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

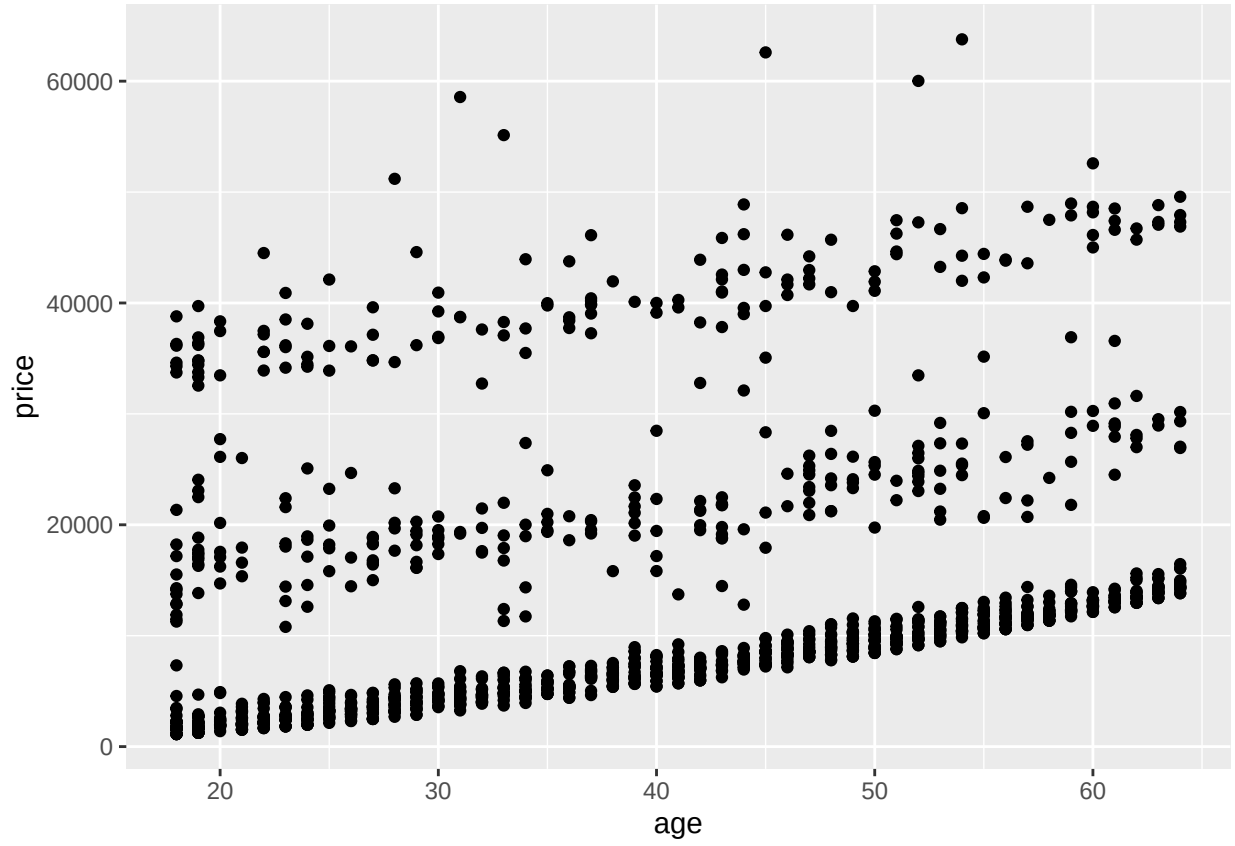
To read the data in `R`, save the file in your working directory (make sure you have changed the directory if different from the R working directory) and read the data using the `R` function `read.csv()`

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

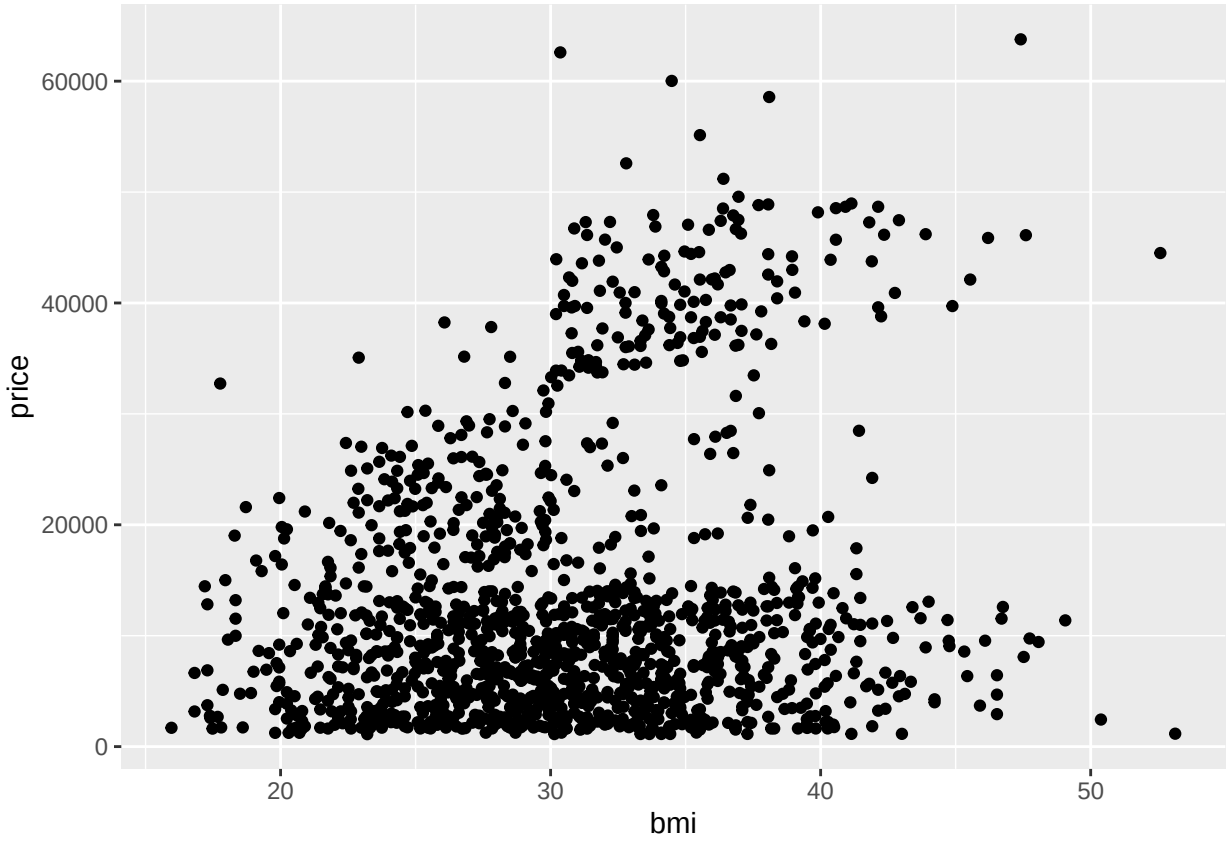
Question 1: Exploratory Data Analysis [15 points]

a. **4 pts** Create scatterplots of the response, *price*, against three quantitative predictors *age*, *bmi*, and *children*. Describe the general trend (direction and form) of each plot. It should be 3 separate scatter plots.

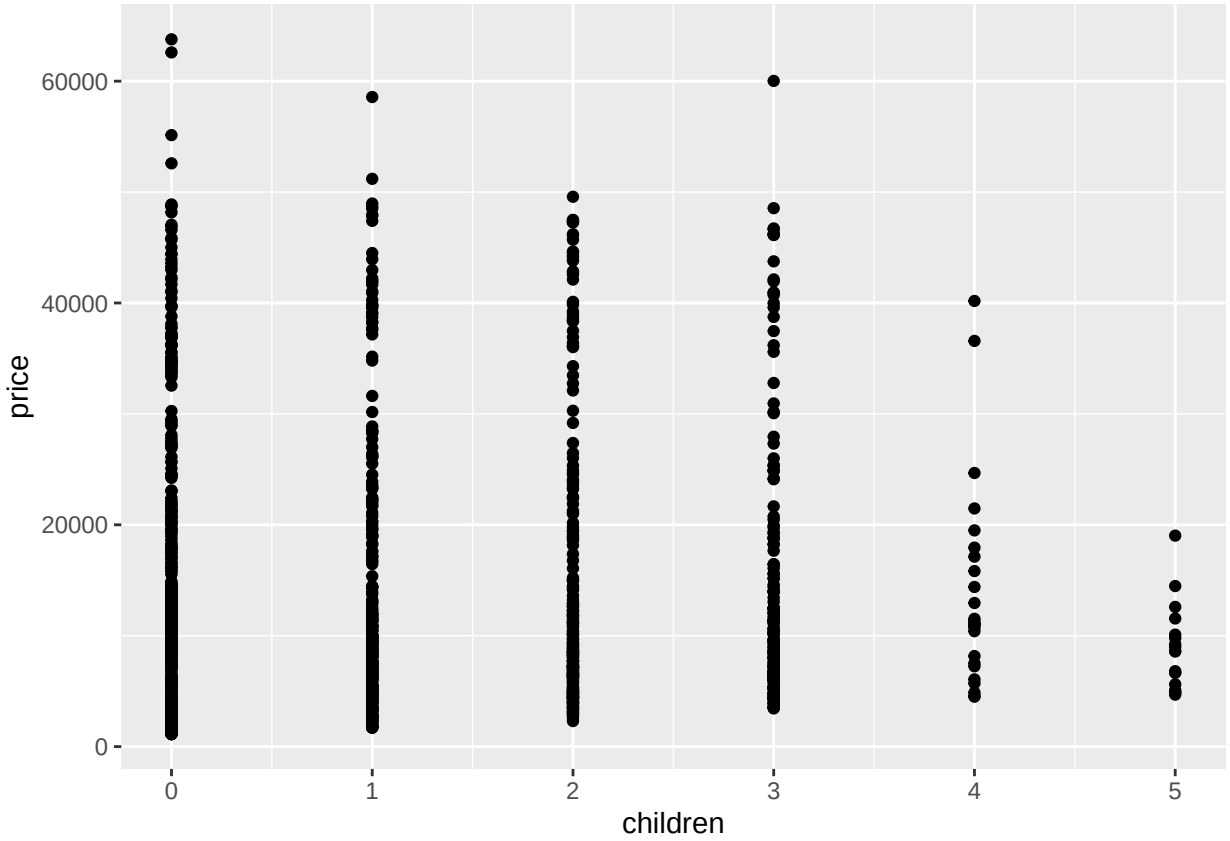
```
library(ggplot2)
age_plot <- ggplot(insurance, aes(x=age, y=price)) + geom_point()
bmi_plot <- ggplot(insurance, aes(x=bmi, y=price)) + geom_point()
child_plot <- ggplot(insurance, aes(x=children, y=price)) + geom_point()
age_plot
```




```
bmi_plot
```




```
child_plot
```



For the age plot we can see a slight positive linear relationship.

For the bmi plot we can see a positive linear relationship. However, we can also see that the data for bmi is mostly from BMI 20 to 40. Within this range we can see that it is not always a positive linear relationship and there may be other predictors or factors influencing the price.

For the child plot there is no relationship or possibly slight decrease in price as the number of children increases as shown in 5 children having the lowest price.

b. **4 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)?

```
price = insurance$price
age = insurance$age
bmi = insurance$bmi
child = insurance$children
cor(age,price)
```

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

```
cor(bmi, price)
```

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

```
cor(child,price)
```

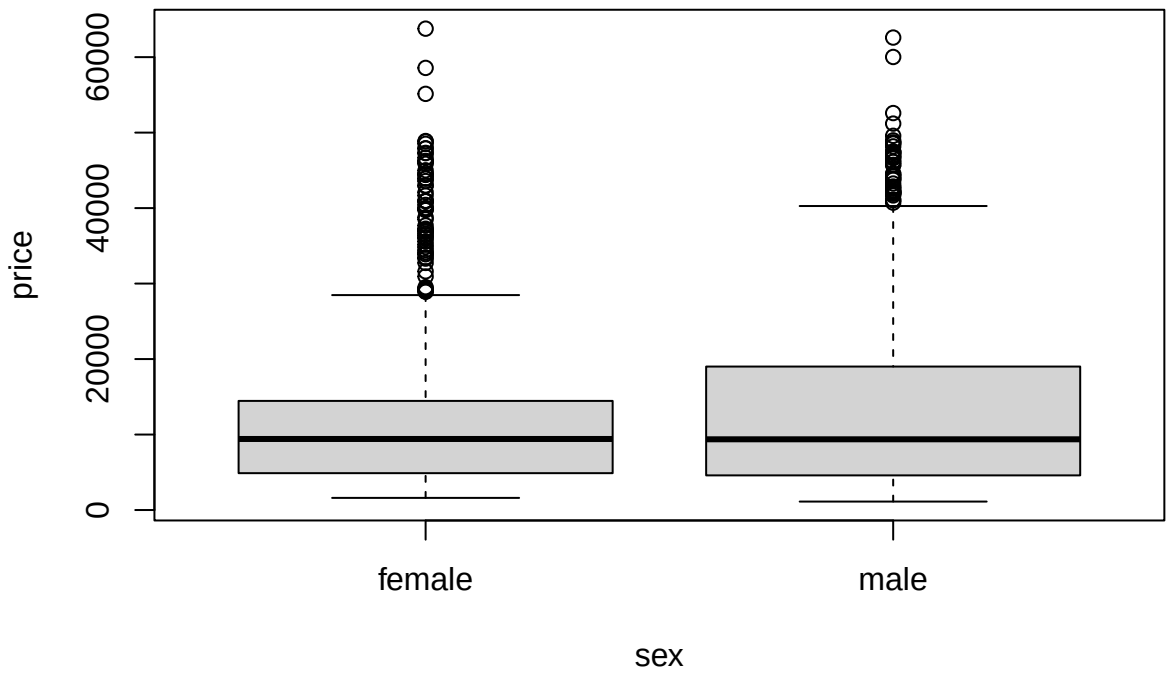
```
## [1] 0.06799823
```

The correlation coefficient for age, bmi and children is 0.299, 0.198, and 0.067 respectively. These values explain a positive correlation, with age having the strongest correlation with price and children having the least. The values for the correlation coefficient agrees with my comments in part(a).

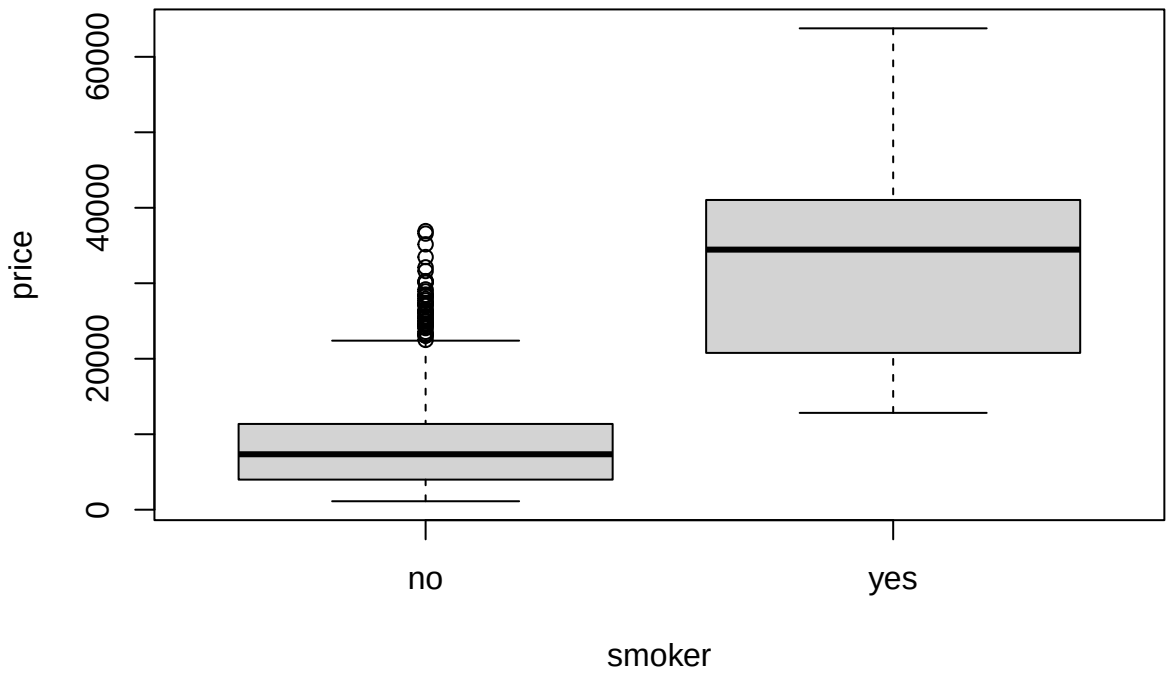
c. **4 pts** Create box plots of the response, *price*, and the three qualitative predictors *sex*, *smoker*, and *region*. Based on these box plots, does there appear to be a relationship between these qualitative predictors and the response?

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

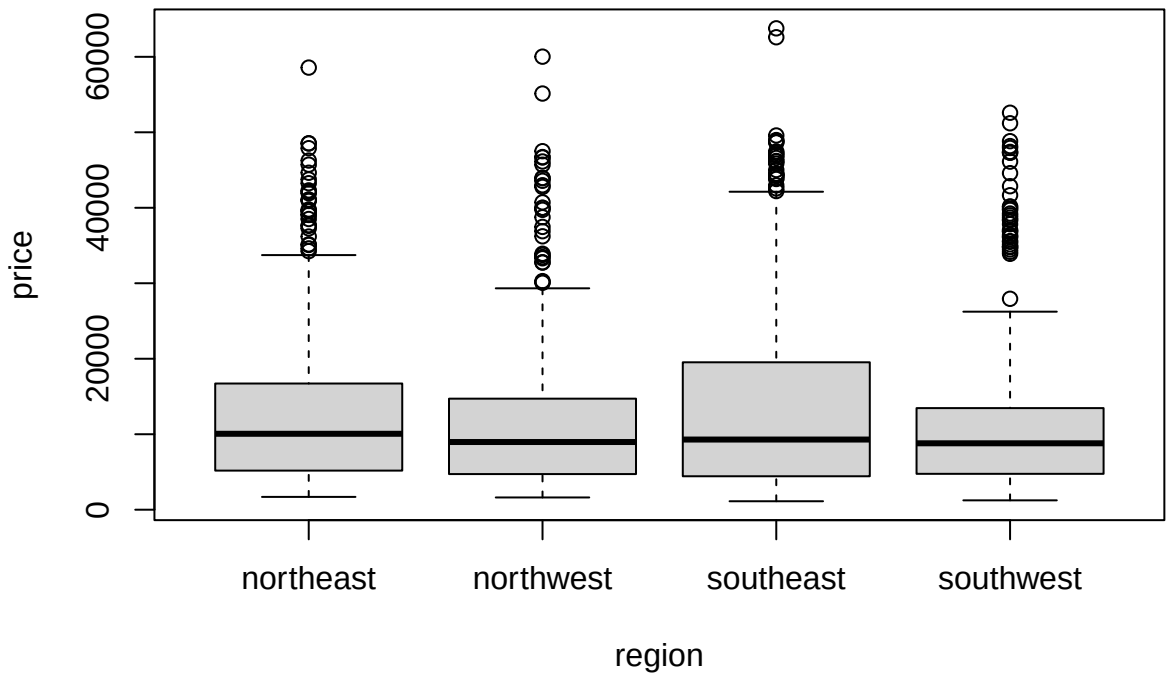
```
#make categorical variables into factors
insurance$sex<-as.factor(insurance$sex) #makes female the baseline level
insurance$smoker<-as.factor(insurance$smoker) #makes no the baseline level
insurance$region<-as.factor(insurance$region) #makes northeast the baseline level
boxplot(price~sex, data =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




```
boxplot(price~smoker, data = insurance)
```




```
boxplot(price~region, data = insurance)
```



There is a little to no relationship between price/sex and price/region. We do see a relationship between smoker/price as the price of the insurance greatly increases if the person is a smoker.

d. **3 pts** Based on the analysis above, does it make sense to run a multiple linear regression with all of the predictors?

Yes, we can run a full multiple linear regression model and then use a partial F-test or remove predictor variables to see which predictors have strong influence on response variable.

Question 2: Fitting the Multiple Linear Regression Model [12 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. **6 pts** Report the coefficient of determination (R-squared) for the model and give a concise interpretation of this value.

```
model1 <- lm(price ~ age+sex+bmi+children+smoker+region, data = insurance)
summary(model1)
```

```
##
## Call:
```

```
## lm(formula = price ~ age + sex + bmi + children + smoker + region,
##     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
```

The R square is 0.7509. Thus, about 75.09% of the variability in Price is explained by the predicting variables considered for this model.

b. **6 pts** Is the model of any use in predicting price? Using $\alpha = 0.05$, provide the following elements of the test of overall regression of the model: null hypothesis H_0, alternative hypothesis H_a, F-statistic or p-value, and conclusion.

H_0: B_1=B_2=B_3=B_4=B_5=B_6=0

H_a: At least one $B_i \neq 0$

F-Statistic = 500.8 on 8 and 1329 DF

p-value: 2.2eˆ-16 < 0.05

We reject H_0, that all the predictors are zero as the p-value is less than $\alpha = 0.05$. Therefore, at least one of the predictors is statistically significant for the response variable.

Question 3: Model Comparison [14 points]

a. **5 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 α-level of 0.05, can we reject the null hypothesis that the means of the regions are equal? Please interpret.

```
model1_aov <- aov(price~region, data = insurance)
summary(model1_aov)
```

```
##               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
```

Yes, p-value $= 0.0309 < \alpha = 0.05$, so we can reject the null hypothesis that the mean insurance price are equal among the regions.

b. **5 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 an α-level of 0.05?

```
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
```

```
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
```

P-value $= 0.09622 > \alpha = 0.05$ means we do not reject the null hypothesis that the mean insurance price are equal among the regions.

c. **4 pts** What can you conclude from 3a and 3b? Do they provide the exact same results?

3a rejects the null hypothesis of equal means and 3b does not reject the null hypothesis that the regression coefficients for region are zero. Which means that 3a and 3b do not provide the same results. From this we can see that by removing region we changed the P-value, giving region a strong relationship and influence with price.

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

Question 4: Coefficient Interpretation [7 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 assumptions you make about other predictors clearly when stating the interpretation.*

The coefficient for 'sexmale' is -131.3. This means that if we fix the other predictors coefficient values, for each unit of increase in sexmale predictor, the price(response) of insurance will decrease by 131.3. However, this predictor value is is not significant to the when compared to the other predictors as P-value is the highest among the predictors.

b. **4 pts** If the value of the *bmi* in *model1* is increased by 0.01 and the other predictors are kept constant, what change in the response would be expected?

This increase of 0.01 * 339.2(bmi) = $3.392 increase in the response variable if all other predictors are held constant.

Question 5: Confidence and Prediction Intervals [12 points]

a. **6 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
```

The width at 95% CI is wider than the width at 90% CI. There is a 90% chance that the value of age coefficient will be between 237.27 and 276.44. There is a 95% chance that the value of age coefficient will be between 233.51 and 280.19. This means that the range of in upper and lower bound increases as the confidence interval increases.

b. **3 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 = insurance[1,1:6]
predict(model1, first, interval = 'confidence', level = 0.95)
```

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

The price of insurance given that the predictor values stay the same will be between an upper bound of 26443.44 and lower bound of 24143.98 with a predicted price of 25293.71.

c. **3 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.

```
first[1] = 50
predict(model1, first, interval = 'confidence', level = 0.95)
```

```
##        fit      lwr      upr
## 1 33256.26 32157.63 34354.89
```

The price of insurance given that the age = 50 and the other predictor values stay the same will be between an upper bound of 34354.89 and lower bound of 32157.63 with a predicted price of 33256.26.

The cost of insurance increased when age increased to 50.