DATA303-A2

Michael Fry 300570669

2023-03-31

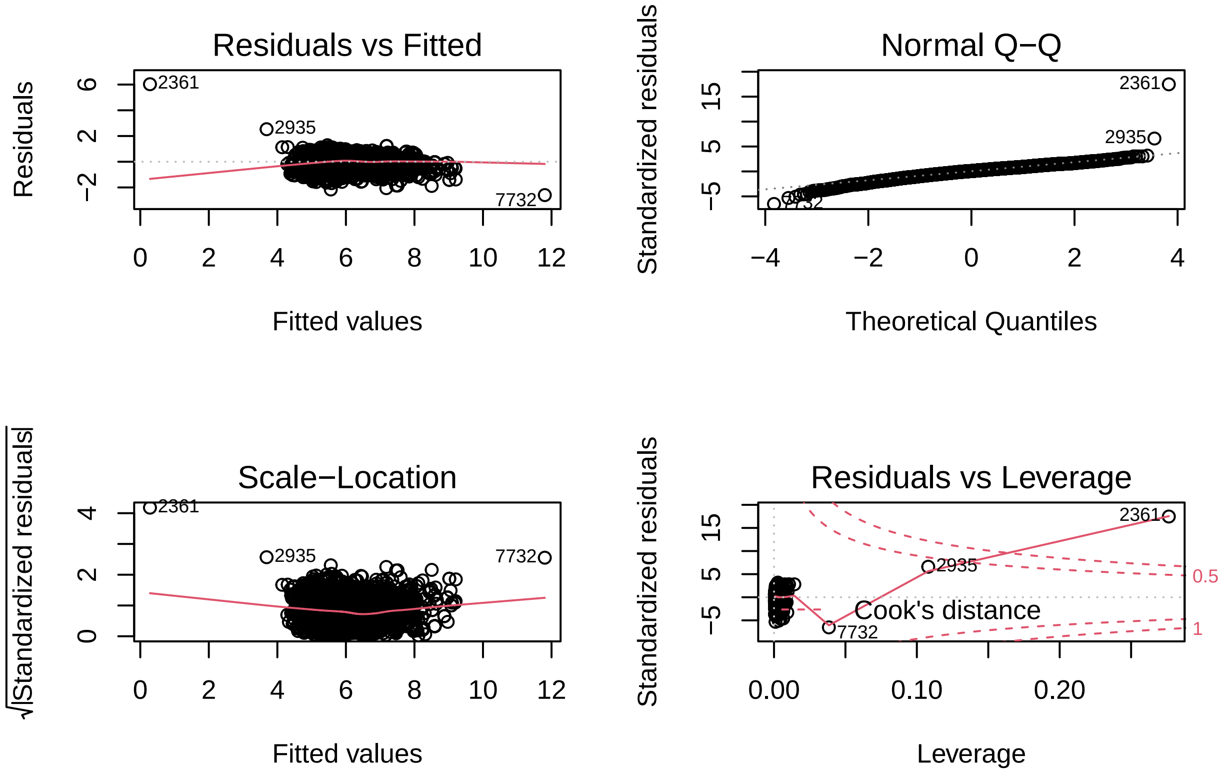
Question 1

a.

```
library(pander)
library(zoo)
```

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

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

```
car2 <- read.csv('cardekho2.csv')

model1 <- lm(log(price) ~ make + kms + fuel + seller + tx + owner +
               mileage + esize + power , data = car2)

# Not needed in report, used to look at model.
# pander(summary(model1))

par(mfrow=c(2,2))
plot(model1)
```



Residuals vs Fitted

Residuals are relatively evenly distributed to a horizontal line when the fitted values are above 4. Below 5 there is deviation from the horizontal line. There are very few fitted values below 5. Outliers seen at low, and high fitted values. With the removal of some outliers, assumption of equal variance seems to be met within the majority range of the fitted values. It would be interesting to see how horizontal the line is once these outliers were removed. (Specifically observation 7732, 2361 and 2935)

QQ Plot

Some variation from QQ line from -4 to -2. Large deviation from QQ at high X values, however only from two points, 2361 and 2935. These observations were also seen in the residuals graph. Investigation into these values should be conducted. Assumption of normality does seem to mostly met when looking at the graph without the outliers mentioned above apart from the deviation seen at the bottom left of the line.

Scale Location

Residuals are evenly spread along the range of fitted values from 4 to 9, no funneling is seen. Smooth line is slightly curved in shape, not linear but possibly skewed by the same outliers mentioned above as most of the variation seems well contained. some evidence of heteroscedasticity / departure from the assumption of equal variance however removal of outliers may change this.

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 data point should be removed from the analysis. Observations 2935 and 7732 also seen as outliers to the rest of the group, but not over the threshold of 1.

Overall

Observations 2361, 2935 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. These points, especially observation 2361 may be skewing the interpretation of the models, and model assumptions.

To further test the diagnostics covered above, hypothesis tests will be used to confirm or deny the findings above with no change of the dataset.

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 = 0.23096, 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. This somewhat aligns the qualitative diagnostics performed above as there was some deviations from the QQ line at the bottom, and large deviation at the top caused by an influencial point. Transformations of predictor variables may help to solve this, but outliers should be investigated first.

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

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 2145.2, 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.

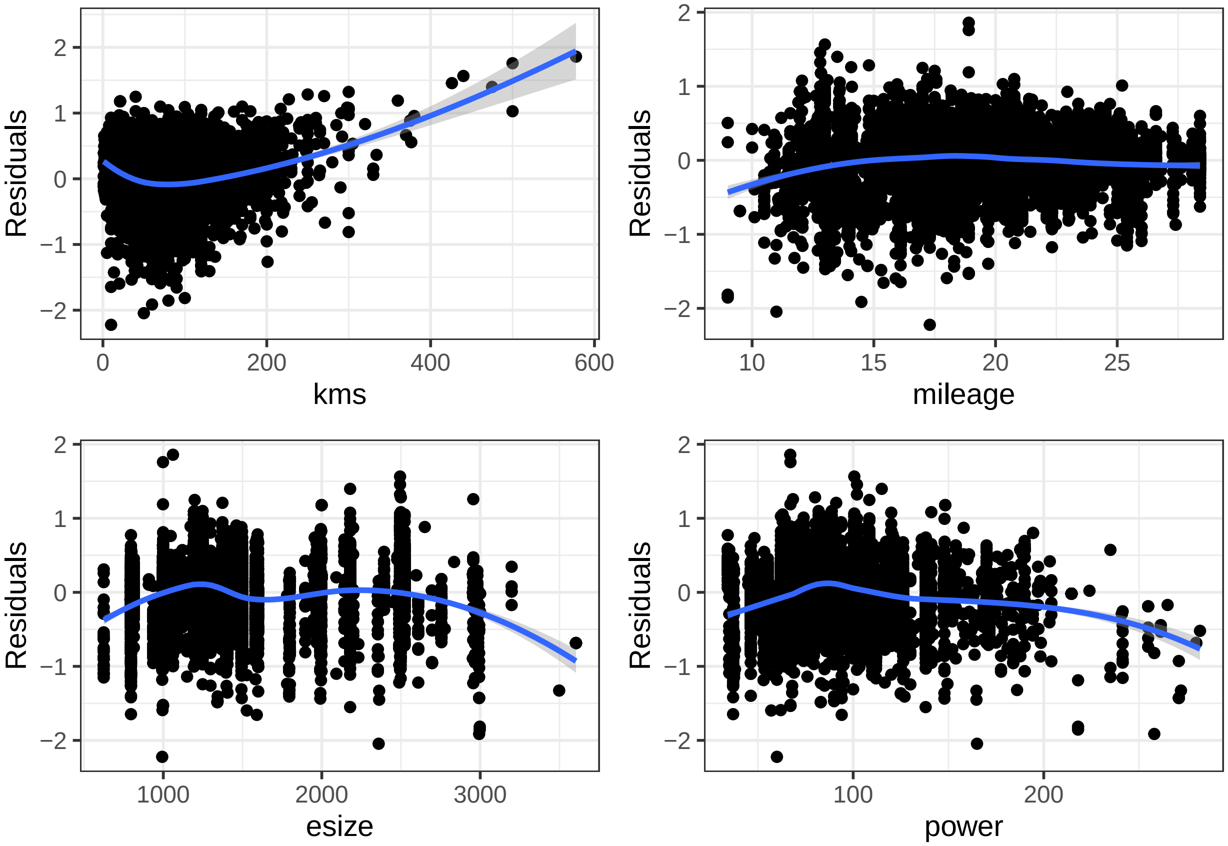
The results of non normality somewhat aligns with the subjective analysis of the scatter plots in part above. However the test concluding that heteroscedasticity was present surprised me.

Note: As mentioned above, there are some outliers that may be skewing results, to get more accurate test results these points would need to be investigated and possibly removed. I believe that the with these outliers removed, the results from the studentized Breusch-Pagan test may be different.

b.

```
car3 <- read.csv('cardekho3.csv')

model2 <- lm(log(price) ~ make + kms + fuel + seller +
               tx + owner + mileage + esize + power , data = car3)

library(broom)

car3$.resid<-model2$residuals
library(ggplot2)

a<-ggplot(model2,aes(x=kms, y=.resid))+
geom_point()+
geom_smooth(method='loess')+
labs(x="kms", y="Residuals")+
theme_bw()

b<-ggplot(model2,aes(x=mileage, y=.resid))+
geom_point()+
geom_smooth(method='loess')+
labs(x="mileage", y="Residuals")+
theme_bw()

c<-ggplot(model2,aes(x=esize, y=.resid))+
geom_point()+
geom_smooth(method='loess')+
labs(x="esize", y="Residuals")+
theme_bw()

d<-ggplot(model2,aes(x=power, y=.resid))+
geom_point()+
geom_smooth(method='loess')+
labs(x="power", y="Residuals")+
theme_bw()

library(gridExtra)
grid.arrange(a,b,c,d, nrow=2)
```

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



There are no clear patters in the residuals for mileage, esize and power however there is a clear upward curve in the residuals in kms. This suggests that the predictor kms could benefit from a transformation to adhere to the assumptions of regression.

c.

```
pander(summary(model2))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.037	0.06883	58.66	0
makeHonda	0.06815	0.02787	2.446	0.01448
makeHyundai	0.1103	0.02313	4.767	1.901e-06
makeMahindra	0.1983	0.02663	7.446	1.064e-13
makeMaruti	0.09294	0.02235	4.159	3.229e-05
makeOther	0.02856	0.02322	1.23	0.2188
makeTata	-0.3009	0.02497	-12.05	3.668e-33
makeToyota	0.4428	0.03011	14.71	2.58e-48
kms	-0.00375	0.0001148	-32.67	1.624e-219
fuelPetrol	-0.1906	0.01393	-13.68	4.153e-42
sellerIndividual	-0.07989	0.01433	-5.573	2.581e-08
sellerTrustmark Dealer	-0.01619	0.03029	-0.5345	0.593
txManual	-0.2311	0.01724	-13.41	1.509e-40
ownerSecond	-0.2852	0.01118	-25.52	4.53e-138
ownerThird or above	-0.4737	0.01742	-27.2	1.133e-155
mileage	0.05422	0.001841	29.46	6.763e-181
esize	0.0003402	2.153e-05	15.8	2.364e-55

	Estimate	Std. Error	t value	Pr(>|t|)
power	0.01261	0.0002307	54.68	0

Table 2: Fitting linear model: log(price) ~ make + kms + fuel + seller + tx + owner + mileage + esize + power

Observations	Residual Std. Error	R^2	Adjusted R^2
7794	0.3937	0.7731	0.7726

To interpret the difference in **price** for a petrol car compared to a diesel car when all other predictors are held constant we have to take into consideration that we have transformed the response variable to be log(price).

The model is now:

$Log(Y)\ \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_p X_p + e$

Due to this, it is easier to understand the interpretation of the proportional change in Y, rather than they level of Y outright.

Proportional Change of Y =

$e^{\beta_1} - 1$

Looking at the regression coefficient for model2 for fuel type **petrol** we have $e^{-0.1906} - 1 = -0.1735$

The price of a petrol car, compared to a diesel car, when all other factors are held constant is:

Price + Price * -0.1735

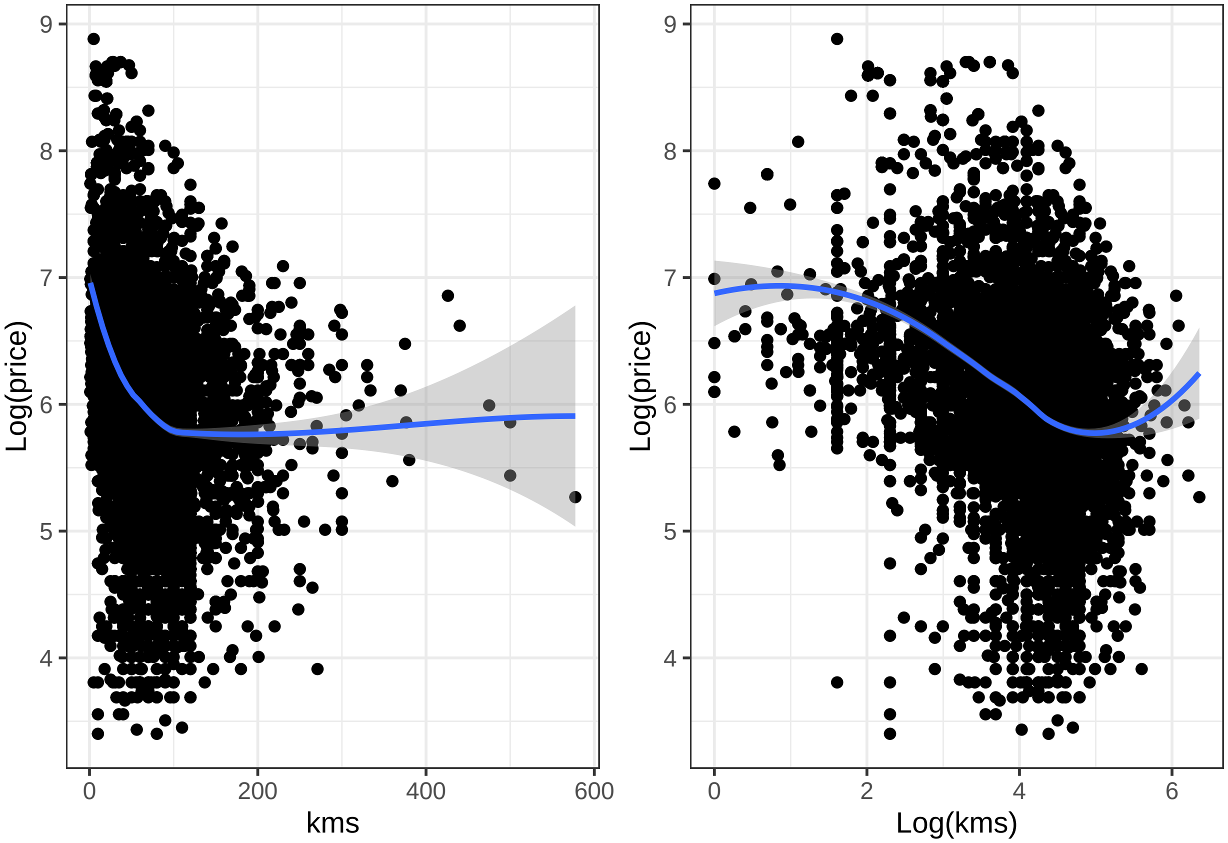
We expect price to decrease by a multiplicative factor of 0.1735 or 17.5% when a car is petrol compared to when it is diesel when all other factors are held constant.

d.

To show that a log transformation is the most appropriate transformation for kms, in the model with log(price) I will create a scatter plot of kms vs log(price) and then another scatter plot of log(kms) vs log(price). If this transformation is appropriate I would expect to see a more linear relationship, reduction of the skew and reduction of Heteroscedasticity / non equal variance.

```
library(ggplot2)

e <- ggplot(car3,aes(x=kms, y=log(price)))+
geom_point()+
geom_smooth(method='loess')+
labs(x="kms", y="Log(price)")+
theme_bw()

f <- ggplot(car3,aes(x=log(kms), y=log(price)))+
geom_point()+
geom_smooth(method='loess')+
labs(x="Log(kms)", y="Log(price)")+
theme_bw()

library(gridExtra)
grid.arrange(e,f, nrow=1)
```

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



If a log transformation was the most appropriate for kms, I would expect the second plot (log(price) against log(kms)) to show a more linear relationship between the two variables than the first plot, a higher level of Homoscedasticity. Although the log transformation of kms has improved some of the funneling seen in the first graph as well as the heavy left skew the relationship still does not seem linear. A polynomial transformation of kms / log(kms) should be tested. It appears that a log transformation of kms did not have the results we were looking for.

e.

```
#AIC
step(model2, direction="both")
```

```
## Start:  AIC=-14513.2
## log(price) ~ make + kms + fuel + seller + tx + owner + mileage +
##     esize + power
##
##             Df Sum of Sq    RSS    AIC
## <none>                   1205.2 -14513
## - seller    2      5.16 1210.3 -14484
## - tx        1     27.87 1233.1 -14337
## - fuel      1     29.00 1234.2 -14330
## - esize     1     38.68 1243.9 -14269
## - mileage   1    134.48 1339.7 -13691
## - make      7    157.42 1362.6 -13570
## - kms       1    165.44 1370.6 -13513
## - owner    2    168.63 1373.8 -13496
## - power    1    463.33 1668.5 -11980
```

```
##
## Call:
## lm(formula = log(price) ~ make + kms + fuel + seller + tx + owner +
##     mileage + esize + power, data = car3)
##
## Coefficients:
##             (Intercept)               makeHonda             makeHyundai
##               4.0374331               0.0681513               0.1102773
##            makeMahindra              makeMaruti               makeOther
##               0.1983128               0.0929383               0.0285578
##                makeTata              makeToyota                     kms
##              -0.3009262               0.4427864              -0.0037497
##              fuelPetrol        sellerIndividual  sellerTrustmark Dealer
##              -0.1905640              -0.0798936              -0.0161888
##                txManual             ownerSecond      ownerThird or above
##              -0.2311297              -0.2852456              -0.4737161
##                 mileage                   esize                   power
##               0.0542173               0.0003402               0.0126117
```

To analyse these results the Hilbe AIC rules of thumb will be conducted to ascertain which model to pick.

$AIC(A) - AIC(B) = -14484 - (-14513) = 29

As the difference is in the interval [>10], applying the Hilbe AIC rules of thumb means our preferred model is model B which has the lower AIC. This is the model that includes all predictors.

Because the preferred model is the model with all predictors, according to the stepwise AIC no parameters should be excluded from the model.

f.

To investigate the effect of mileage on log(price) depends on the value of tx I will fit a model that includes an interaction term between mileage and tx.

```
model5 <- lm(log(price) ~ make + log(kms) + fuel + seller +
               tx + owner + mileage + esize + power + mileage:tx , data = car3)

pander(summary(model5))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.441	0.08778	50.6	0
makeHonda	0.06774	0.0268	2.527	0.01151
makeHyundai	0.1149	0.02226	5.161	2.519e-07
makeMahindra	0.1899	0.02569	7.394	1.576e-13
makeMaruti	0.0974	0.0215	4.53	5.983e-06
makeOther	0.01573	0.02235	0.7041	0.4814
makeTata	-0.3088	0.02407	-12.83	2.712e-37
makeToyota	0.4261	0.02894	14.72	2.031e-48
log(kms)	-0.2569	0.006114	-42.02	0
fuelPetrol	-0.2199	0.01344	-16.36	3.371e-59
sellerIndividual	-0.0696	0.01379	-5.048	4.56e-07
sellerTrustmark Dealer	-0.03202	0.02932	-1.092	0.2747
txManual	0.127	0.06734	1.886	0.05933
ownerSecond	-0.2512	0.01084	-23.16	7.476e-115

	Estimate	Std. Error	t value	Pr(>|t|)
ownerThird or above	-0.4443	0.01678	-26.47	6.663e-148
mileage	0.06927	0.0035	19.79	4.387e-85
esize	0.0003333	2.075e-05	16.07	3.614e-57
power	0.01286	0.000223	57.68	0
txManual:mileage	-0.01574	0.00356	-4.423	9.888e-06

Table 4: Fitting linear model: log(price) ~ make + log(kms) + fuel + seller + tx + owner + mileage + esize + power + mileage:tx

Observations	Residual Std. Error	R^2	Adjusted R^2
7794	0.3788	0.79	0.7895

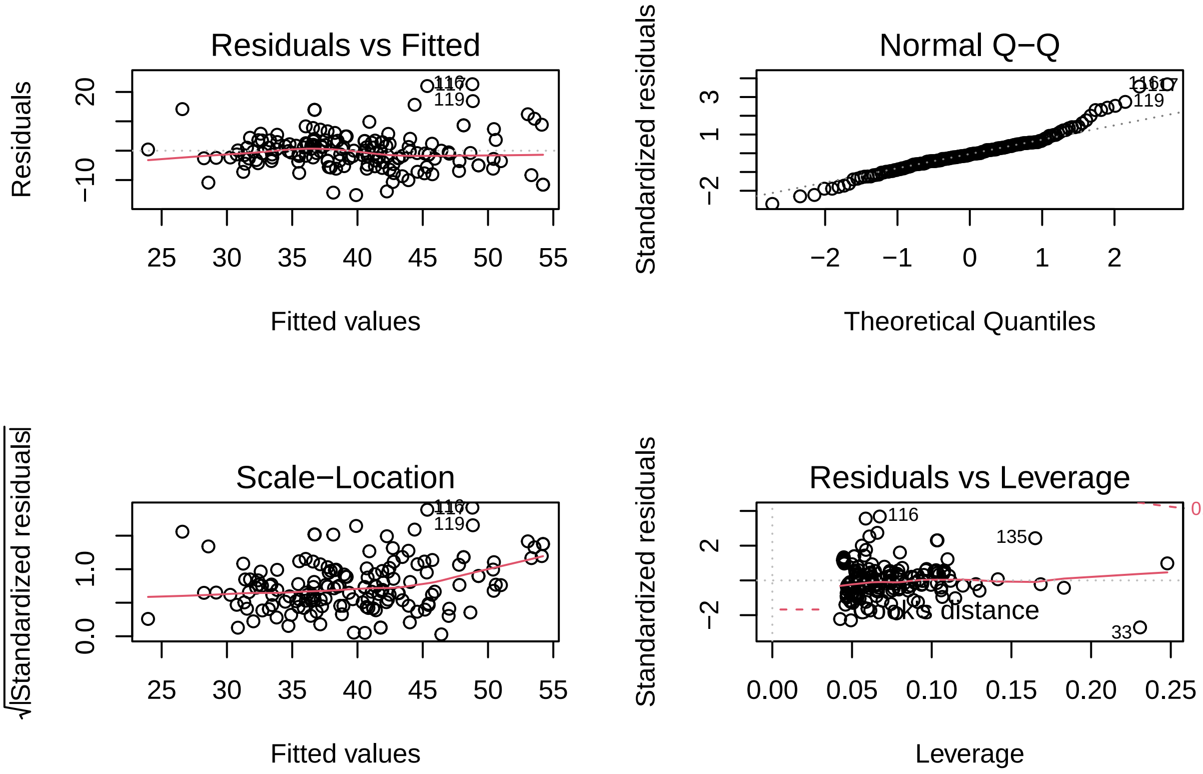
i. For a car with automatic transmission,(ie, txManual = 0) a one-unit increase in mileage is associated with a 0.06927 increase in the log(price) of the car.

ii. For a car with manual transmission (ie, txManual = 1), a one-unit increase in mileage is associated with a (0.06927 - 0.01574) = 0.05353 increase in the log(price) of the car.

Question 2

a.

```
ship <- read.csv('cruise_ship.csv')

ship_model_1 <- lm(pass.density ~ line_grp + age.2013 +
                     passengers.100 + length, data = ship)

 par(mfrow=c(2,2))
 plot(ship_model_1)
```



Residuals vs Fitted

Residuals are relatively evenly distributed to a horizontal line. This line seems to deviate from 0 the further from the middle of the graph on each side, with a small bump in the middle. No distinctive pattern of residuals, though it could be argued that there is some funneling seen. No substantial deviation from a linear line, I am not as worried about the deviation from 0 at the lower and higher end of fitted values as there are less data points at the extremes of the fitted values which could skew results. Assumption of equal variance could be voilated, tests would need to be concluded.

QQ Plot

Some substantial variation from QQ line below -1. Large deviation from QQ at high X values. Assumption of normality does not seem to be met.

Scale Location

Residuals are relatively evenly spread along the range of fitted values, no funneling is seen. Smooth line is slightly curved in shape towards the higher end of fitted values, not linear but most of the variation seems well contained. Little evidence of heteroscedasticity / departure from the assumption of equal variance.

Leverage / Cooks Distance

No highly influential observations, Point 33, 135 and 116 are identified in the plot but these points are clearly well within the threshold of 1 for an influential point. No evidence of highly influential points.

Overall

All diagnostic plots apart from normality seem to hold relatively well, some slight deviations from residuals, no severe funneling or influential points.

To further test the diagnostics covered above, hypothesis tests will be used to confirm or deny the findings above.

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(ship_model_1$res, "pnorm") ##K-S test
```

```
## Warning in ks.test(ship_model_1$res, "pnorm"): ties should not be present for
## the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  ship_model_1$res
## D = 0.34594, 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)
bptest(ship_model_1) ##Breusch-Pagan test
```

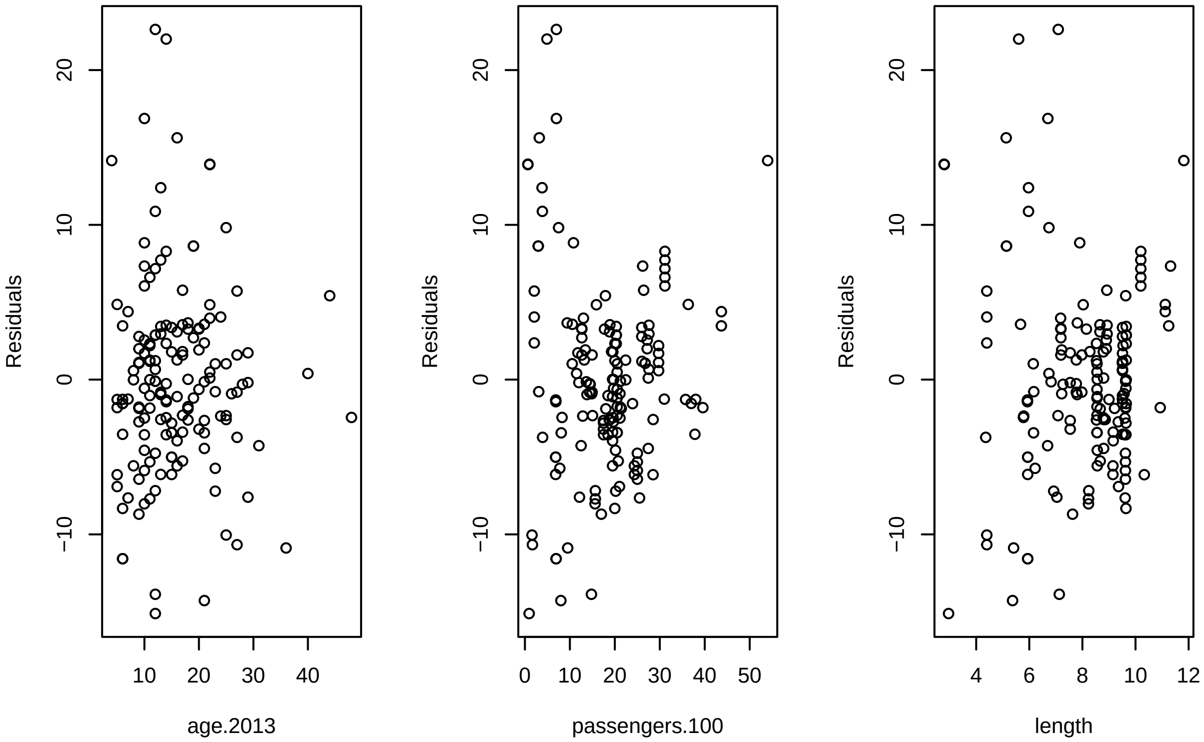
```
##
##  studentized Breusch-Pagan test
##
## data:  ship_model_1
## BP = 63.496, df = 11, p-value = 2.065e-09
```

A p value of 2.065e-09 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 above. I noted that the response variable price did not appear normally distributed, and that there was some amount of variance around some of the predictors.

We can also examine residual plots for each predictor variable to look at the individual relationships of each predictor and to identify if any of these need transformations.

```
# Generate residual plots for each predictor variable
par(mfrow = c(1,3)) # Set up a 1x3 layout for the plots
plot(ship$age.2013, residuals(ship_model_1), xlab = "age.2013", ylab = "Residuals")
plot(ship$passengers.100, residuals(ship_model_1), xlab = "passengers.100", ylab = "Residuals")
plot(ship$length, residuals(ship_model_1), xlab = "length", ylab = "Residuals")
```



Strong funneling, decreasing variance of residuals as both age.2013 and passengers.100 increase. To me this is evidence that a log transformation of either predictors mentioned or the response variable is needed.

There is also evidence of some decreasing variance funneling for length, but not as strong as age.2013 and passengers.100. Again a possible solution for this could be a log transformation of the predictors, or the response.

As some funneling is present in all three numerical predictors the non normality and Heteroscedasticity present in the model diagnostics as well as predictor diagnostics suggest to me that there may be a need for a transformation of the response variable rather than all of the predictor variables.

b.

```
#BIC

ship_model_2 <- lm(log(pass.density) ~ line_grp + age.2013 +
                     passengers.100 + length, data = ship)

step(ship_model_2, direction="both", k=log(nrow(ship)))
```

```
## Start:  AIC=-543.42
## log(pass.density) ~ line_grp + age.2013 + passengers.100 + length
##
##                  Df Sum of Sq    RSS     AIC
## - line_grp        8   0.56032 4.0116 -560.15
## <none>                        3.4512 -543.42
## - length          1   0.57950 4.0307 -523.96
## - passengers.100  1   1.09270 4.5439 -505.02
## - age.2013        1   1.25701 4.7082 -499.41
##
## Step:  AIC=-560.15
## log(pass.density) ~ age.2013 + passengers.100 + length
```

```
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         4.0116 -560.15
## + line_grp         8   0.56032 3.4512 -543.42
## - length           1   0.70052 4.7121 -539.78
## - age.2013         1   1.49609 5.5077 -515.13
## - passengers.100   1   1.92885 5.9404 -503.18
```

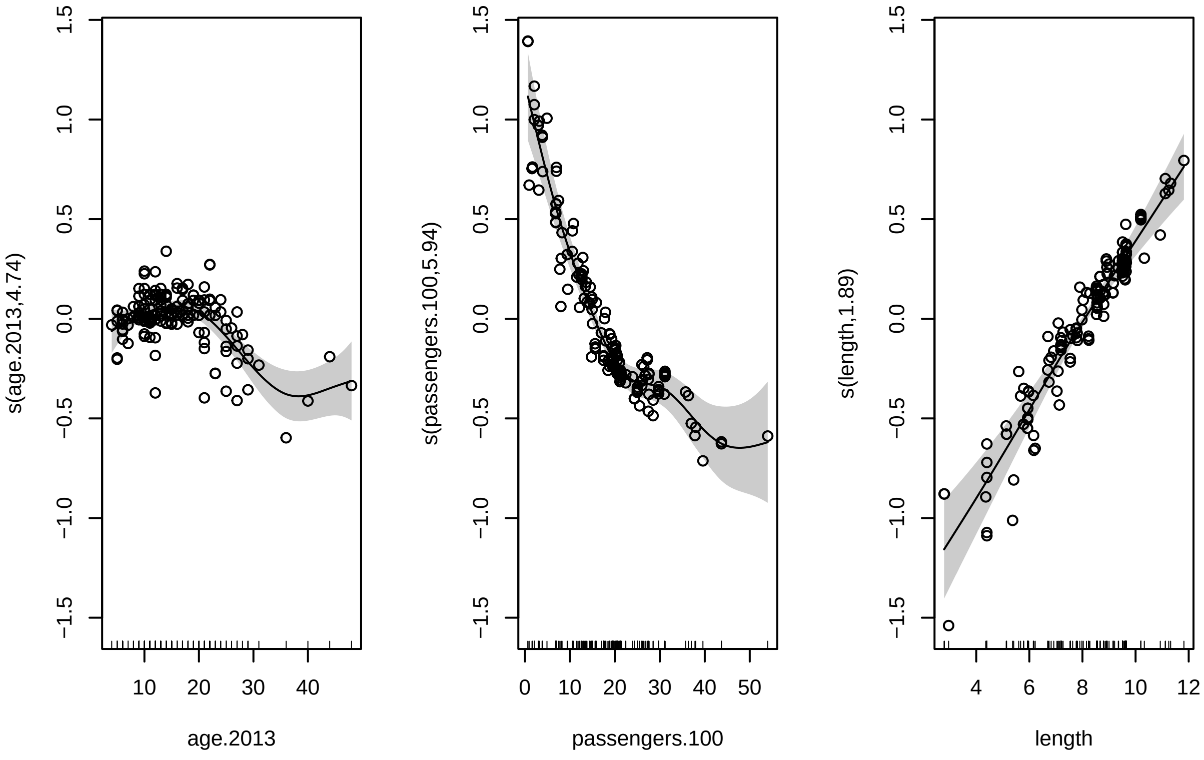
```
##
## Call:
## lm(formula = log(pass.density) ~ age.2013 + passengers.100 +
##     length, data = ship)
##
## Coefficients:
##    (Intercept)        age.2013  passengers.100          length
##        3.69873        -0.01524        -0.02462         0.08102
```

$BIC(A) - BIC(B) = -543.42 - (-560.15) = 16.73

As the difference is in the interval [>10], applying the Raftery BIC rules of thumb means our preferred model is model B which has the lower BIC. This is the model that excludes line_grp.

c.

```
gam_1 <- mgcv::gam(log(pass.density) ~ line_grp + s(age.2013) +
                     s(passengers.100) + s(length), data = ship, method="REML")

par(mfrow=c(1,3))
plot(gam_1, residuals = TRUE, pch = 1,rug=TRUE, scheme = 1)
```



```
summary(gam_1)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(pass.density) ~ line_grp + s(age.2013) + s(passengers.100) +
##     s(length)
##
## Parametric coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.653154   0.028045 130.261   <2e-16 ***
## line_grpCelebrity       -0.038686   0.047145  -0.821   0.4133
## line_grpCosta            0.008168   0.044305   0.184   0.8540
## line_grpHolland American  0.057971   0.047717   1.215   0.2265
## line_grpNorwegian        0.003291   0.042843   0.077   0.9389
## line_grpOther            0.018296   0.050396   0.363   0.7171
## line_grpP&O group        0.040206   0.040297   0.998   0.3202
## line_grpPrincess         0.064140   0.038553   1.664   0.0985 .
## line_grpRoyal Caribbean -0.059244   0.036847  -1.608   0.1102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df     F p-value
## s(age.2013)       4.736  5.785 12.71  <2e-16 ***
## s(passengers.100) 5.939  7.047 17.93  <2e-16 ***
## s(length)         1.886  2.408 48.13  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.709   Deviance explained = 74.7%
## -REML = -78.312  Scale est. = 0.013165  n = 158
```

The significance of each predictor in the model above can be evaluated by the p value of the f test in the model output above.

The non linearity of each predictor in the model above can be evaluated by looking at the graphical output. A significant smooth term is one where you can not draw a horizontal line through the 95% confidence interval band.

Age.2013 and passengers.100 are both significant and non linear. Both have p values which are significantly less than the 0.05 threshold (<2e-16), and a horizontal line cannot be drawn through either of their 95% confidence bands.

Length is significant and linear. The p value associated with it is significantly less than the 0.05 threshold (<2e-16) but a horizontal line can be drawn through the 95% confidence band.

d.

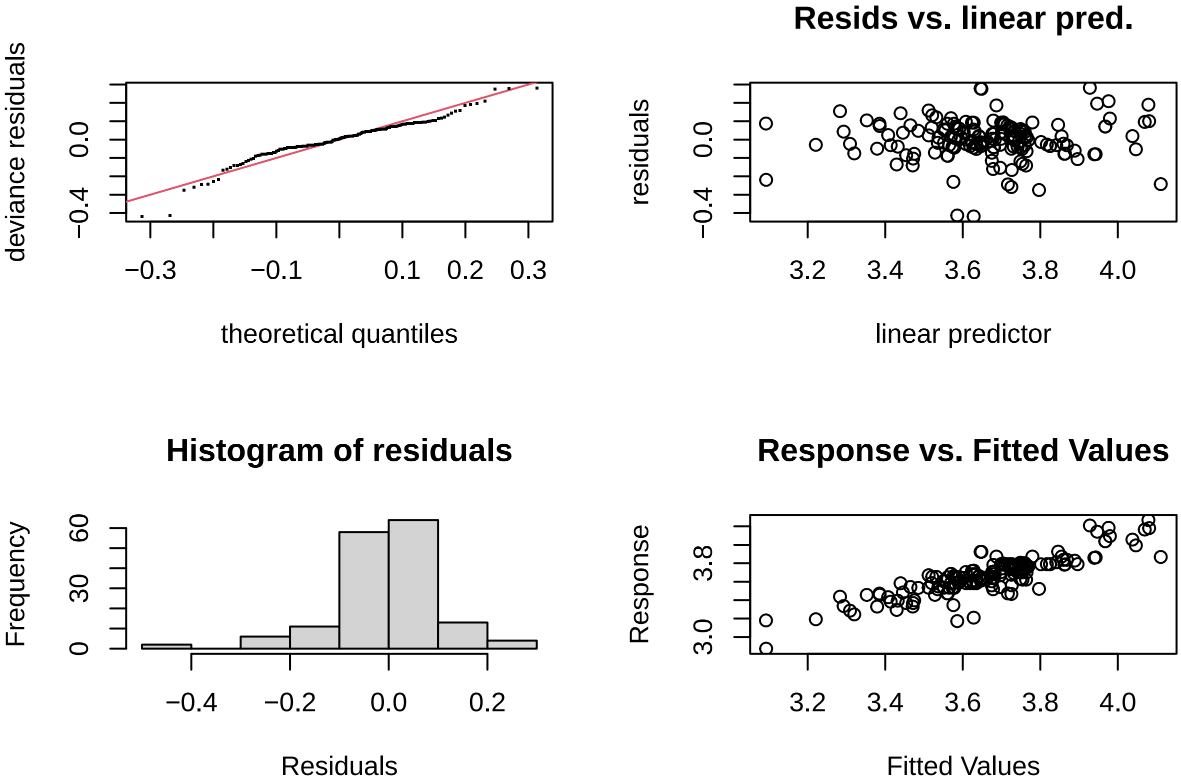
Looking at the graphs from part c the number of basis functions for the smooth terms of passengers.100 and age.2013 seem appropriate. The smooth terms seem to capture a good amount of the variation in the data without over fitting or under fitting.

Using gam.check below I will look at this in further detail.

```
library(mgcv)
```

```
## Loading required package: nlme

## This is mgcv 1.8-41. For overview type 'help("mgcv-package")'.
```

```
par(mfrow=c(2,2))
gam.check(gam_1)
```



```
##
## Method: REML   Optimizer: outer newton
## full convergence after 9 iterations.
## Gradient range [-1.087751e-06,6.394607e-06]
## (score -78.31227 & scale 0.01316489).
## Hessian positive definite, eigenvalue range [0.02700447,73.13671].
## Model rank =  36 / 36
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                      k'  edf k-index p-value
## s(age.2013)       9.00 4.74    1.08   0.825
## s(passengers.100) 9.00 5.94    0.69  <2e-16 ***
## s(length)         9.00 1.89    0.83   0.005 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Full convergence after 6 iterations, results are more reliable than one with non-convergence.

s(passengers.100) with k = 9.00 has a p value of <2e-16 which is very low. This suggests that the residuals are not randomly distributed. The k index is significantly below 1 and edf is the closest to k out of all of the variables. Even though edf is not extremely close to k, both the p value and k index seem to suggest more basis functions may be needed.

Doubling K, and refitting should be completed and compared with the model above.

e.

```r
library(pander)

gam_2 <- fit.gam<-mgcv::gam(log(pass.density) ~ line_grp + age.2013 +
                              passengers.100 + length, data = ship, method="REML")

##Print results in table

#Get BIC values
bic.smooth<-BIC(gam_1)
bic.lin<-BIC(gam_2)

modname<-c("Linear Terms", "Smooth Terms")

BIC_Values<-c(bic.lin, bic.smooth)
mod.compare<-data.frame(modname,BIC_Values)
pander(mod.compare,digits=3, align='c')
```

modname	BIC_Values
Linear Terms	-90
Smooth Terms	-131

$BIC(A) - BIC(B) = -90 - (-131) = 40

As the difference is in the interval [>10], applying the Raftery BIC rules of thumb means our preferred model is model B which has the lower BIC. This is the model that includes smooth terms for age.2013, passengers.100 and length.

f.

BIC is valid to use in the comparison in part e as both models have been fitted with the same estimation method 'REML' and both include the same predictor variables.