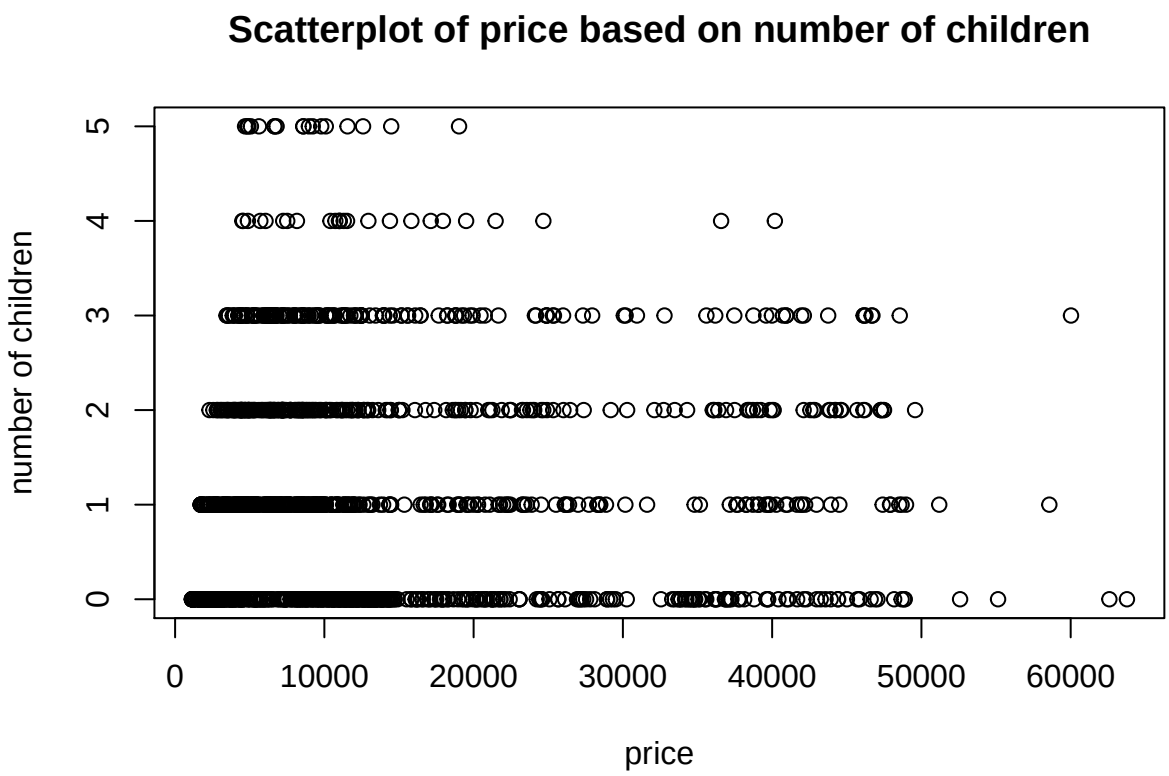
```
#age
plot(insurance$price~insurance$age, data=insurance, main="Price vs. Age", col="grey", pch = 16)
abline(lm(insurance$price~insurance$age, data=insurance), col="red")
```



```r
#bmi
plot(insurance$price, insurance$bmi, main="Scatterplot of price based on BMI", xlab="price",ylab="BMI")
```




```r
#children
plot(insurance$price, insurance$children, main="Scatterplot of price based on number of children", xlab=
```



The relationship between price and age seems moderately linear with some general increasing/positive trend. The relationship between price and BMI seems moderately linear with no significant apparent trend. We can see that as the bmi increases the variance of the insurance price appears to increase as well. The relationship between price and number of children does not seem linear with no apparent trend.

Correlation coefficients

```
cor(insurance$price, insurance$age)
```

```
## [1] 0.2990082
```

```
cor(insurance$price, insurance$bmi)
```

```
## [1] 0.198341
```

```
cor(insurance$price, insurance$children)
```

```
## [1] 0.06799823
```

The very low values of correlation coefficients further prove that there exist very low or no correlation between price and age, price and BMI, and price and number of children. This means that price of health insurance is just as likely to be affected by young people as by older people. Similarly, the price of health insurance is just as likely to be affected by people

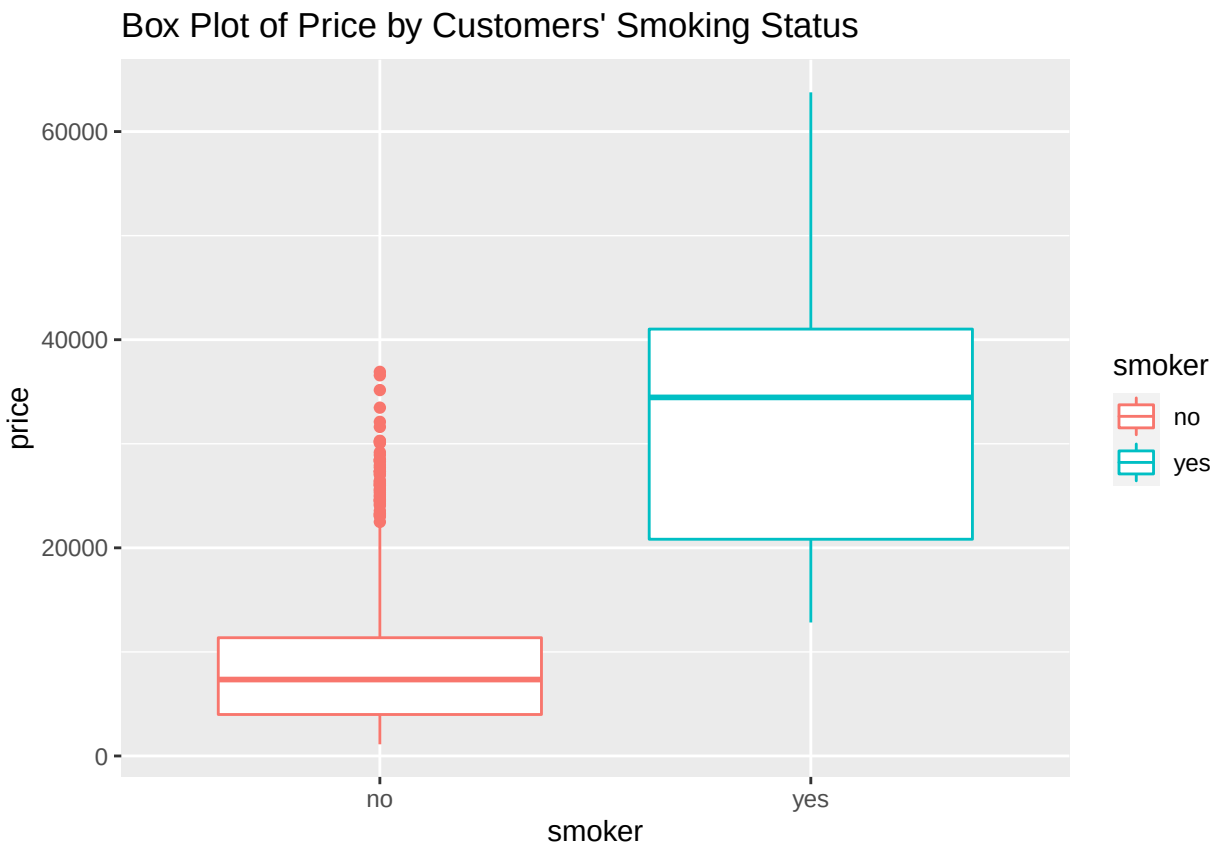
with low BMI as by people with high BMI. Lastly, the price of health insurance is likely to be affected by people with any number of children or no children. Outside of that, our analysis aligns with our hypothesis that the response is positively correlated with each of the predictor variables.

Box plots of the response, *price*, and the three qualitative predictors *sex*, *smoker*, and *region*

```
#make categorical variables into factors
insurance$sex<-as.factor(insurance$sex) #makes female the baseline level
insurance$smoker<-as.factor(insurance$smoker) #makes no the baseline level
insurance$region<-as.factor(insurance$region) #makes northeast the baseline level
```

```
#boxplot
library(ggplot2)
#sex
ggplot(insurance, aes(x=sex, y=price, color=sex)) +
geom_boxplot() +
ggtitle("Box Plot of Price by Customers' Sex")
```


```
#smoker
ggplot(insurance, aes(x=smoker, y=price, color=smoker)) +
geom_boxplot() +
ggtitle("Box Plot of Price by Customers' Smoking Status")
```




```
#region
ggplot(insurance, aes(x=region, y=price, color=region)) +
geom_boxplot() +
ggtitle("Box Plot of Price by Customers' Region")
```

The price of health insurance is very similar for male vs female. The price of health insurance for smokers is higher than for non-smokers. The price of health insurance is comparative for people regardless of the region they live in. The box plots suggest price would significantly differ between smokers and non-smokers, but not between males and females or between people who live in different regions

Should still use multiple linear regression?

There are definite relationships for several of the predictors, and although some predictors such as children,sex, and region don't seem to be marginally associated with the response, they still could be useful in predicting the response variable when considering other predictors in the model.

Fitting the Multiple Linear Regression Model

Full model

```
model1 <- lm(price~.,data=insurance)
model1
```

```
##
## Call:
## lm(formula = price ~ ., data = insurance)
##
## Coefficients:
```

```
##     (Intercept)             age          sexmale             bmi
##        -11938.5           256.9           -131.3           339.2
##        children        smokeryes  regionnorthwest  regionsoutheast
##           475.5         23848.5           -353.0          -1035.0
## regionsouthwest
##          -960.1
```

```
#extract R-squared:
cat("R^2:", summary(model1)$r.squared)
```

```
## R^2: 0.750913
```

Interpreting coefficient of determination (R-squared) for the model

The Multiple R-squared of 0.7509 suggests that the model explains 75% of variability of the response variable (health insurance price) around its mean which is moderately good. Even after adjusting for the number of predictions, the adjusted R-squared of 0.7494 still explains 75% of variability.

Overall adequacy of the model, using $\alpha = 0.05$

```
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ ., data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## ---
## 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: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

We can conduct an F-Test in order to assess the overall adequacy of the model. The elements of the test of overall regression are as follows: null hypothesis H0 : B1 = ... = B8 = 0. Null hypothesis is that none of the variables have predicting power on price of health insurance price.

Alternative hypothesis Ha: At least one of the slope coefficients is nonzero. Alternative hypothesis is that at least one of the predicting variables has a predicting power on price of health insurance.

Test statistic: F(8, 1329) = 500.8 p-value: < 2.2e-16 The F-value is 500.8 and the p-value is approximately 0, so less than the alpha-level of 0.05, which means that we reject the null hypothesis and conclude that at least one of the predicting variables has a predictive power.The overall model appears to be statistically useful in predicting price

Model Comparison

Assuming a marginal relationship between *region* and *price*, an ANOVA F-test on the mean insurance prices among the different regions.

```
aov.model <- aov(price~region,data=insurance)
summary(aov.model)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## region         3 1.301e+09 433586560    2.97 0.0309 *
## Residuals   1334 1.948e+11 146007093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model.tables(aov.model,type="means")
```

```
## Tables of means
## Grand mean
##
## 13270.42
##
##  region
##     northeast northwest southeast southwest
##         13406     12418     14735     12347
## rep       324       325       364       325
```

We conduct an Anova F-Test on the mean insurance prices among the different regions. The elements of the test are as follows: H0 : 1 = ... = 4. Ha: At least 2 of the population means are not equal

Test statistic: F(3, 1334) = 2.97. p-value: 0.0309

Based on resulting p-value of F-test of 0.0309 which is less than alpha-level of 0.05, we can reject the null hypothesis that the means of the regions are equal, so at least one of the regions or all affect price of health insurance.Therefore, the mean health insurance price is not the same for all three regions, so we conclude that region does affect price.

Second multiple linear regression model, called *model2*, using *price* as the response variable, and all variables except *region* as the predictors.

```
model2 <- lm(price~age+sex+bmi+children+smoker,data=insurance)

summary(model2)
```

```
##
## Call:
## lm(formula = price ~ age + sex + bmi + children + smoker, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11837.2  -2916.7   -994.2   1375.3  29565.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12052.46     951.26 -12.670  < 2e-16 ***
## age            257.73      11.90  21.651  < 2e-16 ***
## sexmale       -128.64     333.36  -0.386 0.699641
## bmi            322.36      27.42  11.757  < 2e-16 ***
## children       474.41     137.86   3.441 0.000597 ***
## smokeryes    23823.39     412.52  57.750  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6070 on 1332 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7488
## F-statistic:   798 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Partial F-test

```
anova(model2, model1)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ age + sex + bmi + children + smoker
## Model 2: price ~ age + sex + bmi + children + smoker + region
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1   1332 4.9073e+10
## 2   1329 4.8840e+10  3 233431209 2.1173 0.09622 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We conduct a Partial F-Test comparing model2 with model1. The elements of the test are as follows: H0 : Bregionnorthwest = Bregionsoutheast =Bregionsouthwest = 0 Ha: At least one of Bregionnorthwest, Bregionsoutheast, Bregionsouthwest is not equal to zero. Test statistic: Fpartial(3, 1329) = 2.1173 p-value: 0.09622 The p-value for the partial F-test is 0.09622 which is more than the B-level of 0.05, so we cannot reject the null hypothesis that the regression coefficients for regions are zero given all other predictors in model1, at B-level of 0.05

Comparing the two models above, from the first model we can conclude that all regions do not have the same mean insurance prices, and thus that region is useful for predicting insurance prices when not considering other factors. From the second model we can conclude that region did not add explanatory power given that all other predictors were already included in the model.

Coefficient Interpretation using model

```
model1$coefficients[3]
```

```
##   sexmale
## -131.3144
```

The reference group for the categorical variable sex is female. The coefficient of sexmale is -131.3144, which means that in average the price of insurance policies for males are $131.31 cheaper than policies for females, provided all other predictors are held constant.

```
model1$coefficients[4]
```

```
##      bmi
## 339.1935
```

Since the coefficient for bmi is 339.1935, the change in price is 0.01 $\times$ 339.1935 = 3.391935. A 0.01 unit increase of the predictor bmi corresponds with an increase in price by $3.391935, provided all other predictors are held constant.

Confidence and Prediction Intervals

Computing 90% and 95% confidence intervals (CIs) for the parameter associated with *age* for *model1*.

```
confint(model1,"age",level=0.90) #for 90%
```

```
##          5 %     95 %
## age 237.2708 276.4419
```

```
confint(model1,"age",level=0.95) #for 95%
```

```
##        2.5 %   97.5 %
## age 233.5138 280.1989
```

The 90% confidence interval is narrower than the 95% confidence interval. This behavior is expected since the width of the interval depends on the degree of confidence required. As the degree of confidence increases, the width of the interval increases because, in order to be more confident that the true population value falls within the interval, we will need to allow more potential values within the interval..They both do not include 0, which implies that the coefficient for age is statistically significant at both significance levels.

Estimating the average price for all insurance policies with the same characteristics as the first data point in the sample

```
newdata <- insurance[1,]
newdata
```

```
##   age    sex  bmi children smoker    region    price
## 1  19 female 27.9        0    yes southwest 16884.92
```

```
predict(model1, newdata, interval="confidence")
```

```
##        fit      lwr      upr
## 1 25293.71 24143.98 26443.44
```

The average estimated price (mean response for price) is $25293.71. For insurance policy with the same characteristic as the first data point in the sample, the average estimated health insurance prices are $25293.71 with a lower bound of $24143.98 and an upper bound of $26443.44. We are 95% confident that the mean price for all insurance policies with these specific characteristics is between $24,143.98 and $26,443.44. The 95% confidence is a confidence that approximately 95% of the CIs will contain the true population mean if we were to apply the same procedure repeatedly to different samples. Looking at the first data point, the actual cost was 16884.924. This is nowhere close to the fit value and also not in the confidence interval.

Changing the age of the patient to be 50

```
insurance.new=newdata
insurance.new[1]=50
predict(model1,insurance.new,interval="prediction")
```

```
##        fit      lwr      upr
## 1 33256.26 21313.29 45199.23
```

After changing the age to 50 years old now, We can be 95% confident that the mean estimated price of health insurance will increase to $33256.26 with lower bound of $21313.29 and an upper bound of $45199.23. This makes sense as the older customers might need more coverage as young people are relatively healthy.