M3-Part 1: Multiple Linear Regression

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
```

```
summary(insurance)
```

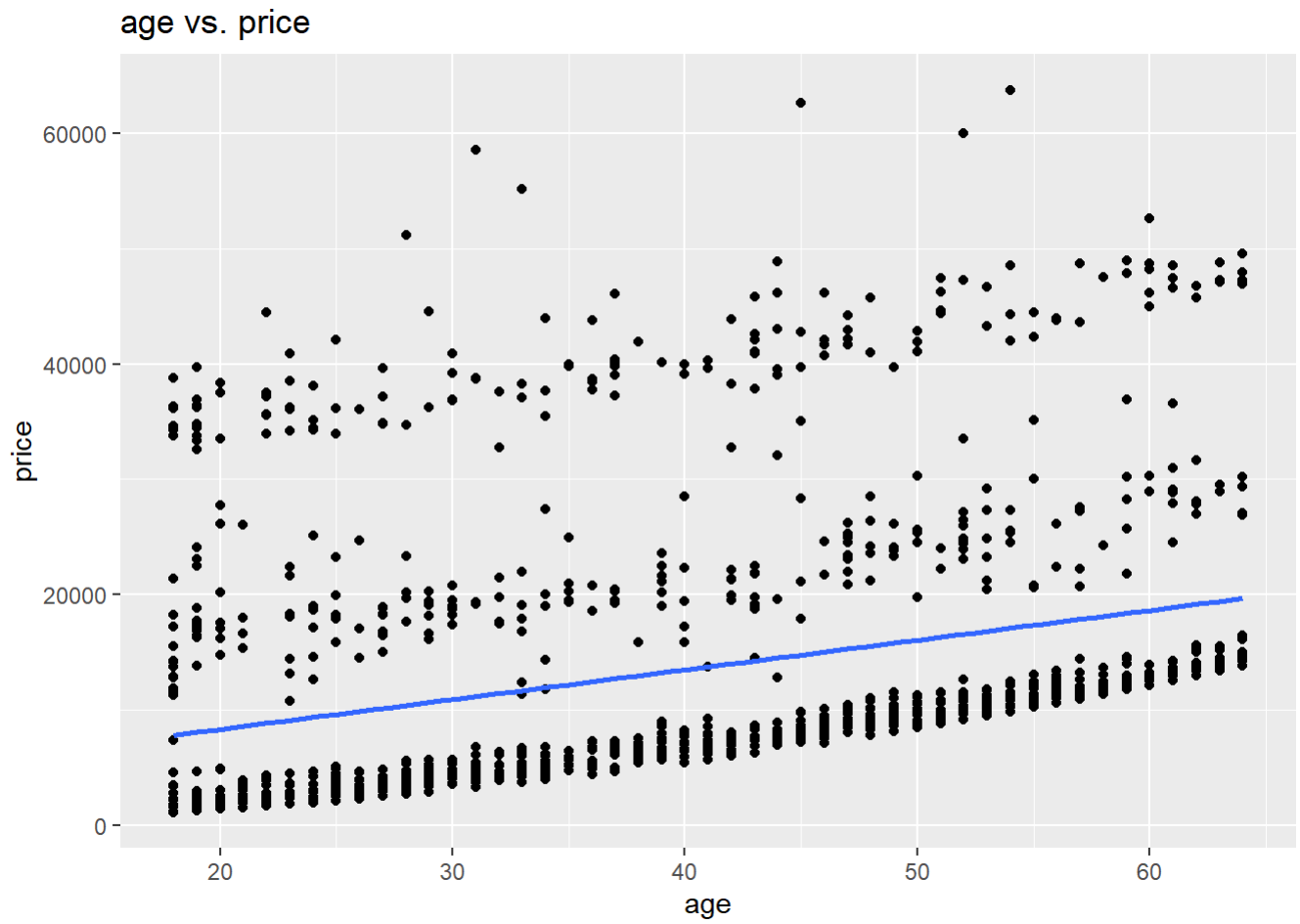
```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
##     smoker             region             price
##  Length:1338        Length:1338        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

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.

```
library(ggplot2)

ggplot(insurance, aes(x=age, y=price)) + geom_point() + ggtitle("age vs. price") +
  scale_colour_hue(l=50) + # Use a slightly darker palette than normal
  geom_smooth(method="lm",   # Add linear regression line
              se=FALSE,    # Don't add shaded confidence region
              fullrange=TRUE) # Extend regression lines
```

```
## `geom_smooth()` using formula 'y ~ x'
```

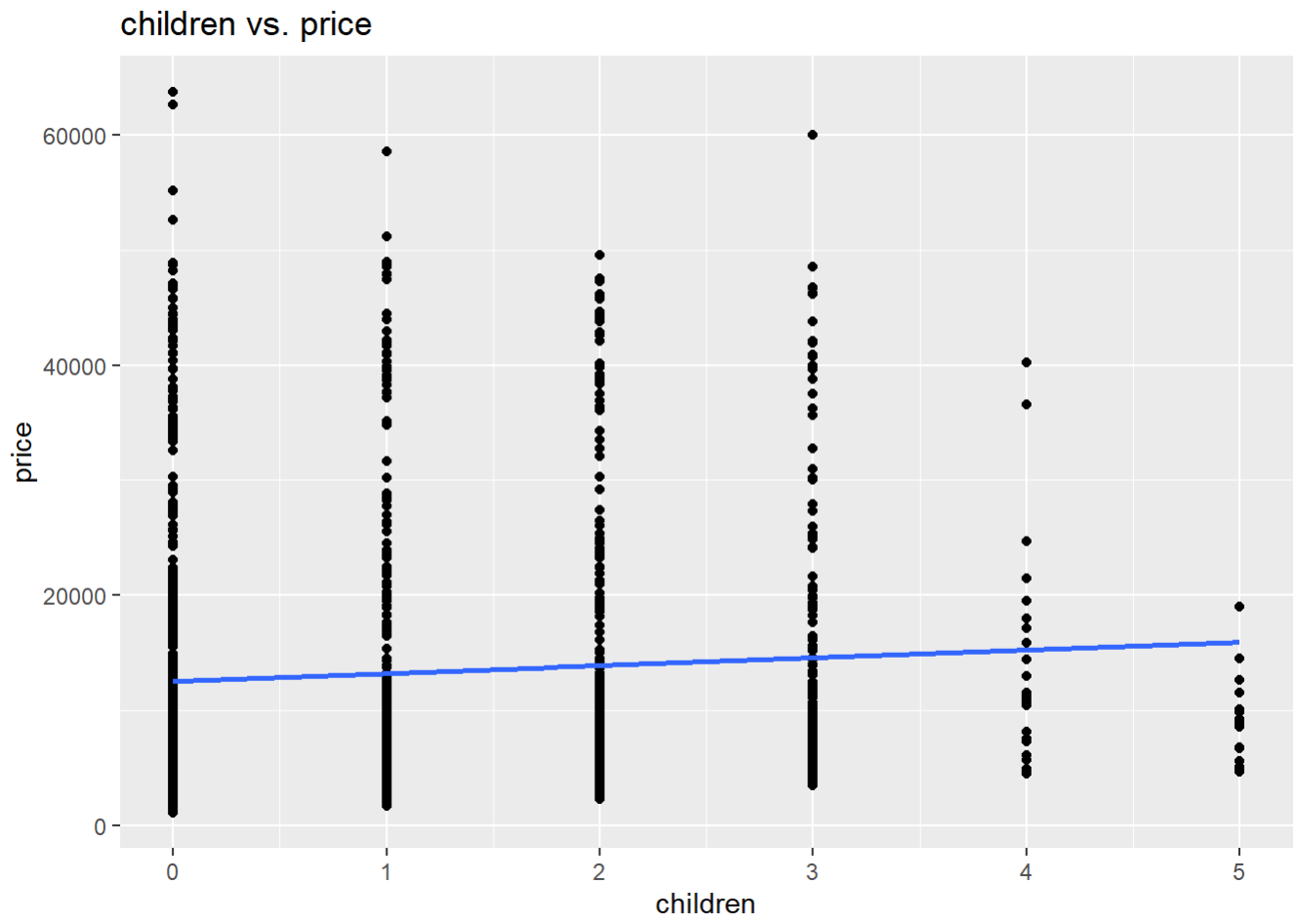



```
ggplot(insurance, aes(x=bmi, y=price)) + geom_point() + ggtitle("bmi vs. price") +
  scale_colour_hue(l=50) + # Use a slightly darker palette than normal
  geom_smooth(method="lm",   # Add linear regression line
              se=FALSE,    # Don't add shaded confidence region
              fullrange=TRUE) # Extend regression lines
```

```
## `geom_smooth()` using formula 'y ~ x'
```


```
ggplot(insurance, aes(x=children, y=price)) + geom_point() + ggtitle("children vs. price") +
  scale_colour_hue(l=50) + # Use a slightly darker palette than normal
  geom_smooth(method="lm",   # Add linear regression line
              se=FALSE,    # Don't add shaded confidence region
              fullrange=TRUE) # Extend regression lines
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Answer

We can see there is a weak direct relationship between the predictors *age* and *bmi* and the response *price*, meaning that as the predictor increases, there is a slight increase in price. Looking at the relationship between *children* and *price*, we can see an ***extremely*** weak direct relationship. The relationship between the predictor *children* and *price* is much lower than the other two predictors as we can see in the *age* and *bmi*'s steeper slope of the regression line while the *children*'s slope is almost 0.

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).

```
r_age = cor(insurance$age, insurance$price)
print(paste("The correlation coefficient between age and price is: ", r_age))
```

```
## [1] "The correlation coefficient between age and price is:  0.299008193330648"
```

```
r_bmi = cor(insurance$bmi, insurance$price)
print(paste("The correlation coefficient between bmi and price is: ", r_bmi))
```

```
## [1] "The correlation coefficient between bmi and price is:  0.198340968833629"
```

```
r_children = cor(insurance$children, insurance$price)
print(paste("The correlation coefficient between children and price is: ", r_children))
```

```
## [1] "The correlation coefficient between children and price is:  0.0679982268479048"
```

Answer

The value of the age correlation coefficient is: 0.299 (~ 0.3)

The value of the bmi correlation coefficient is: 0.198 (~ 0.2)

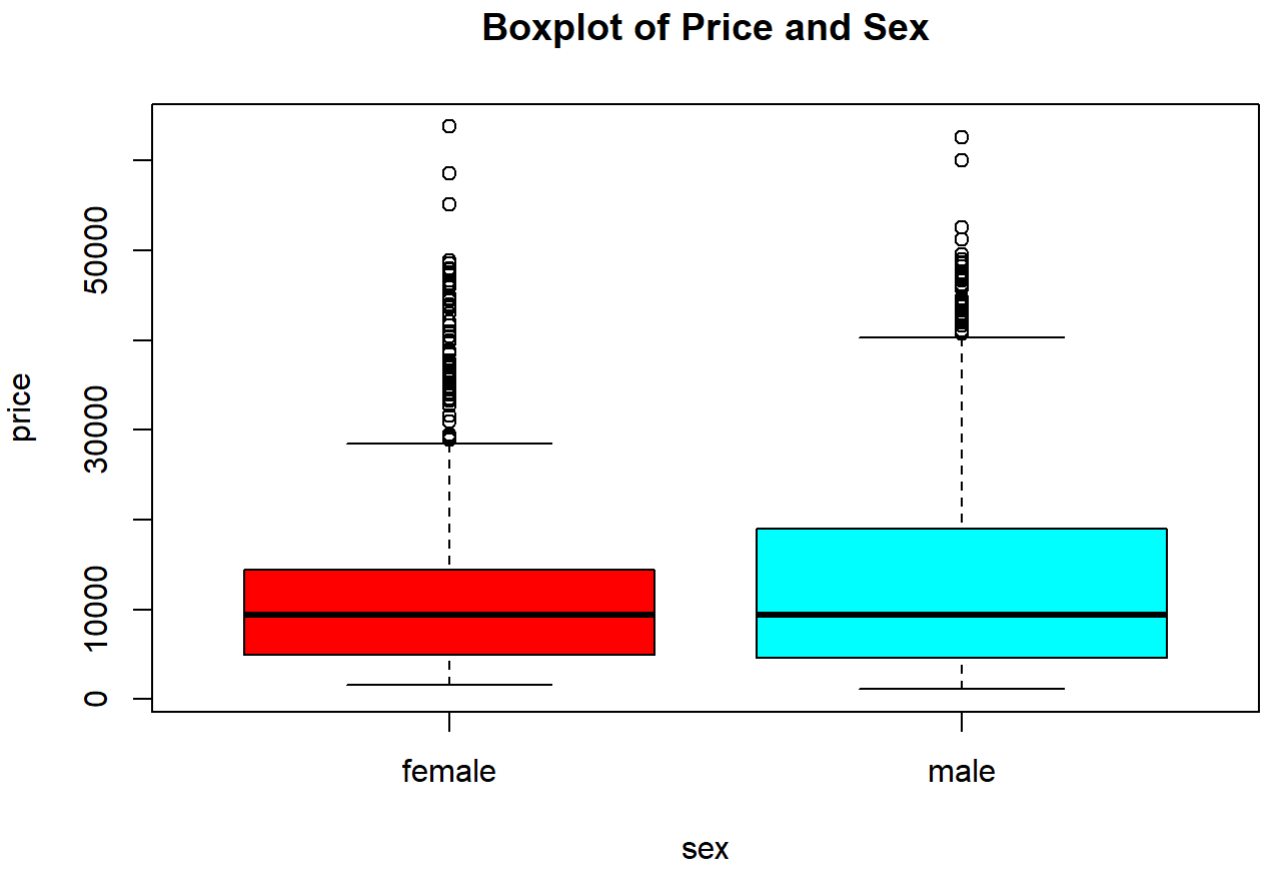
The value of the children correlation coefficient is: 0.0679 (~ 0.068)

As observed in part a, the relationship between the predictors *age* and *bmi* and the response *price* is a weak direct relationship as indicated by the correlation coefficients of 0.3 and 0.2 for *age* and *bmi*, respectively. Moreover, we can see the extremely weak relationship between *children* and *price* as indicated by the correlation coefficient around 0. The relationship between *children* and *price* is so weak, indicated by the correlation coefficient of 0.068, that it can be argued that there is no relationship.

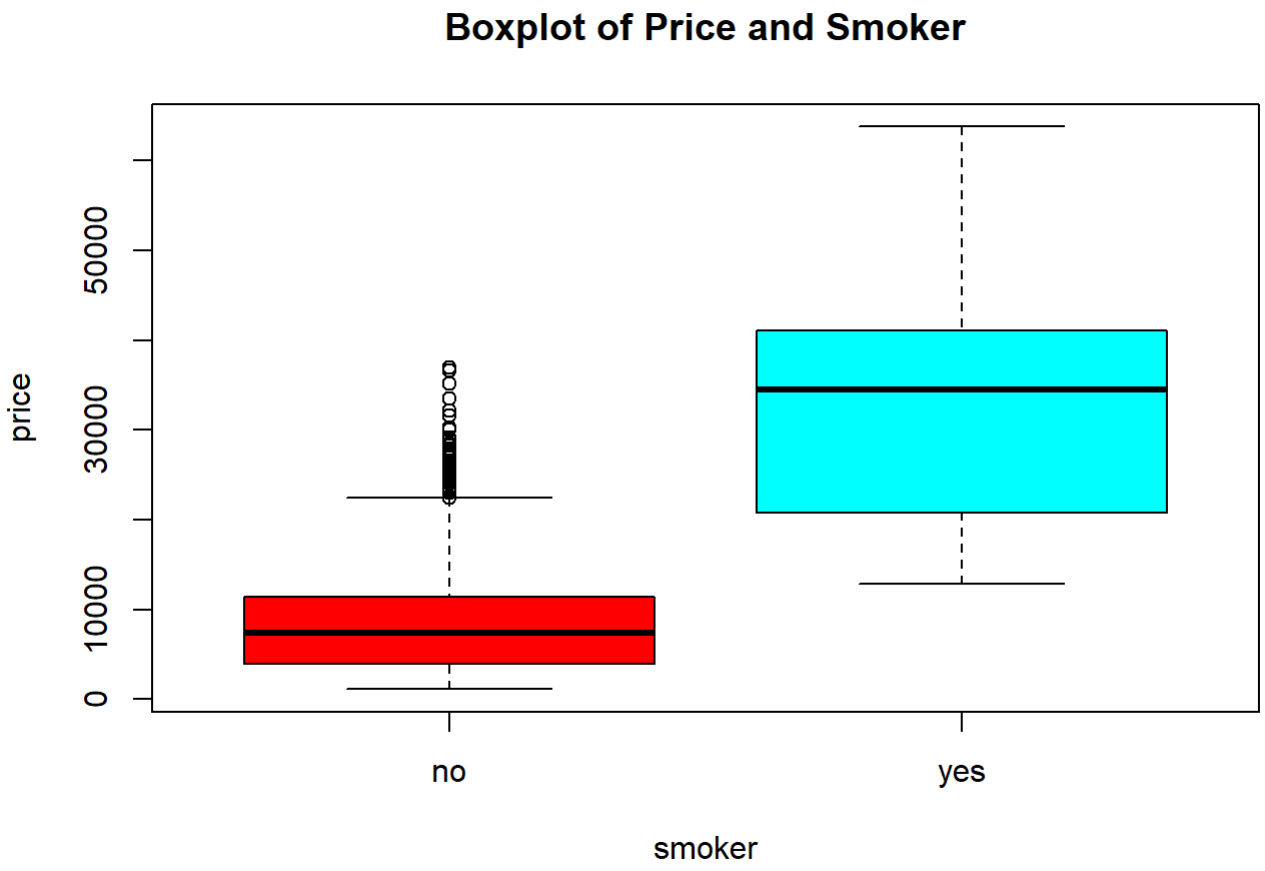
c. **3 pts** Create box plots of the response, *price*, and the three qualitative predictors *sex*, *smoker*, and *region*. Based on these box plots, does there appear to be a relationship between these qualitative predictors and the response?

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

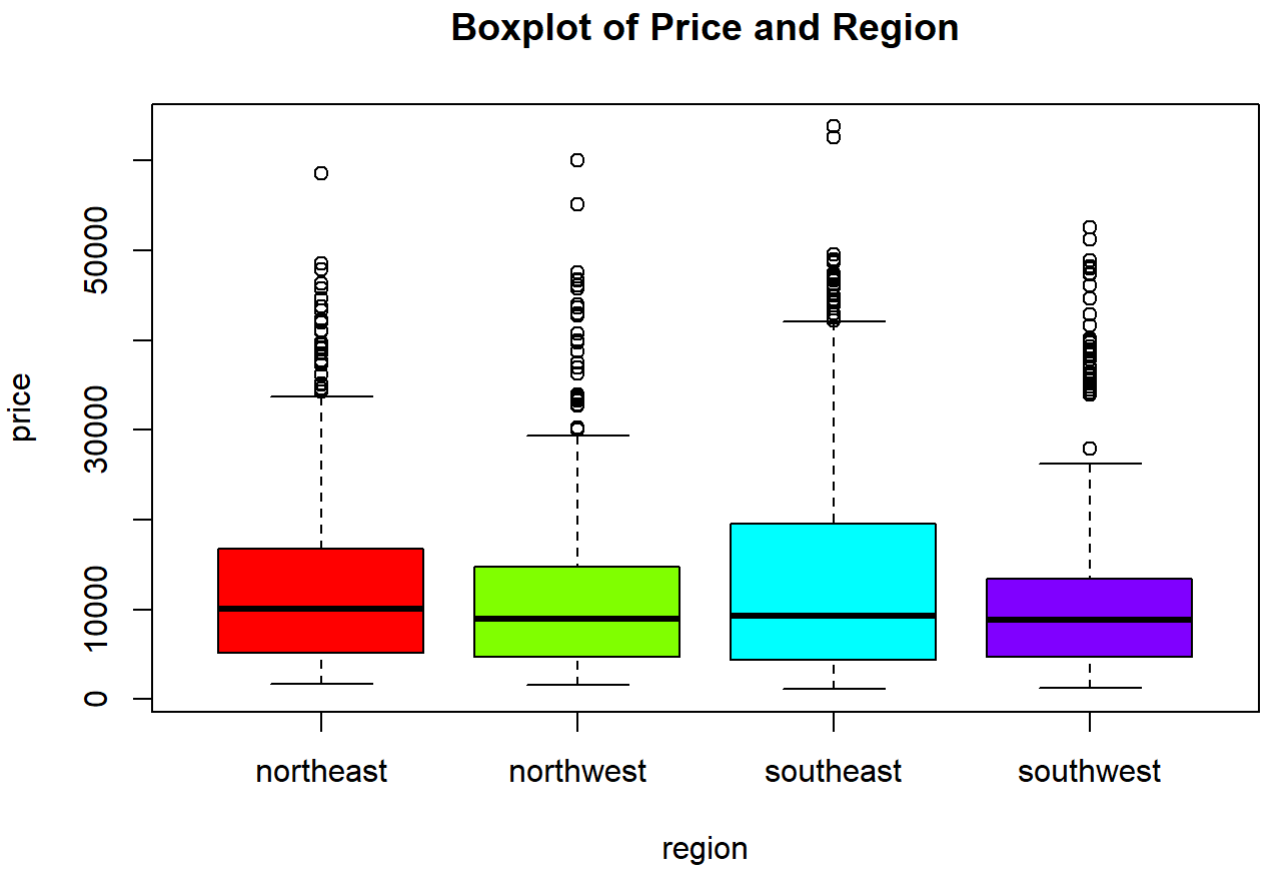
```
#make categorical variables into factors
insurance$sex<-as.factor(insurance$sex) #makes female the baseline level
insurance$smoker<-as.factor(insurance$smoker) #makes no the baseline level
insurance$region<-as.factor(insurance$region) #makes northeast the baseline level

boxplot(price ~ sex, data = insurance, main="Boxplot of Price and Sex", col= rainbow(2))
```




```
boxplot(price ~ smoker, data = insurance, main="Boxplot of Price and Smoker", col= rainbow(2))
```




```
boxplot(price ~ region, data = insurance, main="Boxplot of Price and Region", col= rainbow(4))
```



Answer

Based on the box plots above we see *sex* does not have much effect on price since, while price of male has more variability (range of price male pay for insurance is wider than female), the mean of price seems to be the same for male and female. *Region* also does not seem to be affecting price since the mean of *price* seems to be about the same, even with different price variability between the regions. However, price of insurance does increase significantly for smokers as we can see in the *smoker* box plot that the mean of *price* is much greater for smoker than for non-smoker.

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

Answer

Based on the analysis above age, bmi and smoker should be included in a multiple linear regression model. It seems like the region and sex do not have such a strong predictive power since there is not much difference in the price mean when looking at male-female and the four different regions. But since the variability is different, meaning price variability of male is greater than price variability of female, and price variability of region southeast is greater than the other three regions, it might be worthwhile to include those factors in the multiple linear regression model and assess their statistical significance. Regarding the children predictor, since there is a relationship, even though weak, I would also suggest to include in the model and assess for significance.

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 ~ ., data = 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, or R^2, is the proportion of total variability in Y that can be explained by the linear regression model, or the amount of variance accounted for in the relationship between two (or more) variables i.e. our response variable price and the predictors. Simply put, as R^2 increases, the Y values would be closer to the regression line/ plane. R^2 is calculated as:

R^2 = SSR / SST = 1 - (SSE - SST), where:

SST = sum of squares total = total deviation

SSE = sum of squared errors = unexplained deviation

SSR = sum of squares for regression = explained deviation

Observe above that the R^2 equals 0.7509, or 75%. This means that 75% of the variation of price can be explained by the predictors.

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.

```
p = length(coefficients(model1)) - 1 # exclude intercept
n = nrow(insurance)

SST = var( insurance$price ) * (n - 1)
SSE = sum( model1$resid^2 )
SSR = SST - SSE

MSS_Reg = SSR / p
MSS_E = SSE / (n - p - 1)

F_stat = MSS_Reg / MSS_E
pv = pf(F_stat, p, (n - p - 1), lower.tail = FALSE)

print(paste("F statistic is: ", F_stat))
```

```
## [1] "F statistic is:  500.810741628387"
```

```
print(paste("P_value is approximately: ", pv))
```

```
## [1] "P_value is approximately:  0"
```

```
# get f-stat from the model1 summary

#print("F-statistic is:")
#m1 = summary(model1)
#m1$fstatistic[1]
#print("P_value is: ")
#pf(m1$fstatistic[1],m1$fstatistic[2],m1$fstatistic[3],lower.tail=FALSE)
```

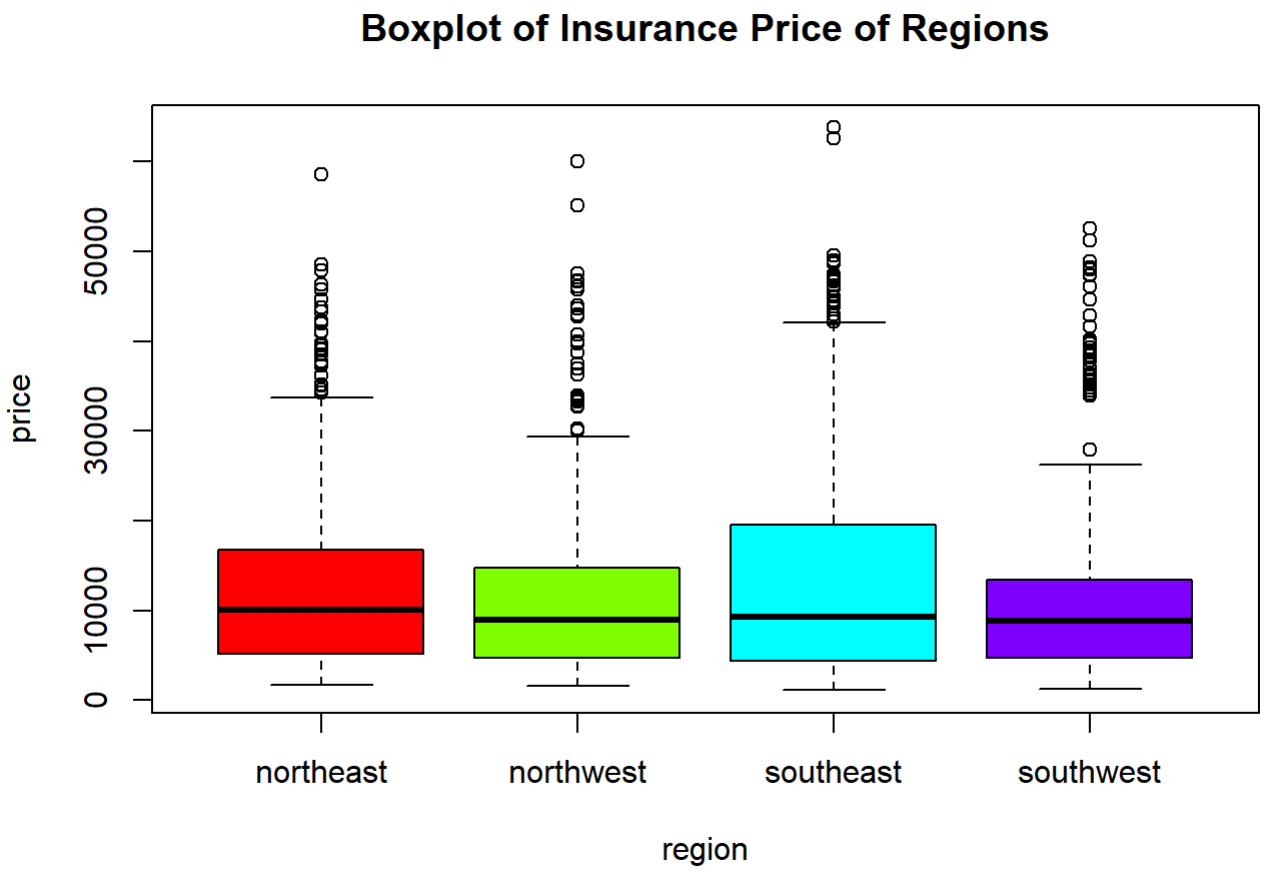
Answer

The null hypothesis is that the regression coefficients, excluding the intercept, equal zero, i.e. H_0 : β_1 = β_2 = β_3 = … = β_p = 0. The alternative hypothesis, H_a : is that at least one of the regression coefficients is not equal to zero, meaning that at least one of the predictors included in the model has predictive power, i.e. $\beta_i \neq 0$, for at least one β_i, i = 1,…,p. We can see in *model1* output (and in the manually calculated output above) that the F-statistic is 500.8 and the p-value is 2.2e-16 which is approximately 0. This means that at an $\alpha = 0.05$, we can reject the null hypothesis since the p-value is smaller than 0.05 and conclude that at least one predictor included in the model is not equal to 0 and has a predictive power.

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.

```
#region = lm(price ~ region, data=insurance)

# Let's see Box-Plot
boxplot(price ~ region, data = insurance, main="Boxplot of Insurance Price of Regions", col= rai
nbow(4))
```




```
anova = aov(price ~ region, data = insurance)
summary(anova)
```

```
##               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

The ANOVA's null hypothesis, H_0 :, is that the mean price of insurance of all four regions are equal, meaning $\mu_{northeat}$ = $\mu_{northwest}$ = $\mu_{southeast}$ = $\mu_{southwest}$ and the alternative hypothesis, H_a is that at least one of the price means differs from all of the others. Observe above in the anova output that the P-value is 0.0309. Since 0.0309 is

smaller than 0.05, we can reject the null hypothesis at $\alpha - level$ = 0.05 and conclude that at least one region's insurance price mean is different than the others. Note that in order to know wich region's mean is different, we would need to perform the TukeyHSD test.

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?

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

Answer

Here, the null hypothesis is that the region coefficients are all 0 i.e. $\mu_{northeat}$ = $\mu_{northwest}$ = $\mu_{southeast}$ = $\mu_{southwest}$ = 0 and the alternative hypothesis is that at least one region coefficient is not 0, at $\alpha - level$ of 0.05. In other words, if we reject the null hypothesis, it means we can conclude that at least one region coefficient has predictive power (note that even if only one region coefficient has predictive power, we would still include all regions in the model). Observe above the F-statistic is 2.1173 and p-value is 0.09622. Because the p-value is greater than 0.05, we ***cannot*** reject the null hypothesis that the region coefficients are all 0, with the other variables been taken under consideration. The conclusion of this test is that at $\alpha - level$ = 0.05, region does not contribute significant information to the insurance price, given age, sex, bmi, children and smoker variables.

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

Answer

The answer to question *a* suggests that the mean price of the four regions are not all equal. Since the means are not equal, one could expect that the region variable would have predictive power. However, in *b* we could not reject the null hypothesis that the four region coefficients are all zero, which means region has no predictive power, given the other variables in the model. Note however that in question *a* we assumed marginal relationship between region and price. The marginal model captures the association of one predicting variable to the response variable marginally and without considering other variables. In question *b* however, we assumed conditional relationship which captures the association of the variable region to price conditional to the inclusion of the age, sex, bmi, children and smoker factors.

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.*

```
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 Intercept is the base case of sex = female, smoker = no, and region = northeast. In the case where the observation/ data point is female, the $\beta_{sexmale} * X_{sexmale}$ term will be zero because X_{sex} = male, and the intercept would be the base case, including X_{sex} = female, which is -11938.5 (holding all other constant). The $\beta_{sexmale}$ coefficient is the ***difference*** in the average insurance price between female and male. $\beta_{sexmale}$ equals -131.3 which means that male, on average, would pay $113.3 less than female for insurance, holding all other constant. Also see output code below.

```
test = insurance[1,-7] # sex = female
p_f = predict.lm(model1, test, interval = "predict")[1] # predict with female

test$sex = "male" # sex = male
p_m = predict.lm(model1, test, interval = "predict")[1] # predict with male

print(paste("The average difference price between male and female, holding all other constant i
s: ", p_m - p_f))
```

```
## [1] "The average difference price between male and female, holding all other constant is:  -1
31.31435939511"
```

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

When the value of *bmi* increases by 0.01, mean price of insurance will increase by 3.392 (U.S. dollars) which equals β_{bmi} * 0.01 (339.2 * 0.01), holding all other constant. Also, see output code below.

```
flux = test
flux$bmi = flux$bmi + 0.01

p_b = predict.lm(model1, flux, interval = "predict", level = 0.95)

print(paste("When the value of bmi increases by 0.01, keeping other predictors constant, mean pr
ice of insurance will increase by: ", p_b[1] - p_m[1]))
```

```
## [1] "When the value of bmi increases by 0.01, keeping other predictors constant, mean price o
f insurance will increase by:  3.39193453610642"
```

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?

```
b_age90 = confint(model1, "age", level=0.90)
print("90% confidence interval of b_age parameter is:")
```

```
## [1] "90% confidence interval of b_age parameter is:"
```

```
print(b_age90)
```

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

```
b_age95 = confint(model1, "age", level=0.95)
print("95% confidence interval of b_age parameter is:")
```

```
## [1] "95% confidence interval of b_age parameter is:"
```

```
print(b_age95)
```

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

Answer

Observe above that β_{age} 90% confidence interval is [237.2708, 276.4419] and at the 95% level is [233.5138, 280.1989]. As expected, the 95% interval is wider than the 90% interval.

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 data point, excluding response column
newdata = insurance[1,-7]

print("95% confidence interval of the average price for all insurance policies with the same cha
racteristics as the first data point:")
```

```
## [1] "95% confidence interval of the average price for all insurance policies with the same ch
aracteristics as the first data point:"
```

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

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

Answer

At the 95% confidence interval, the average estimated insurance price for individuals with the same characteristics as the first data point in the sample is $25,293.71, with a lower bound of 24,143.98 and un upper bound of 26,443.44.

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.

```
newage = newdata
newage$age = 50

predict(model1, newage, interval="prediction")
```

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

Answer

If the age of an individual with the same characteristics as the first data point is increased to 50, the predicted price of an insurance policy is 33,256.26 U.S. dollars, which is an increase of 7,962.55 U.S. dollars, with a lower bound of 21,313.29 and un upper bound of 45,199.23. Since this is a prediction, the interval is wider than the confidence interval for the same data point (i.e. with age = 50). Note however that the fitted value is the same for the confidence and prediction intervals. See below a comparison between confidence interval and prediction interval for the first data point with age = 50.

```
print("prediction interval for data point 1 and age = 50:")
```

```
## [1] "prediction interval for data point 1 and age = 50:"
```

```
predict(model1, newage, interval="prediction")
```

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

```
print("confidence interval for data point 1 and age = 50:")
```

```
## [1] "confidence interval for data point 1 and age = 50:"
```

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

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