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 an automobile consulting company to understand the factors on which the pricing of cars depends.

Data Description

The data consists of a data frame with 205 observations on the following 8 variables:

1. price: Response variable ($)
2. fueltype: Qualitative variable
3. carbody: Qualitative variable
4. carlength: Quantitative variable
5. enginesize: Quantitative variable
6. horsepower: Quantitative variable
7. peakrpm: Quantitative variable
8. highwaympg: Quantitative 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()`

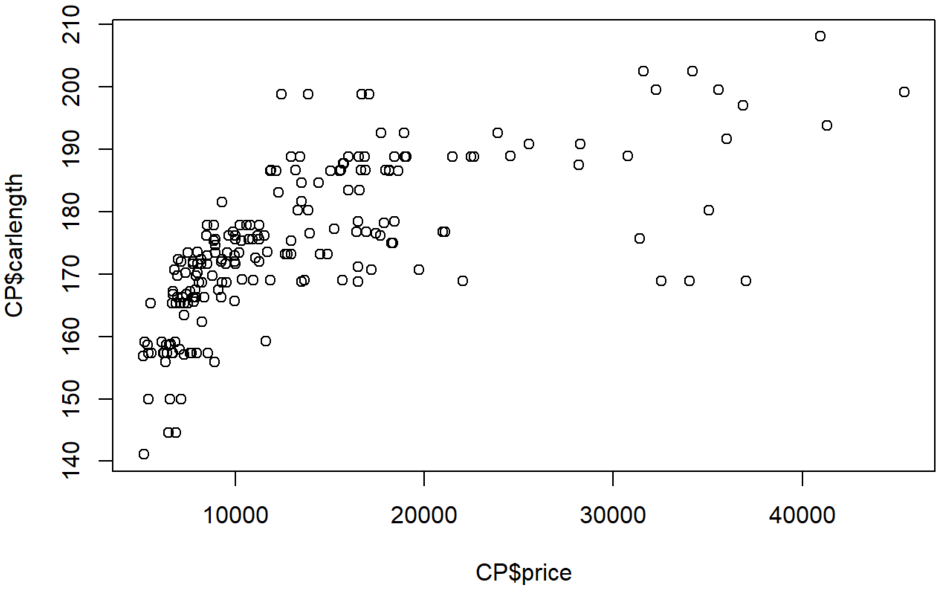
```
CP = read.csv("carprice.csv", head = TRUE)
head(CP)
```

```
##   fueltype     carbody carlength enginesize horsepower peakrpm highwaympg price
## 1      gas convertible     168.8        130        111    5000         27 13495
## 2      gas convertible     168.8        130        111    5000         27 16500
## 3      gas   hatchback     171.2        152        154    5000         26 16500
## 4      gas       sedan     176.6        109        102    5500         30 13950
## 5      gas       sedan     176.6        136        115    5500         22 17450
## 6      gas       sedan     177.3        136        110    5500         25 15250
```

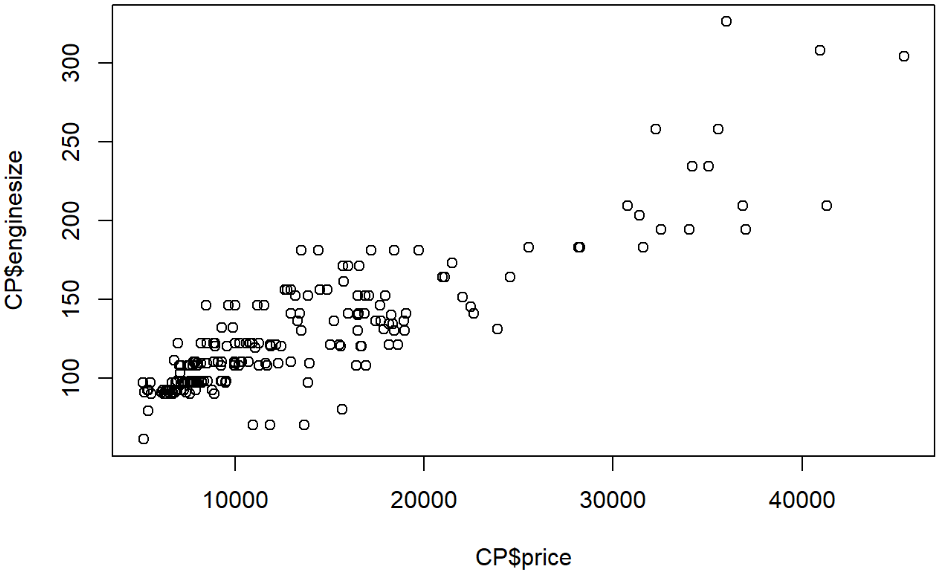
Question 1: Exploratory Data Analysis [12 points]

a. **3 pts** Create plots of the response, *price*, against three quantitative predictors (for simplicity) *carlength*, *enginesize*, and *horsepower*. Describe the general trend (direction and form) of each plot.

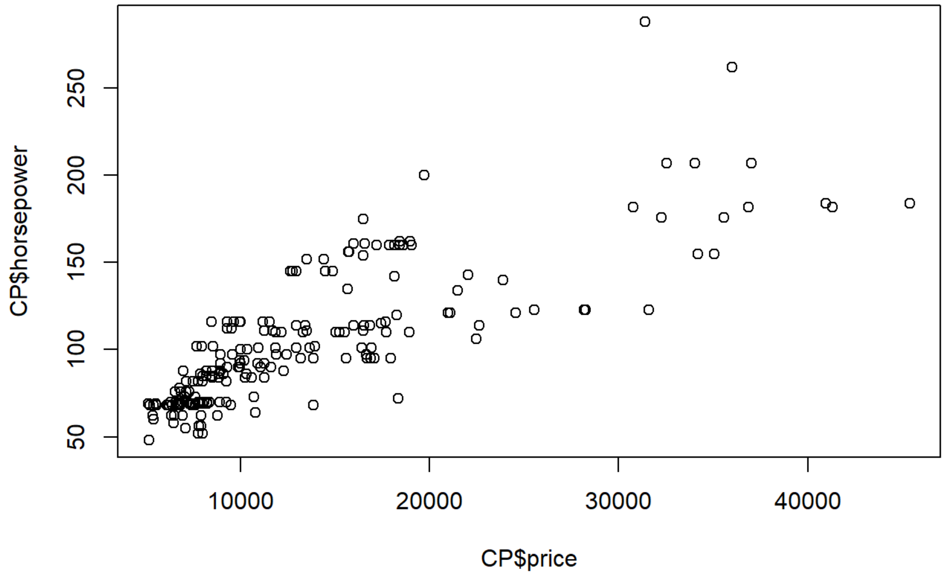
```
plot(CP$price,CP$carlength)
```



```
plot(CP$price,CP$enginesize)
```



```
plot(CP$price,CP$horsepower)
```



All of the above graphs show positive linear relationship between response variable and predicting variables.

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(CP$price,CP$carlength)
```

```
## [1] 0.68292
```

```
cor(CP$price,CP$enginesize)
```

```
## [1] 0.8741448
```

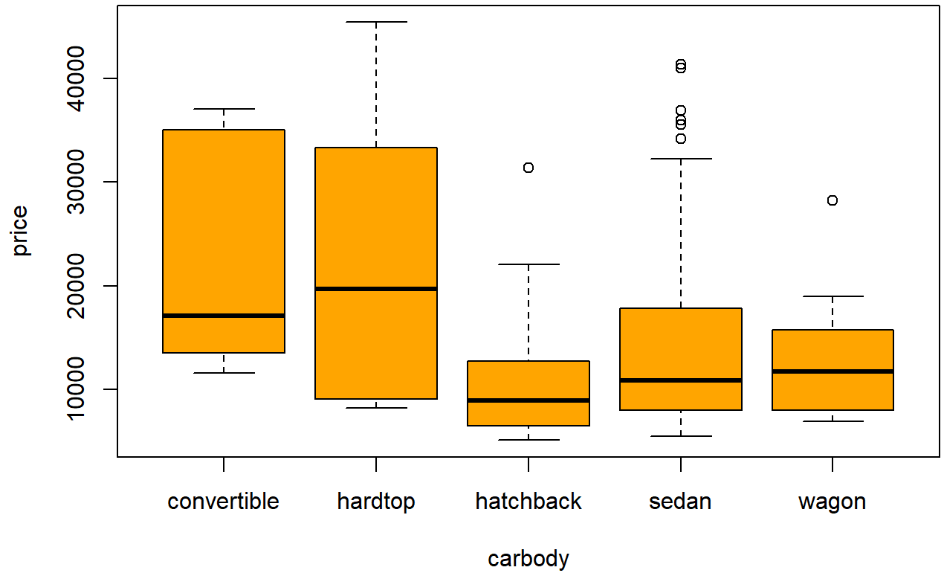
```
cor(CP$price,CP$horsepower)
```

```
## [1] 0.8081388
```

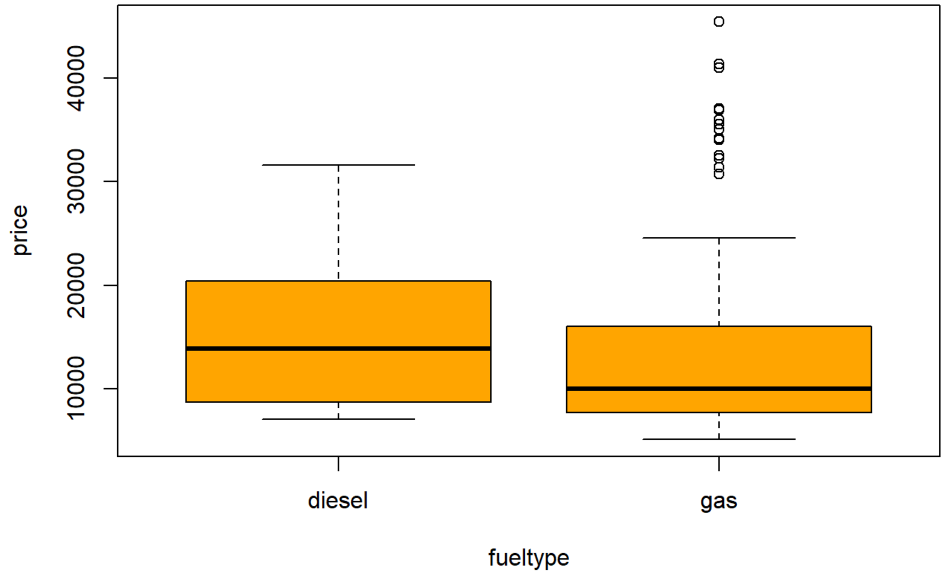
From the above result it seems that there is very strong relationship between pair of "enginesize" and price and a pair of "horsepower" and price. While there is moderate to strong relationship between "carlength" and price.

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

```
boxplot(price~carbody, col= "orange", data = CP)
```



```
boxplot(price~fueltype, col= "orange", data = CP)
```



From the above boxplot of price vs carbody we can say that convertible and hardtop car prices have very wide range compare to other three car types. And there seems to be some outliers for hatchback, sedan and wagon.
From the price vs fueltype boxbot we can say that there many outliers for gas type car.

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

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 7 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.

```
CP$fueltype <- as.factor(CP$fueltype)
CP$carbody <- as.factor(CP$carbody)
model_CP <- lm(price~.,data= CP)
summary(model_CP)
```

```
##
## Call:
## lm(formula = price ~ ., data = CP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9012.6 -1848.1   -48.1  1658.0 13011.4
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.235e+04  9.325e+03  -2.397 0.017502 *
## fueltypegas      -3.810e+03  9.596e+02  -3.970 0.000101 ***
## carbodyhardtop   -2.904e+03  1.790e+03  -1.622 0.106401
## carbodyhatchback -5.128e+03  1.436e+03  -3.571 0.000449 ***
## carbodysedan     -4.305e+03  1.477e+03  -2.914 0.003985 **
## carbodywagon     -5.504e+03  1.618e+03  -3.402 0.000811 ***
## carlength         9.564e+01  3.915e+01   2.443 0.015471 *
## enginesize        1.032e+02  1.277e+01   8.082 6.69e-14 ***
## horsepower        4.703e+01  1.383e+01   3.400 0.000818 ***
## peakrpm           2.126e+00  6.605e-01   3.218 0.001512 **
## highwaympg       -6.235e+01  6.868e+01  -0.908 0.365114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3272 on 194 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.8323
## F-statistic: 102.2 on 10 and 194 DF,  p-value: < 2.2e-16
```

Here the R squared(coefficient of determination) is 0.8405 which indicates that the model fits our observations. That means fitted value are close to data points and they fall very close to the regression line.However we also need to check the Residuls vs Fitted plot to check for bias.

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.

From the above result we can see that p-value is 2.2e-16 approximately zero, which is less than $\alpha = 0.05$. That means that at least one of the variable have predictive power.
Null hypothesis H_0: All coefficents except Intercept are equal to zero.fueltypegas=carbodyhardtop=carbodyhatchback=carbodysedan=carbodywagon=carlength=enginesize=horsepower=peakrpm=highwaymp

alternative hypothesis H_a:
At least one of the regression coefficient is not equal to zero.
F- statistic: 102.2
p-value: 2.2e-16
conclusion: From the above model We can reject the null hypothesis and say that at least one of the predicting variable have predicting power and also model has high r-squared and adjusted r-squared value which means most the variance is explained by our model and its a good fit.

Question 3: Model Comparison [12 points]

a. **4 pts** Assuming a marginal relationship between the car's body type and its price, perform an ANOVA F-test on the means of the car's body types. Using an $\alpha - level$ of 0.05, can we reject the null hypothesis that the means of the car body types are equal? Please interpret.

```
anova_model <- aov(price~carbody, data = CP)
summary(anova_model)
```

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)    
## carbody       4 1.802e+09 450499206   8.032 5.03e-06 ***
## Residuals   200 1.122e+10  56088213                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, we can reject the null hypothesis that the means of the car body types are equal as pvalue 5.03e-06 is smaller than $\alpha - level$ of 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 *carbody* 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 *carbody* are zero at $\alpha - level$ of 0.05?

```
model2 <- lm(price~fueltype+carlength+enginesize+horsepower+peakrpm+highwaympg, data =CP )

anova(model2,model_CP)
```

```
## Analysis of Variance Table
## 
## Model 1: price ~ fueltype + carlength + enginesize + horsepower + peakrpm + 
##     highwaympg
## Model 2: price ~ fueltype + carbody + carlength + enginesize + horsepower + 
##     peakrpm + highwaympg
##   Res.Df        RSS Df Sum of Sq      F   Pr(>F)   
## 1    198 2254940295                                
## 2    194 2076673688  4 178266607 4.1634 0.002941 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above result we can say that pvalue 0.002941 is less than $\alpha - level$ of 0.05 and we can reject the null hypothesis that the regression coefficients for "carbody" 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?

No, the question a and b give different results. Question a. compares the means of the different car body types and shows that they are different. They all affect the price of the car while question b looks at the coefficient of carbody variable and it is not zero that means it is significantly associated with the car price.

Question 4: Coefficient Interpretation [6 points]

a. **3 pts** Interpret the estimated coefficient of *fueltypegas* in the context of the problem. *Mention any assumption you make about other predictors clearly when stating the interpretation.*

When the fuel type is gas the price of the car reduces by $3810 holding all other predictors fixed.

b. **3 pts** If the value of the *enginesize* in the above model is increased by 0.01 keeping other predictors constant, what change in the response would be expected?

The price of the car will increase by (103.2)*(0.01) = 1.032

Question 5: Confidence and Prediction Intervals [10 points]

a. **5 pts** Compute 90% and 95% confidence intervals (CIs) for the parameter associated with *carlength* for the model in Question 2. What observations can you make about the width of these intervals?

```
confint(model_CP,"carlength", level = .90)
```

```
##                 5 %    95 %
## carlength 30.93222 160.356
```

```
confint(model_CP,"carlength", level = .95)
```

```
##                2.5 %   97.5 %
## carlength 18.42162 172.8666
```

From the above intervals we can say that 95 times out of 100 the data will be within (18.42162 - 172.8666) this range when the confidence interval is 95% and 90 out of 100 times data will be within 30.93222 - 160.356 this range when confidence interval is 90%. The bigger the confidence interval the larger the range.

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

```
newdatacp <- CP[1,-8]
predict(model_CP, newdatacp, interval = "confidence",level =.95)
```

```
##        fit      lwr      upr
## 1 17565.19 14885.37 20245.02
```

The average predicted price of the car is 17565.19 with the lower bound of 14885.37 and an upper bound of 20245.02 for 95% confidence interval.

c. **2.5 pts** Suppose that the *carlength* value for the first data point is increased to 200, while all other values are kept fixed. Using *model1*, predict the price of a car with these characteristics. What is the 95% prediction interval for this prediction? Provide an interpretation of your results.

```
newcpobs = newdatacp
newcpobs[3] = 200
predict(model_CP, newcpobs, interval="prediction")
```

```
##        fit      lwr      upr
## 1 20549.29 13133.51 27965.07
```

From the above model we can say that if the car length increases to 200 from 168.8 the predicted price of the car increase by 2984.1(from 17565.19 to 20549.29)