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)
head(insurance)
```

```
##   age    sex    bmi children smoker    region     price
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Question 1: Exploratory Data Analysis [15 points]

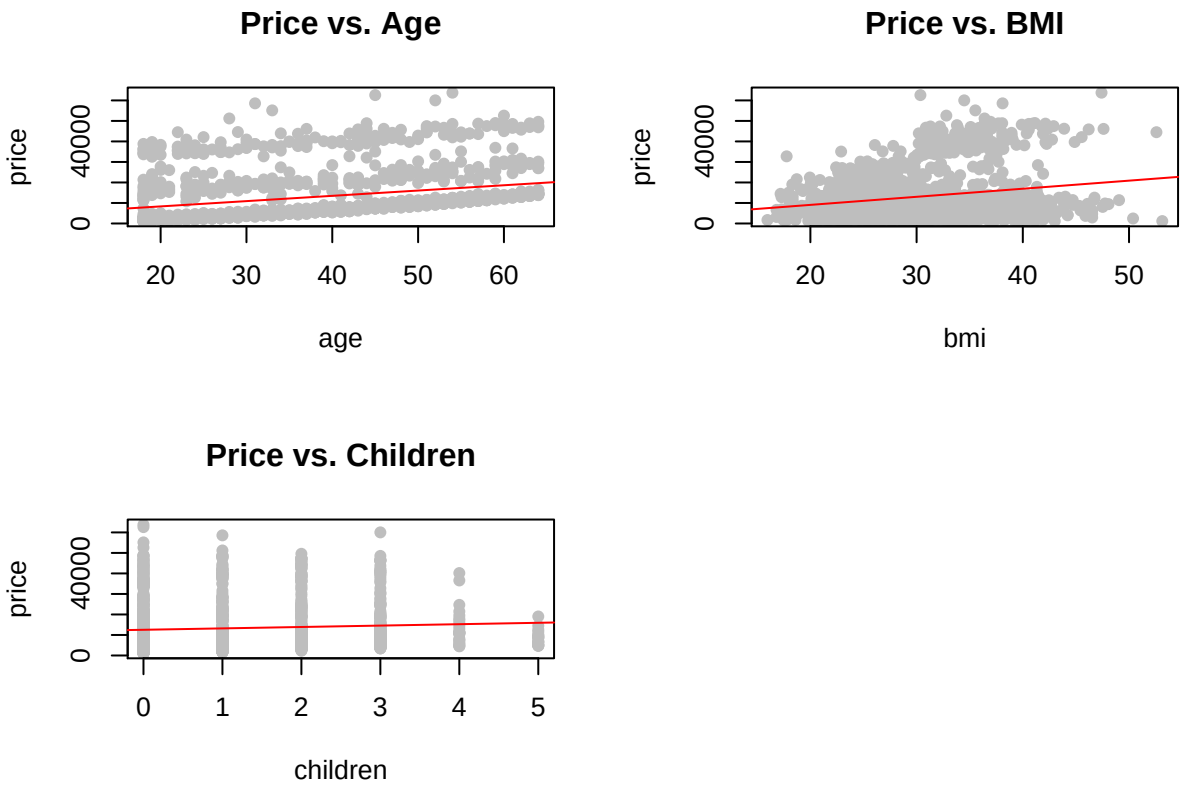
a. **4 pts** Create scatterplots of the response, *price*, against three quantitative predictors *age*, *bmi*, and *children*. Describe the general trend (direction and form) of each plot. It should be 3 separate scatter plots.

```r
# Grid the plots
par(mfrow=c(2,2))

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

# Plot price vs bmi
plot(price~bmi, data=insurance, main="Price vs. BMI",col="grey", pch = 16)
abline(lm(price~bmi, data=insurance), col="red")

# Plot price vs children
plot(price~children, data=insurance,
     main="Price vs. Children",col="grey", pch = 16)
abline(lm(price~children, data=insurance), col="red")
```



General trend:

There appears to be a positive, linear relationship of weak strength between the response, *price*, and the predictor variables *age* and *bmi*. From the scatter plot of *price* vs. *bmi*, we can see that as the bmi increases the variance of the insurance price appears to increase as well.

Additionally, there seems to be a very weak, almost non-existent, positive linear relationship between the response, *price*, and the predictor variable *children*.

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)?

```r
# Print the correlation coefficients between the predictors and the response
cat("cor(price, age):", cor(insurance$price, insurance$age)[1], end="\n")
```

```
## cor(price, age): 0.2990082
```

```r
cat("cor(price, bmi):", cor(insurance$price, insurance$bmi)[1], end="\n")
```

```
## cor(price, bmi): 0.198341
```

```r
cat("cor(price, children):", cor(insurance$price, insurance$children)[1],
    end="\n")
```

```
## cor(price, children): 0.06799823
```

The correlation coefficient between *price* and *age* of 0.2990082, although the highest of the group, suggests a weak positive linear relationship between the two variables. The correlation coefficient between *price* and *bmi* (0.198341) shows a slighter positive linear relationship, and the correlation coefficient between *price* and *children* (0.06799823) shows that there is almost no relationship between the two variables. These results reinforce that our comments in part (a) were correct. Outside of that, our analysis aligns with our hypothesis that the response is positively correlated with each of the predictor variables.

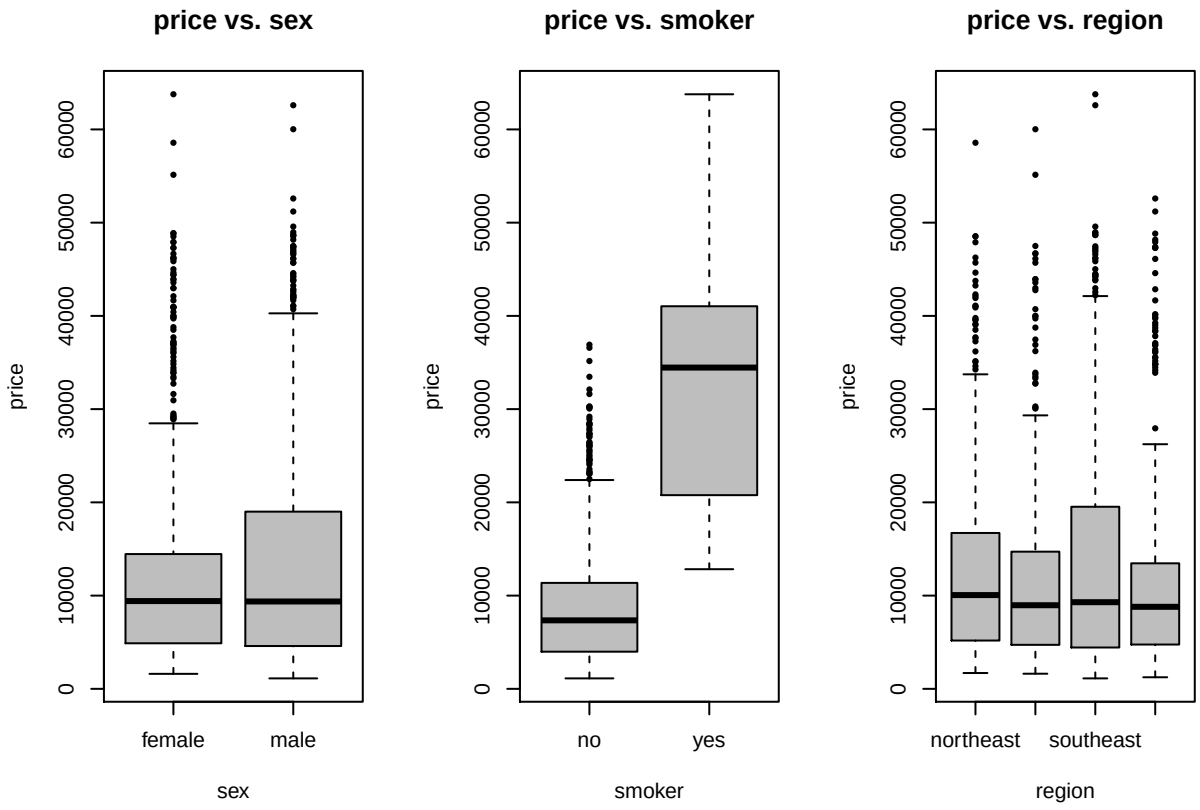
c. **4 pts** Create box plots of the response, *price*, and the three qualitative predictors *sex*, *smoker*, and *region*. Based on these box plots, does there appear to be a relationship between these qualitative predictors and the response?

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

```r
par(mfrow=c(1,3))

#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

# Plot price vs sex
plot(price~sex, data=insurance, main="price vs. sex", col="gray", pch = 16)
# Plot price vs smoker
plot(price~smoker, data=insurance, main="price vs. smoker", col="gray",pch = 16)
# Plot price vs region
plot(price~region, data=insurance, main="price vs. region", col="gray",pch = 16)
```



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. Hence, there does appear to be a marginal relationship between *price* and *smoker*, but not between *price* and *sex* or *price* and *region*. We need further quantitative analysis to confirm this.

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

Yes. Based on this initial assessment, running a multiple linear regression model appears to be reasonable. 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.

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.

```r
# Fit model1
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
```

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

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

R^2 is 0.750913 or 75.09%. We can interpret this as 75.09% of the variation in the response (insurance prices) is explained by the predictors in the 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.

```r
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:

$H_0 : \beta_1 = ... = \beta_8 = 0$

H_a: At least one of the slope coefficients is nonzero

Test statistic: $F(8, 1329) = 500.8$

p-value: $<$ 2.2e-16

Conclusion: Since $\alpha = 0.05$ exceeds the observed significance level, $p <$ 2.2e-16, we reject the null hypothesis. The data provide strong evidence that at least one of the slope coefficients is nonzero. The overall model appears to be statistically useful in predicting price.

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.

```
# One-way ANova
anova_mod = aov(price ~ region, insurance)
summary(anova_mod)
```

```
##               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
```

We conduct an Anova F-Test on the mean insurance prices among the different regions. The elements of the test are as follows:

$H_0 : \mu_1 = ... = \mu_4$

H_a: At least 2 of the population means are not equal

Test statistic: $F(3, 1334) = 2.97$

p-value: 0.0309

The p-value of the F-test is 0.0309, which is less than the α-level of 0.05. We reject the null hypothesis that the mean insurance prices for different regions are equal, and conclude that the mean insurance price in at least one region is different than the means in the other regions at α-level of 0.05.

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?

```r
# Fit model2
model2 <- lm(price ~ .-region, data=insurance)
summary(model2)
```

```
##
## Call:
## lm(formula = price ~ . - region, 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
```

```r
# Partial F-test
anova(model2, model1)
```

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

$H_0 : \beta_{regionnorthwest} = \beta_{regionsoutheast} = \beta_{regionsouthwest} = 0$

H_a: At least one of $\beta_{regionnorthwest}, \beta_{regionsoutheast}, \beta_{regionsouthwest}$ is not equal to zero.

Test statistic: $F_{partial}(3, 1329) = 2.1173$

p-value: 0.09622

The p-value for the partial F-test is 0.09622 which is more than the α-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 α-level of 0.05.

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

Parts 3a and 3b provide different results. The marginal model in part(a) captures the association of the variable region to the response variable marginally, hence without considering the other factors; while in part (b), we examine whether or not region adds explanatory power in addition to the other variables: *age*, *sex*, *bmi*, *children*, and *smoker* being in the model.

We concluded from part a) 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. We concluded from part b) that *region* did not add explanatory power given that all other predictors were already included in the model.

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.*

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

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

Note that 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.

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?

```
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.

Note, that an answer of 3.391935 is enough to receive full credit for this question.

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?

```
#For the 90% CI:
confint(model1, "age", level = 0.9)
```

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

```r
#For the 95% CI:
confint(model1, "age", level = 0.95)
```

```
##        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.

Zero is not included in neither of the intervals, which implies that the coefficient for age is statistically significant at both significance levels.

Note: One observation about the width of these intervals should be sufficient to receive full credit for this question

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.

```r
# Extract first data point
data1 <- insurance[1,1:6]
# Get the estimation and confidence interval
predict(model1, data1, interval="confidence")
```

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

The estimated average price of all insurance policies with the same characteristics as the first data point in the sample is $25,293.71. The 95% confidence interval would be $24,143.98 for the lower bound and $26,443.44 for the upper bound. 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. Note that 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.

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.

```r
# Update first data point
data1[1] <- 50
# Get the estimation and prediction interval
predict(model1, data1, interval="prediction")
```

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

The predicted price of an insurance policy with the above characteristics would be $33,256.26. The 95% prediction interval would be $21,313.29 for the lower bound and $45,199.23 for the upper bound. We can be 95% confident that the price of an insurance policy with the above characteristics is between $21,313.29 and $45,199.23.