Peer Grader Guidance

Please review the student expectations for peer review grading and peer review comments. Overall, we ask that you score with accuracy. When grading your peers, you will not only learn how to improve your future homework submissions but you will also gain deeper understanding of the concepts in the assignments. When assigning scores, consider the responses to the questions given your understanding of the problem and using the solutions as a guide. Moreover, please give partial credit for a concerted effort, but also be thorough. **Add comments to your review, particularly when deducting points, to explain why the student missed the points.** Ensure your comments are specific to questions and the student responses in the assignment.

Background

You have been contracted as a healthcare consulting company to understand the factors on which the pricing of health insurance depends.

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

Instructions on reading the data

To read the data in `R`, save the file in your working directory (make sure you have changed the directory if different from the R working directory) and read the data using the `R` function `read.csv()`

```
insurance = read.csv(
  "C:\\Users\\adri_\\Documents\\Gatech\\ISYE6414\\Homeworks\\2_Multiple_linear_regression\\insurance.cs
```

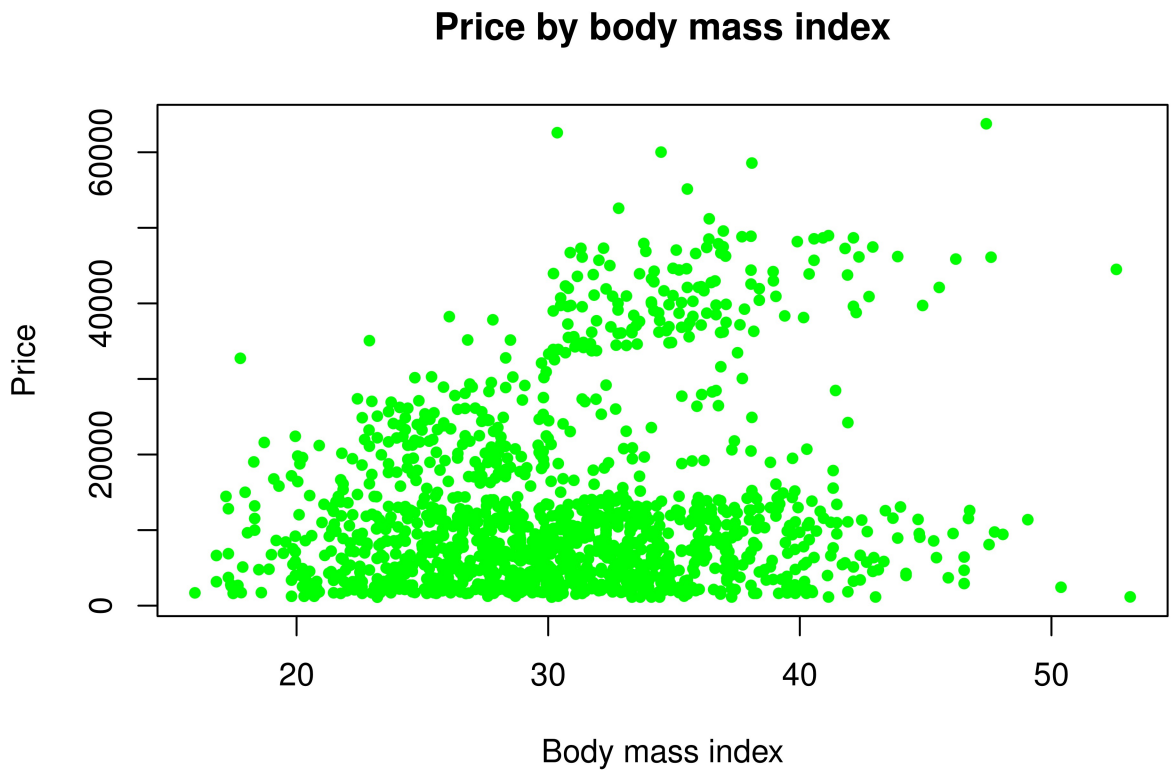
Question 1: Exploratory Data Analysis [12 points]

a. **3 pts** Create plots of the response, *price*, against three quantitative predictors *age*, *bmi*, and *children*. Describe the general trend (direction and form) of each plot.

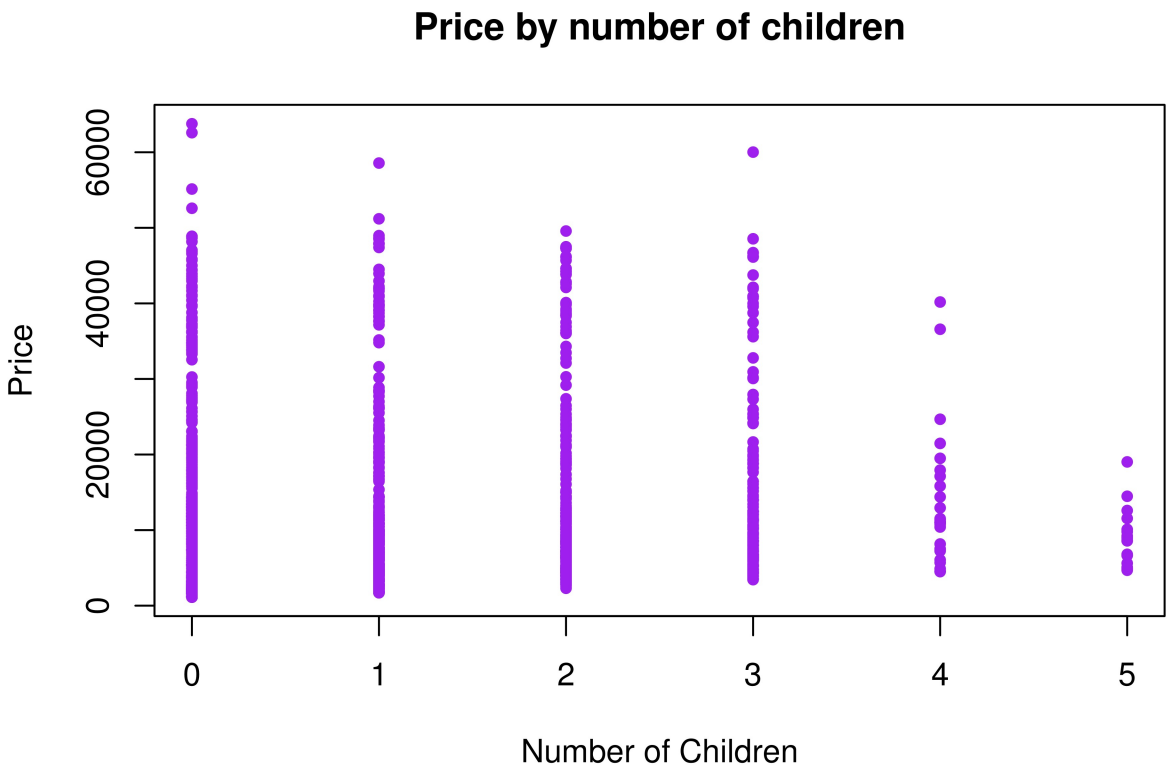
```
price <- insurance[,7]   # price is the response variable
age <- insurance[,1]
bmi <- insurance[,3]
children <- insurance[,4]

plot(age,price,
     xlab="Age",
```

```
     ylab="Price",
     main="Price by Age",
     col = "red",
     pch=20)
```


```
plot(bmi,price,
     xlab="Body mass index",
     ylab="Price",
     main="Price by body mass index",
     col = "green",
     pch=20)
```




```
plot(children,price,
     xlab="Number of Children",
     ylab="Price",
     main="Price by number of children",
     col = "purple",
     pch=20)
```



Price by Age
We can see three (3) positive trends that are approximately parallel to each other, all of them indicating that the price increases linearly with age. At first, it may seem that the plot does not make sense or that the relationship is very weak but this behavior could be an indication that other variables may be affecting the relationship between price and age, causing the three distinct trends to develop.

Price by body mass index
In this plot the data is very scattered around what it appears to be two (2) different clusters. One in which the price is not affected by body mass index (or showing a very weak positive trend) with a maximum of value for price of $15,000 and the other trend shows a clearer and steeper correlation, with price increasing by body mass index with maximum values of price around 50,000 dollars. This behavior could also be an indication that other variables may be playing an important role in the relationship between price and body mass index, potentially causing the separation of the data in two clusters.

Price by number of children
In this plot, the trend is very clear and shows negative correlation, with the price decreasing as more children the person has. Because we are working with quantitative discrete data, we expect to see this behavior with the price piled up around the number of children as an integer value.

b. **3 pts** What is the value of the correlation coefficient for each of the above pair of response and predictor variables? What does it tell you about your comments in part (a).

```
cor(age,price)  # correlation between age and price
```

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

```r
cor(bmi,price)  # correlation between bmi and price
```

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

```r
cor(children,price)  # correlation between children and price
```

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

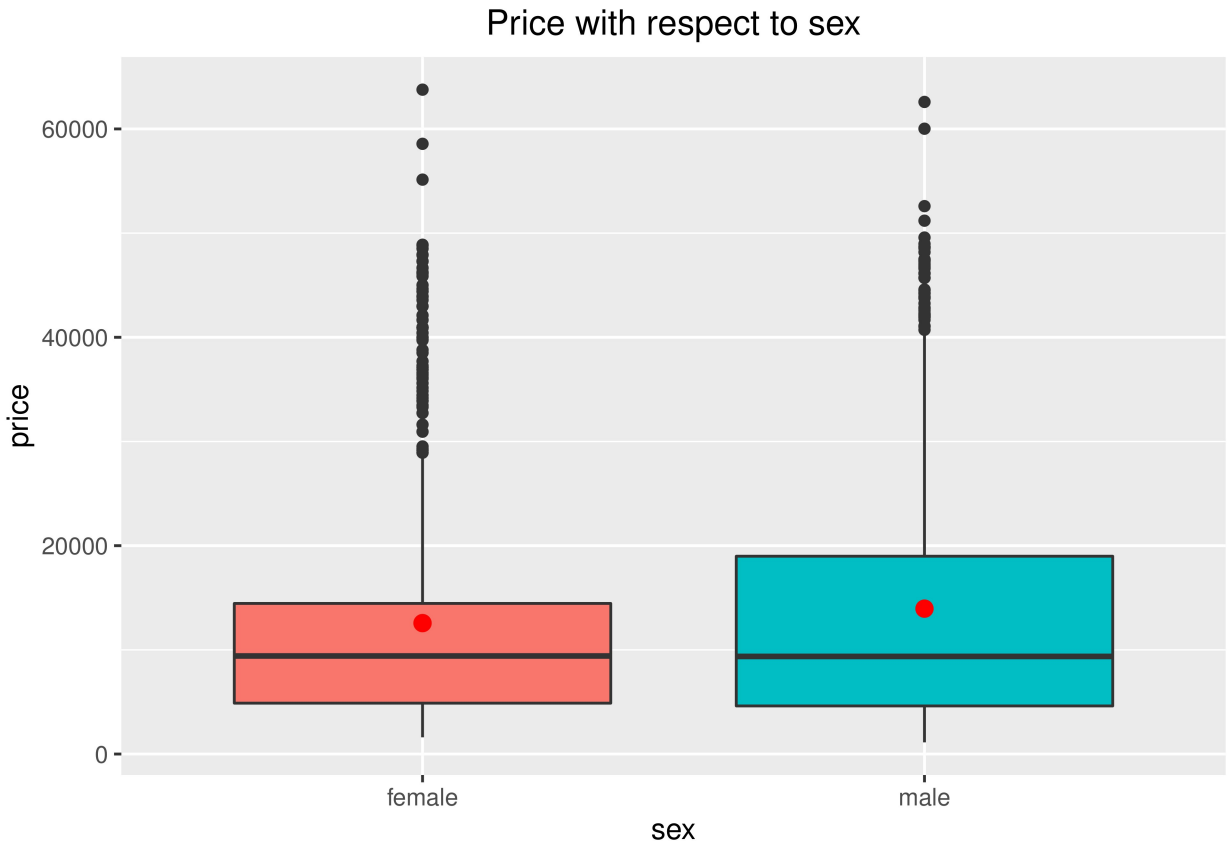
Interpretation #
The values of the correlations are all low, with varying degrees of correlation strength. For the first plot, the correlation of ~0.3 is small because it captures the global correlation in the whole data, not distinguishing each one of the three distinct trends (Which would make the correlation larger when considered in isolation). On the second plot, the correlation is small (~0.2), as the data is quite scattered. For the third plot, because the variable is not continuous but discrete (only integer numbers), the correlation number is weak because there is a large spread of prices for each category (number of children), although visually it can be seen that there is a relationship.

c. **3 pts** Create box plots of the response, *price*, and the three qualitative predictors *sex*, *smoker*, and *region*. Based on these box plots, does there appear to be a relationship between these qualitative predictors and the response?

Hint: Use the given code to convert the qualitative predictors to factors.

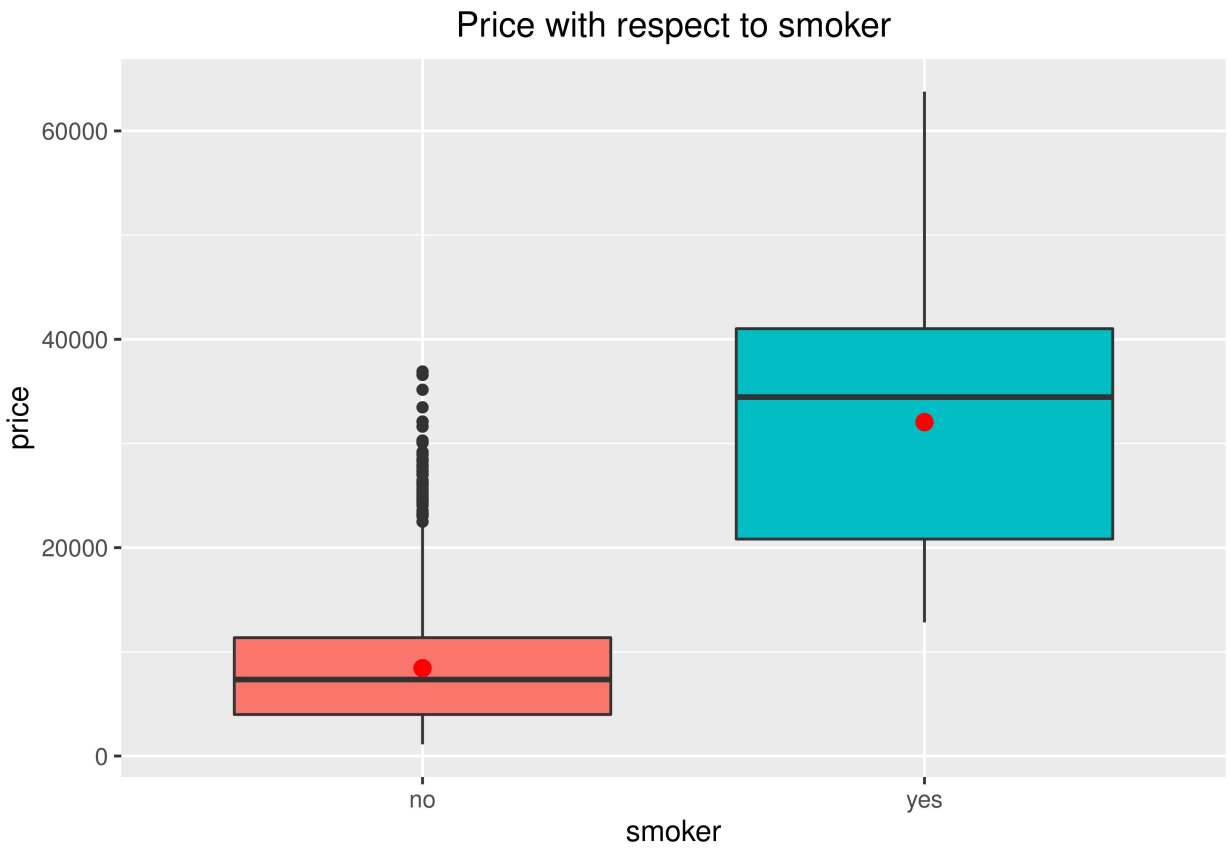
```r
#make categorical variables into factors
insurance$sex<-as.factor(insurance$sex) #makes female the baseline level
insurance$smoker<-as.factor(insurance$smoker) #makes no the baseline level
insurance$region<-as.factor(insurance$region) #makes northeast the baseline level

#Boxplots with mean

library(ggplot2)
ggplot(insurance, aes(
  x=sex, y=price, fill=sex)) + geom_boxplot(alpha = 3) +
  stat_summary(fun.y=mean, geom="point", shape=20, size=4, color="red", fill="red") +
  theme(legend.position = "none") + ggtitle("Price with respect to sex") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```



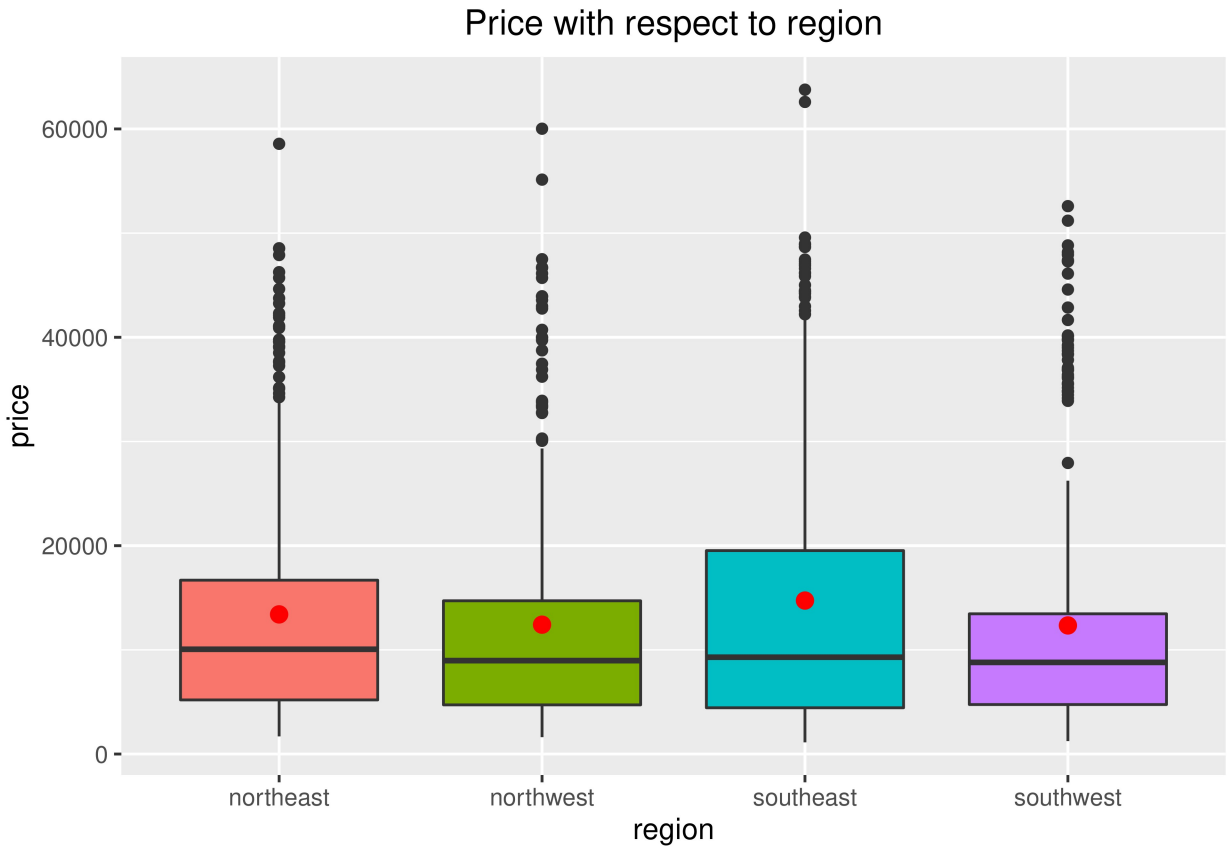

```
ggplot(insurance, aes(
  x=smoker, y=price, fill=smoker)) + geom_boxplot(alpha = 3) +
  stat_summary(fun.y=mean, geom="point", shape=20, size=4, color="red", fill="red") +
  theme(legend.position = "none") + ggtitle("Price with respect to smoker") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```




```
ggplot(insurance, aes(
  x=region, y=price, fill=region)) + geom_boxplot(alpha = 3) +
  stat_summary(fun.y=mean, geom="point", shape=20, size=4, color="red", fill="red") +
  theme(legend.position = "none") + ggtitle("Price with respect to region") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```



Price with respect to Sex
In this plot we can see that the data in males is more wide than it is in females, but the mean and the median on price is very similar on both (females and males) indicating that price is only lightly affected by the gender, with a slightly higher price in the case of males.

Price with respect to smoker
In this plot, we can see that there is a very clear difference in price when it comes to the person being a smoker or not. Smokers have a mean in price above $30,000 and non-smokers have a mean on price below 10,000 dollars.

Price by region
In this plot, we see that the medians are similar but there are some differences in the means. The Southeast has a mean price slightly above the other areas.

d. **3 pts** Based on the analysis above, does it make sense to run a multiple linear regression with all of the predictors?
Yes, it makes sense to run a linear regression model with all the predictors (as a starting point) and then identify which variables are the most important and well as the variables that are not significant.

Note: Please work on non-transformed data for all of the following questions.

Question 2: Fitting the Multiple Linear Regression Model [10 points]

Build a multiple linear regression model, named *model1*, using the response, *price*, and all 6 predictors, and then answer the questions that follow:

a. **5 pts** Report the coefficient of determination for the model and give a concise interpretation of this value.

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

Answer
The coefficient of determination is 0.75 meaning that that 75% of the response variable variation is explained by the model.

b. **5 pts** Is the model of any use in predicting price? Conduct a test of overall adequacy of the model, using $\alpha = 0.05$. Provide the following elements of the test: null hypothesis H_0, alternative hypothesis H_a, F- statistic or p-value, and conclusion.

```
#Same analysis as above, checking that gives the same results as alpha = 0.05
model1 <- lm(price ~ ., insurance, level = 0.95)
```

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'level' will be disregarded
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ ., data = insurance, level = 0.95)
```

```
##
## 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
```

Answer Is the model of any use in predicting price?What is the null hypothesis H_0 the Alternative hypothesis H_a, F- statistic,p-value, and conclusions
Yes, the model is of great use in predicting price as per its p-value which is $<$ 2.2e-16 (close to 0 and smaller than the alpha value 0.05). The F-statistic of the model is 500.8. The null hypothesis in this case is that there is no model because the regression coefficient values would be zero, in such case, I would use the mean on the observations to predict the price, while the alternative hypothesis is that the model is predictive and has at least one regression coefficient value different than zero, meaning that I can use the model in predicting the price.

Question 3: Model Comparison [12 points]

a. **4 pts** Assuming a marginal relationship between *region* and *price*, perform an ANOVA F-test on the mean insurance prices among the different regions. Using an $\alpha - level$ of 0.05, can we reject the null hypothesis that the means of the regions are equal? Please interpret.

```r
#Marginal relationship between  price and region using ANOVA
anova_marginal_model = aov(price ~ region, insurance)
model.tables(anova_marginal_model, type = "means")
```

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

```
summary(anova_marginal_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
```

Answer
Yes, we can reject the null hypothesis that the means of the regions are equal because the p-value is 0.0309, which is below to 0.05.

b. **4 pts** Now, build a second multiple linear regression model, called *model2*, using *price* as the response variable, and all variables except *region* as the predictors. Conduct a partial F-test comparing *model2* with *model1*. What is the partial-F test p-value? Can we reject the null hypothesis that the regression coefficients for *region* variables are zero at $\alpha - level$ of 0.05?

```
# Model #2 without the regions
model2 <- lm(price ~ age + sex + bmi + children + smoker, 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(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ age + sex + bmi + children + smoker + region
## Model 2: price ~ age + sex + bmi + children + smoker
```

```
##   Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
## 1   1329 4.8840e+10
## 2   1332 4.9073e+10 -3 -233431209 2.1173 0.09622 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer

What is the partial-F test and the p-value?Can we reject the null hypothesis that the regression coefficients for *region* variables are zero at $\alpha - level$ of 0.05?
The partial F-statistic is 2.1173 and the p-value is 0.09622 which is bigger than 0.05, therefore we cannot reject the null hypothesis that the regression coefficients for *region* variables are zero at $\alpha - level$ of 0.05.

c. **4 pts** What can you conclude from a and b? Do they provide the exact same results?

Answer
The two tests contradict each other. While the ANOVA F-test shows the regions means are statistically significantly different with an alpha of 0.05, the partial F-test with the regression model shows that region is not statistically significant (hence all the means would be the same). In both cases, the p-value is close to the alpha level (0.05), either 0.03 (ANOVA) or 0.096 (partial F-test), so I would interpret this as there is a potential difference between the region means and I would include it in the model, but we have a relatively low confidence that this relationship is significant.

Note: Please use model1 for all of the following questions.

Question 4: Coefficient Interpretation [6 points]

a. **3 pts** Interpret the estimated coefficient of *sexmale* in the context of the problem. *Make sure female is the baseline level for sex. Mention any assumption you make about other predictors clearly when stating the interpretation.*

Answer
The price will be 131.3 times lower than the global mean in the case of males, assuming that all the other variables are fixed. In the other hand, the price will be 131.3 times more than the global mean in the case of females.

b. **3 pts** If the value of the *bmi* in *model1* is increased by 0.01 keeping other predictors constant, what change in the response would be expected? **Answer**
If the value of bmi in model1 is increased by 0.01, assuming other predictors stay the same, the price would increase by 3.39.

Question 5: Confidence and Prediction Intervals [10 points]

a. **5 pts** Compute 90% and 95% confidence intervals (CIs) for the parameter associated with *age* for *model1*. What observations can you make about the width of these intervals?

```
confint(model1, "age", level = 0.90)
```

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

```
confint(model1, "age", level = 0.95)
```

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

Answer
With 90% of confidence the price will be between 237.2708 and 276.4419. In this case, the width of the interval is 39.17 and with 95% of confidence the price will be between 233.5138 and 280.1989 making the interval width at 46.68. We can conclude that the width is narrower when the confidence is 90% and wider with the confidence at 95%.

b. **2.5 pts** Using *model1*, estimate the average price for all insurance policies with the same characteristics as the first data point in the sample. What is the 95% confidence interval? Provide an interpretation of your results.

```
## First row of data
newdata = insurance[1,-7]
newdata
```

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

```
predict(model1, newdata, interval="confidence", level = 0.95)
```

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

Answer
The 95% confidence interval is 24143.98 - 26443.44, this means that for insurance policies with the same characteristics as the first, the average estimated prices are 25,293.71 dollars, with a lower bound of 24,143.98 dollars and an upper bound of $26,443.44 dollars.

c. **2.5 pts** Suppose that the *age* value for the first data point is increased to 50, while all other values are kept fixed. Using *model1*, predict the price of an insurance policy with these characteristics. What is the 95% prediction interval? Provide an interpretation of your results.

```
#Changing only age to 50 in the first row of data
newdata_2 = replace(insurance[1,-7],1, 50)
newdata_2
```

```
##   age    sex  bmi children smoker    region
## 1  50 female 27.9        0    yes southwest
```

```
## Prediction Interval
predict(model1, newdata_2, interval="prediction", level = 0.95)
```

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

Answer
The 95% prediction interval is 21313.29 - 45199.23 and this means that if the age of the applicant is 50, the predicted price would be 33,256.26 dollars.