DATA303-A1

Michael Fry 300570669

2023-03-17

Question 1

a)

```
data <- read.csv('cardekho.csv')
summary(data)
```

```
##      price             make                kms              fuel
##  Min.   :   30.0   Length:8028        Min.   :   1.00   Length:8028
##  1st Qu.:  260.0   Class :character   1st Qu.:  35.00   Class :character
##  Median :  450.0   Mode  :character   Median :  60.00   Mode  :character
##  Mean   :  640.4                      Mean   :  69.77
##  3rd Qu.:  680.0                      3rd Qu.:  98.00
##  Max.   :10000.0                      Max.   :2360.46
##
##     seller              tx               owner              mileage
##  Length:8028        Length:8028        Length:8028        Min.   : 0.00
##  Class :character   Class :character   Class :character   1st Qu.:16.78
##  Mode  :character   Mode  :character   Mode  :character   Median :19.30
##                                                           Mean   :19.39
##                                                           3rd Qu.:22.32
##                                                           Max.   :42.00
##                                                           NA's   :214
##      esize          power
##  Min.   : 624   Min.   :  0.00
##  1st Qu.:1197   1st Qu.: 68.85
##  Median :1248   Median : 82.40
##  Mean   :1463   Mean   : 91.82
##  3rd Qu.:1582   3rd Qu.:102.00
##  Max.   :3604   Max.   :400.00
##  NA's   :214    NA's   :208
```

i) Incorrect Values

The following variables have potentially incorrect values.

price:

Maximum price of 10000.0 is significantly larger than I would expect from looking IQR, mean and median. It is possible that this value is real, eg from an ultra luxury car, however this would need to be further investigated.

kms:

It is unlikely for a car to have traveled 2,360,460km's as the data would suggest. However this is not impossible,

and would need to be investigated further. This value is also significantly larger than IQR, mean and median would suggest is logical.

mileage:

A value of 0 kilometers per liter is an obvious mistake.

power:

Unless a car has been sold without an engine, or not running, it is impossible for a car to have 0bhp.

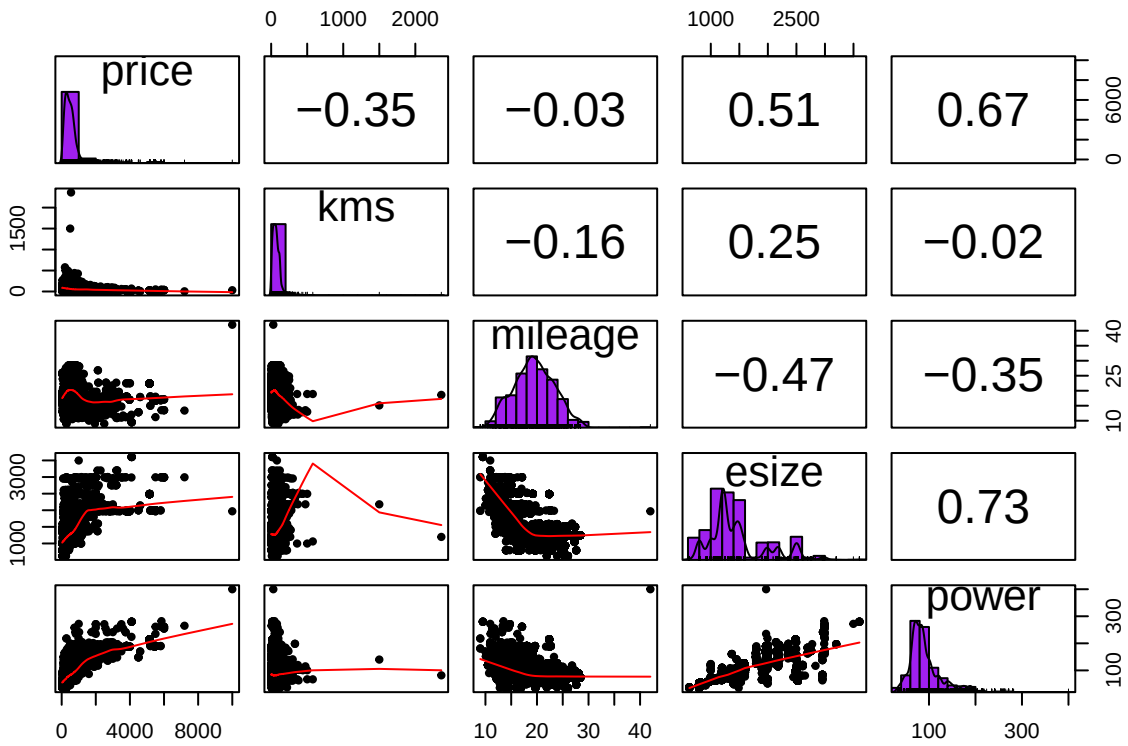
ii) Missing Values

The variables mileage, esize and power all have a number of missing values. 214, 214 and 208 respectively shown by the NA values in each column of the summary.

b)

```
data_cleaned <- read.csv('cardekho2.csv')

data_cleaned%>%
  select(where(is.numeric))%>%  #select numerical variables (includes integers)
  pairs.panels(method = "spearman", # correlation method
               hist.col = 'purple', # histogram color
               density = TRUE,  # show density plots
               ellipses = FALSE # do not show correlation ellipses
               )
```




```
## Each predictor was looked at in more detail to assertain whether
# there was a linear relationship with price. Unnecessary for the
# final report. So these graphs have been left out.
#plot(price ~ kms, data=data_cleaned)
#plot(price ~ mileage, data=data_cleaned)
```

```
#plot(price ~ esize, data=data_cleaned)
#plot(price ~ power, data=data_cleaned)
```

i)

The two predictors that stand out to me as not having a linear relationship with price are mileage and kms. Both of these variables would not be good candidates for fitting a linear model. This can be tested by using pearsons correlation coefficient which measures the strength of the linear relationship between two variables.

```
# Pearsons correlation coefficient for Mileage vs Price
mil_pear <- cor.test(data_cleaned$mileage, data_cleaned$price, method = "pearson")

# Pearsons correlation coefficient for kms vs Price
kms_pear <- cor.test(data_cleaned$kms, data_cleaned$price, method = "pearson")

pander(data.frame(mil_pear$estimate,kms_pear$estimate))
```

	mil_pear.estimate	kms_pear.estimate
cor	-0.1267	-0.2222

The output of the pearsons correlation above confirms my qualitative assessment of non linearity by looking at the graph. Numbers close to zero, like the ones above suggest that there is not, or at least a very weak linear relationship between each of the variables and price.

Both esize and power seem to have some sort of positive, somewhat linear relationship with price however I would suspect they would not meet the regression assumptions of equal variance and normal errors. Looking at the graphs closely, both seem to have funneling occurring which could suggest that a log transformation may be needed to align the variables with the linear regression assumptions.

ii)

Price does not seem to follow a normal distribution, with a large left skew of data towards cheaper cars. Transformation of the predictor variable may lead to better results.

Further analysis of regression assumptions, and tests should be performed before making this decision.

c)

```
# Fit a LM for all predictors.
model1 <- lm(price ~ kms + mileage + esize + power + make + fuel
             + seller + tx + owner, data=data_cleaned)

# Create Summary of LM model
summary1 <- summary(model1)

# Print summary
summary1
```

```
##
## Call:
## lm(formula = price ~ kms + mileage + esize + power + make + fuel +
##     seller + tx + owner, data = data_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2620.0  -188.9     3.0   171.2  3970.6
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -745.84341   77.73880  -9.594  < 2e-16 ***
## kms                      -1.58575    0.10333 -15.346  < 2e-16 ***
## mileage                  30.89608    2.08048  14.850  < 2e-16 ***
## esize                     0.09409    0.02463   3.820 0.000134 ***
## power                    13.82643    0.26011  53.156  < 2e-16 ***
## makeHonda              -150.95095   31.91104  -4.730 2.28e-06 ***
## makeHyundai              18.50214   26.50858   0.698 0.485218
## makeMahindra            108.30980   30.47386   3.554 0.000381 ***
## makeMaruti              149.07261   25.60333   5.822 6.03e-09 ***
## makeOther               262.86123   26.61647   9.876  < 2e-16 ***
## makeTata                -50.08944   28.61375  -1.751 0.080065 .
## makeToyota              235.48398   34.48390   6.829 9.21e-12 ***
## fuelPetrol               -5.70032   15.65198  -0.364 0.715725
## sellerIndividual       -235.50855   16.38584 -14.373  < 2e-16 ***
## sellerTrustmark Dealer -284.31597   34.71053  -8.191 3.00e-16 ***
## txManual               -414.75405   19.70954 -21.043  < 2e-16 ***
## ownerSecond            -112.88321   12.72001  -8.874  < 2e-16 ***
## ownerThird or above    -129.54557   19.78783  -6.547 6.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 451.2 on 7779 degrees of freedom
## Multiple R-squared:  0.6901, Adjusted R-squared:  0.6894
## F-statistic:  1019 on 17 and 7779 DF,  p-value: < 2.2e-16
```

```
# Ascertain estimate of sigma^2 from the output above.
model1_error_var <- (summary1$sigma)^2
```

Estimate of σ^2:

```
model1_error_var
```

```
## [1] 203606.8
```

d)

The predictor values for a car where $\widehat{E[Y|X]} = \hat{\beta}0$ = -745843.41 are all 0.

The value -745843.41 INR corresponds to the intercept term of -745.84341(thousand rupees). This is the value of Y (Price) when all predictors are 0. For a mix of categorical and numerical variables as seen in the example, this intercept value represents the price when all numerical values are 0, and all categorical variables are equal to their baseline value, aka the one not shown in the table above. IE: Make = Ford, Fuel = Diesel, Seller = Dealer, tx = Automatic and owner = First.

Even though this intercept is meant to help with interpretation of the co-efficients, in this example, it is not logical to predict such a car, as the values 0 are not in the range of values for the predictor variables and thus it does not make sense to use them in an interpretation.

e)

Interpretation of following co-efficient from part c):

i) txManual

The difference in price Y given all other predictors are kept constant of a car being manual vs the baseline of a car being Automatic. Practical interpretation says that a car being manual results in a decrease in price of -414.75 Thousand Rupees compared to the same car if it were automatic.

Real interpretation: On average, automatic cars cost more, Ceteris Paribus. (All other things held constant)

ii) mileage

A unit increase in mileage, results in a 30.896 Thousand Rupee increase in price, Ceteris Paribus.

f)

95% Confidence and prediction intervals for the last three observations in the data set.

```
#Get predictor values to predict for as a dataframe.
xdata<-subset(tail(data_cleaned,3), select = -price)
## Select last three rows and excludes price
rownames(xdata) <- NULL
```

95% Confidence Interval

```
##Confidence intervals for estimating the mean response
pander(predict(model1, newdata=xdata, interval="confidence"),
caption="Confidence intervals", round=2)
```

Table 2: Confidence intervals

fit	lwr	upr
2294	2256	2332
2623	2581	2665
2034	1987	2081

95% Prediction Interval

```
##Prediction intervals for predicting the response for new predictor values
pander(predict(model1, newdata=xdata, interval="prediction"),
round=2,caption="Prediction intervals")
```

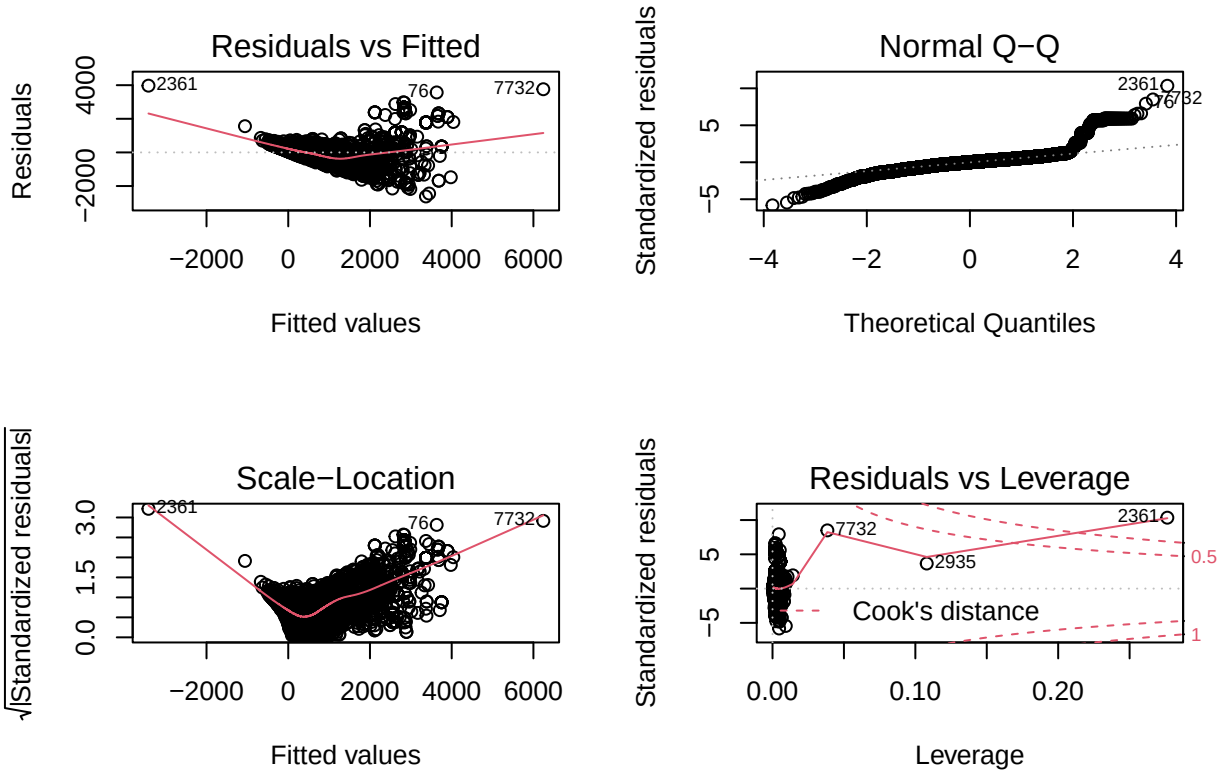
Table 3: Prediction intervals

fit	lwr	upr
2294	1408	3179
2623	1737	3508
2034	1148	2920

The prediction intervals are always wider than the confidence intervals as this is reflecting the greater uncertainty about the individual price of a given car compared to uncertainty about the average price over many cars.

g)

```
par(mfrow=c(2,2)) ##Places the next four graphs in a single window in a 2 x 2 configuration
plot(model1) ## Obtain diagnostic plots
```




```
# Closer inspection of plots, not needed in report.
#plot(model1) ## Obtain diagnostic plots
```

Residuals vs Fitted

Residuals are not evenly distributed to a horizontal line, clear evidence of funneling, Very uneven variation. Non linearity is seen, possible logarithmic bias. Outliers seen at very low, and very high fitted values.

QQ Plot

Substantial variation from QQ line from -4 to -2 and from 2-4. Assumption of normality does not seem to be met.

Scale Location

Residuals are not evenly spread along the range of fitted values, clear funneling is seen. Smooth line is parabolic in shape, clearly not linear. Clear evidence of heteroscedasticity / departure from the assumption of equal variance.

Leverage / Cooks Distance

One highly influential observation, number 2361. This point is clearly outside the threshold of 1 for an influential point. Further investigation should be conducted to see if this datapoint should be removed from the analysis.

Overall

Observations 2361 and 7732 are present as outliers in all diagnostic plots. This suggests that further investigation should be conducted on these observations, and they should be considered for exclusion.

h)

Test of Normality

The KS test will be used to test for normality as we have more than 50 samples in the dataset.

H_0 : The sample comes from a normal distribution.

H_1 The sample does not come from a normal distribution.

```
ks.test(model1$res, "pnorm") ##K-S test
```

```
## Warning in ks.test(model1$res, "pnorm"): ties should not be present for the
## Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  model1$res
## D = 0.49948, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

A p value of <2.2e-16 is significantly under the 5% significance level ($Pvalue < 0.05$) and therefor I have sufficient evidence to reject the null hypothesis in favor of the alternative and conclude that the sample does not come from a normal distribution.

Test of heteroscedasticity

The Breush- Pagan test will be used to determine whether or not heteroscedasticity is present in a regression model.

H_0 Homoscedasticity is present (the residuals are distributed with equal variance)

H_1 Heteroscedasticity is present (the residuals are not distributed with equal variance)

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(model1) ##Breusch-Pagan test
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 2600.1, df = 17, p-value < 2.2e-16
```

A p value of <2.2e-16 is significantly under the 5% significance level ($Pvalue < 0.05$) and therefor I have sufficient evidence to reject the null hypothesis in favor of the alternative and conclude that heteroscedasticity is present in the sample.

The results of non normality and non equal variance aligns with the subjective analysis of the scatter plots in part g). In part g I noted that the response variable price did not appear normally distributed, and that there was a large amount of variance around some of the predictors.

The results of the above tests confirm my suspicion that transformations of both the response and predictor variables will be needed to meet the regression assumptions.

i)

```
pander(vif(model1), digits=2, caption="VIF values")
```

Table 4: VIF values

	GVIF	Df	GVIF^(1/(2*Df))
kms	1.3	1	1.2
mileage	2.5	1	1.6
esize	5.9	1	2.4
power	3.3	1	1.8
make	3.5	7	1.1
fuel	2.3	1	1.5
seller	1.5	2	1.1
tx	1.7	1	1.3
owner	1.2	2	1.1

As the model includes categorical predictors, GVIF^(1/(2*Df)) values are used for the interpretation.

There is no evidence of severe multicollinearity, all GVIF^(1/(2*Df)), and GVIF values significantly less than 10.

j)

H_0 $\beta 0, \beta 1 ... \beta p = 0$

H_1 At least one $\beta \neq 0$

The global usefulness test is given in the output of the model summary in terms of the F statistic and associated P value.

From the summary output in 1c) the P value for the associated F test is $<$ 2.2e-16 which is significantly less than the 5% significance level ($Pvalue < 0.05$) and therefor I have sufficient evidence to reject the null hypothesis in favor of the alternative and conclude that at least one $\beta \neq 0$

This test leads to the conclusion that the model is useful, and worth going on to further analyse and interpret a model of price against each of the predictors.

Question 2

```
olive<-read.csv("olive.csv", header=T)
#str(olive)

fit2<-lm(palmitic ~  linoleic + stearic, data=olive)
#summary(fit2)

fit3<-lm(palmitic ~  linoleic + stearic + oleic, data=olive)
#summary(fit3)
```

a)

The change in sign of the coefficient for linoleic acid between fit2 and fit3 models could be due to the presence of multicollinearity. From the scatter plot matrix, there is a strong negative relationship between linoleic and oleic acids, showing they have a strong correlation between each other. The addition of oleic acid in model

3 may have caused the above correlation to be picked up by the model, causing the sign of linoleic acid to change.

b)

```
# create data frame with new values
new_data <- data.frame(linoleic=0.3, stearic=2.2, oleic=73.0)

# get predictions and Confidence intervals
pander(predict(fit3, new_data, interval="prediction", level=0.95),
       round=2, caption = 'Prediction Interval')
```

Table 5: Prediction Interval

fit	lwr	upr
18.88	18.17	19.59

```
pander(predict(fit3, new_data, interval="confidence", level=0.95),
       round=2, caption = 'Confidence Interval')
```

Table 6: Confidence Interval

fit	lwr	upr
18.88	18.66	19.1

c)

If all other assumptions hold, for the prediction in part b to be valid, it would have to represent a likely result given the distributions of the response variables. IE: The values used for the predictors must be representative of the population from which the original sample was drawn.

d)

The regression assumption violated by many samples coming from the same region is the assumption of independence. When observations are not independent of one another, the model's assumptions regarding the errors' independence from one another and their constant variance may also be lost. As a result, the study may draw the wrong results due to skewed regression coefficients, unreliable standard errors, and inaccurate p-values.