DATA 303/473 Assignment 2

Quang Thinh Lam

Due 1159pm Friday 31 March

Assignment Questions

Q1.(20 marks) We'll continue to use the CarDekho data from Assignment 1. As a reminder the variables in the `cardekho2.csv` dataset are:

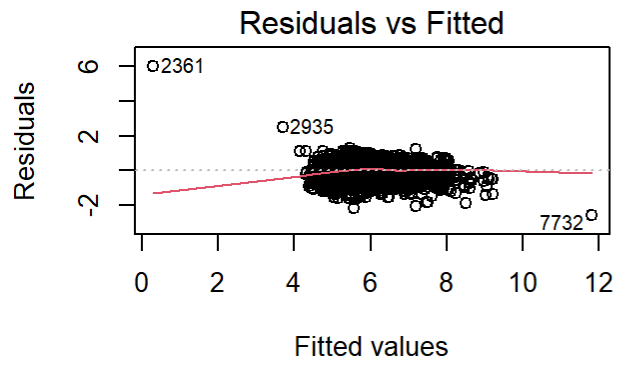
- `price` : Selling price in thousand Indian Rupees (INR)
- `make` : Car make grouped into eight categories: `Ford` , `Honda` , `Hyundai` , `Mahindra` , `Maruti` , `Tata` , `Toyota` , `Other`
- `kms` : Kilometres driven (x 1000)
- `fuel` : Fuel type: `Diesel` or `Petrol`
- `seller` : Seller type: `Dealer` , `Individual` or `Trustmark Dealer`
- `tx` : Transmission type: `Automatic` or `Manual`
- `owner` : Current owner is: `First` , `Second` or `Third or above` owner
- `mileage` : Fuel economy in kilometres per litre (kmpl)
- `esize` : Engine size in cubic centimetres (CC)
- `power` : Maximum engine power in brake horse power (bhp)

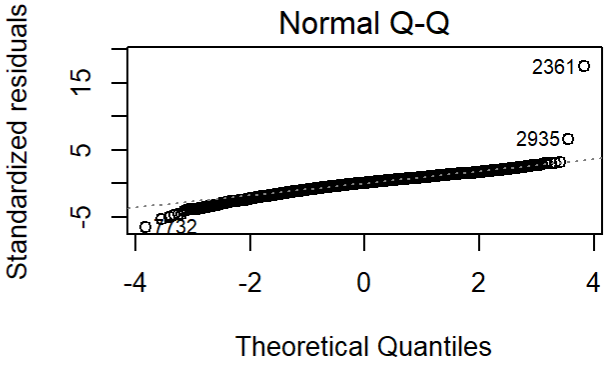
The residual diagnostic plot showed evidence of non-linear relationships between `price` and some predictors, non-normality and non-constant variance. To address non-constant variance, use `log(price)` as the response variable for this assignment.

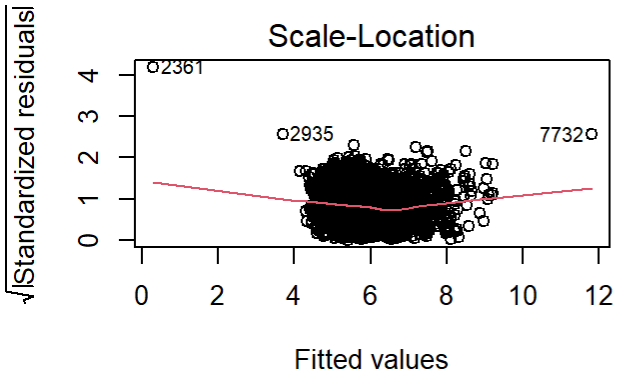
```
#Read the data cardekho2
cardekho2 <- read.csv("cardekho2.csv")
```

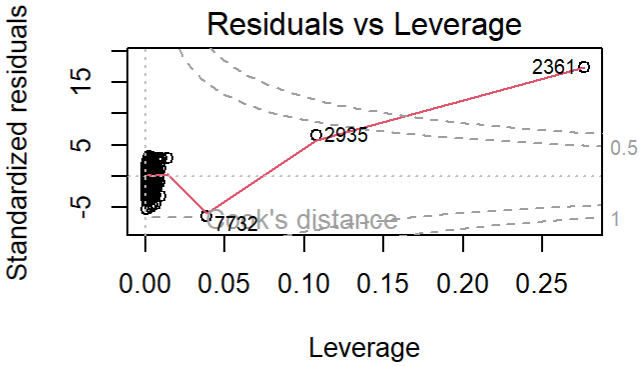
a. **(3 marks)** Fit a model with `log(price)` as the response variable and include all predictors without transformations or interactions. Use the plot function to carry out residual diagnostics for your fitted model. Based on these plots, are there any observations you might consider excluding from further analysis? Explain your answer briefly.

```
#Fit the model
fit1 <- lm(log(price) ~ make + kms + fuel + seller + tx + owner + mileage + esize + power, data
= cardekho2)
par(mfrow = c(2,2))
#Diagnostics plot
plot(fit1)
```









The observations that might be excluded from further analysis: 2361th.

Reasons: The observation 2361th is outside the Cook's Distance threshold, which makes it a highly influential observation in the model.

Although the observations 2935th and 7732th are potential outliers across all four residual diagnostic plots, they are not highly influential, these observations will be retained in further analyses.

Some data cleaning is done and a new dataset, `cardekho3.csv`, (available on Canvas) is created. Use this new dataset to answer the rest of Question 1.

b. **(3 marks)** Read in dataset `cardekho3.csv` and fit the same model as in part (a). Plot the residuals from your fitted model against each of the numerical predictors `kms`, `mileage`, `esize` and `power`. Is there an indication of a non-linear relationships with `log(price)` for any of these predictors? If so, which ones?

```
#Read the new data cardekho3
cardekho3 <- read.csv("cardekho3.csv")
```

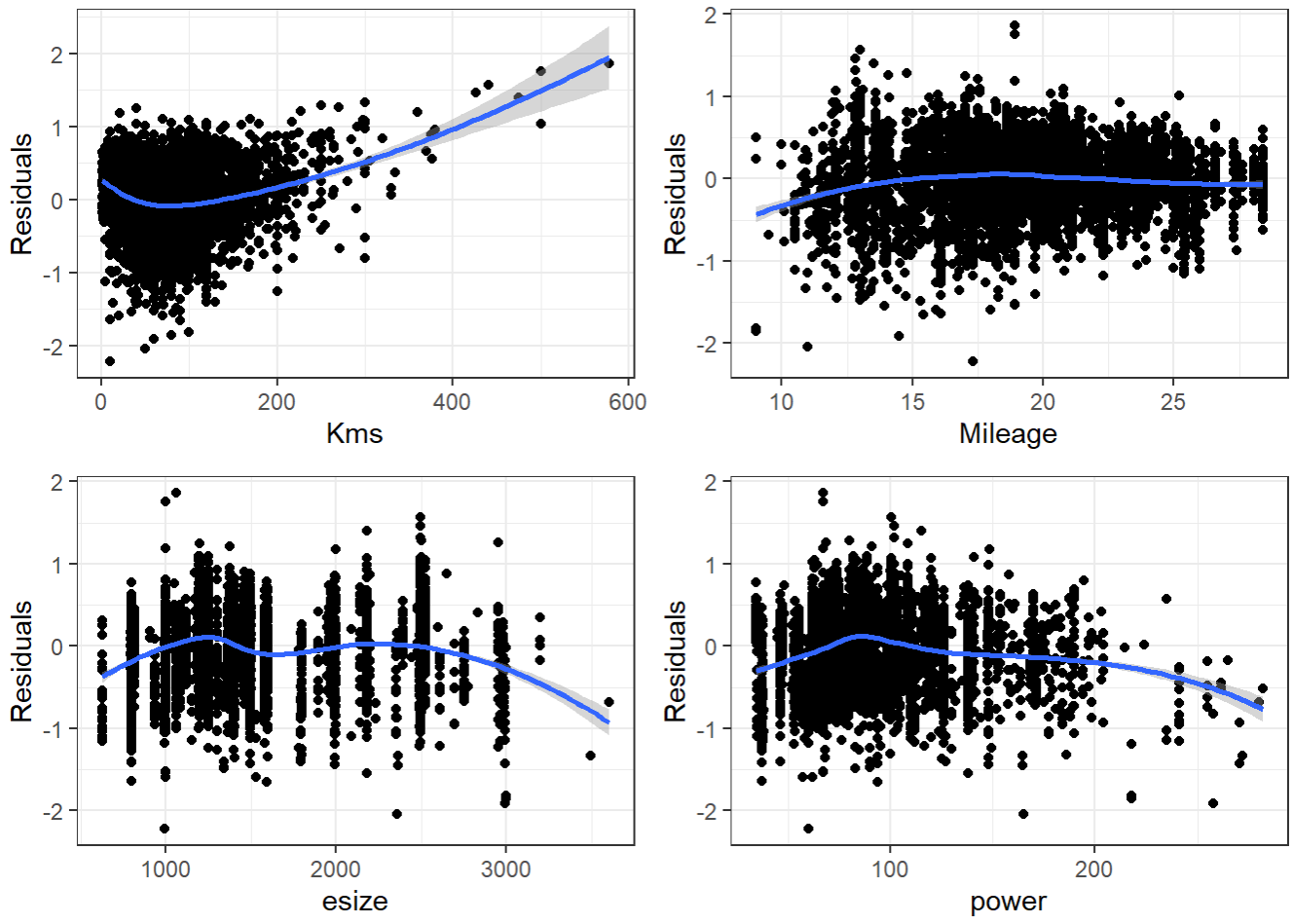
```
fit2 <- lm(log(price) ~ make + kms + fuel + seller + tx + owner + mileage + esize + power, data
= cardekho3)
cardekho3$.resid<-fit2$residuals

kms<-ggplot(cardekho3,aes(x= kms, y=.resid))+
geom_point() + geom_smooth(method='loess')+
labs(x="Kms", y="Residuals")+
theme_bw()

mileage<-ggplot(cardekho3,aes(x=mileage, y=.resid))+
geom_point()+ geom_smooth(method='loess')+
labs(x="Mileage", y="Residuals")+
theme_bw()

esize<-ggplot(cardekho3,aes(x=esize, y=.resid))+
geom_point()+ geom_smooth(method='loess')+
labs(x="esize", y="Residuals")+
theme_bw()

power<-ggplot(cardekho3,aes(x=power, y=.resid))+
geom_point()+ geom_smooth(method='loess')+
labs(x="power", y="Residuals")+
theme_bw()
library(gridExtra)
grid.arrange(kms, mileage, esize, power, nrow=2)
```



There is an indication of non-linear patterns in `kms` , `esize` , and `power` with `log(price)`

c. **(3 marks)** Based on the model fitted in part (b), calculate and give an interpretation for the difference in **price** for a petrol car compared to a diesel car when all other predictors are held constant.

```
summary(fit2)
```

```
##
## Call:
## lm(formula = log(price) ~ make + kms + fuel + seller + tx + owner +
##     mileage + esize + power, data = cardekho3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22253 -0.22384  0.03091  0.25423  1.85885
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.037e+00  6.883e-02  58.658  < 2e-16 ***
## makeHonda                6.815e-02  2.787e-02   2.446   0.0145 *
## makeHyundai              1.103e-01  2.313e-02   4.767 1.90e-06 ***
## makeMahindra             1.983e-01  2.663e-02   7.446 1.06e-13 ***
## makeMaruti               9.294e-02  2.235e-02   4.159 3.23e-05 ***
## makeOther                2.856e-02  2.322e-02   1.230   0.2188
## makeTata                -3.009e-01  2.497e-02 -12.053  < 2e-16 ***
## makeToyota               4.428e-01  3.011e-02  14.707  < 2e-16 ***
## kms                     -3.750e-03  1.148e-04 -32.672  < 2e-16 ***
## fuelPetrol              -1.906e-01  1.393e-02 -13.679  < 2e-16 ***
## sellerIndividual        -7.989e-02  1.433e-02  -5.573 2.58e-08 ***
## sellerTrustmark Dealer -1.619e-02  3.029e-02  -0.535   0.5930
## txManual                -2.311e-01  1.724e-02 -13.409  < 2e-16 ***
## ownerSecond             -2.852e-01  1.118e-02 -25.524  < 2e-16 ***
## ownerThird or above     -4.737e-01  1.742e-02 -27.197  < 2e-16 ***
## mileage                  5.422e-02  1.841e-03  29.456  < 2e-16 ***
## esize                    3.402e-04  2.153e-05  15.797  < 2e-16 ***
## power                    1.261e-02  2.307e-04  54.676  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3937 on 7776 degrees of freedom
## Multiple R-squared:  0.7731, Adjusted R-squared:  0.7726
## F-statistic:  1559 on 17 and 7776 DF,  p-value: < 2.2e-16
```

```
fuel_petrol <- (abs(exp(summary(fit2)$coefficient[10,1]) - 1) * 100)
```

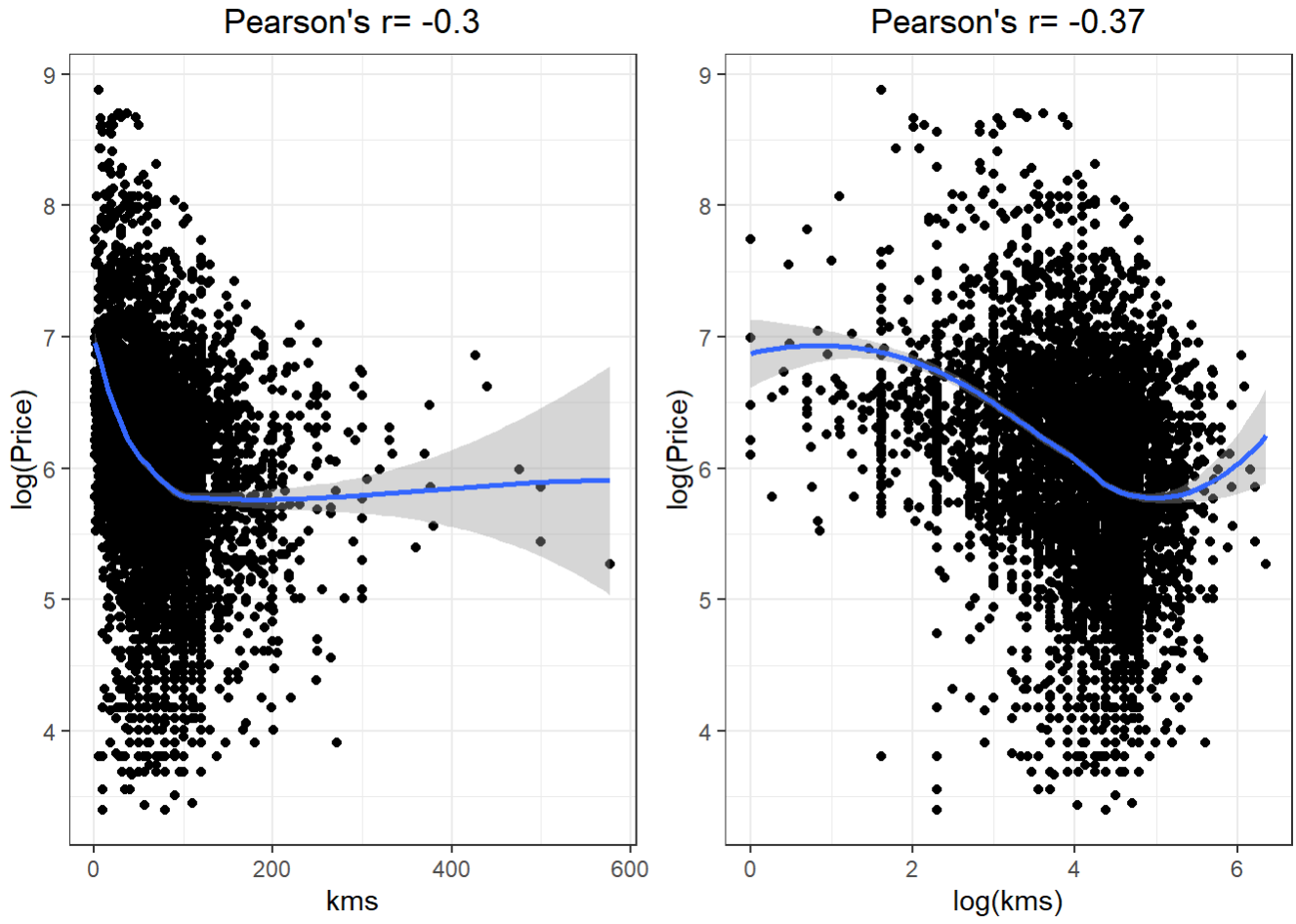
Based on the fitted model, the coefficient of `fuelPetrol` is approximately -0.1906. This indicates that the petrol cars have lower price than diesel cars of approximately 17.3507136%

d. **(4 marks)** Based on the dataset and model in part(b), provide two plots that give graphical evidence that a log transformation is the most appropriate transformation for `kms` in a model for `log(price)` . Explain your reasoning briefly.

```
kms<-ggplot(cardekho3,aes(x=kms, y=log(price)))+
  geom_point()+
  geom_smooth(method='loess')+
  labs(x="kms", y="log(Price)",
       title=paste("Pearson's r=",round(cor(log(cardekho3$price), cardekho3$kms),2)))+
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5))

kms_log<-ggplot(cardekho3,aes(x=log(kms), y=log(price)))+
  geom_point()+
  geom_smooth(method='loess')+
  labs(x="log(kms)", y="log(Price)",
       title=paste("Pearson's r=",round(cor(log(cardekho3$price), log(cardekho3$kms)),2)))+
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(kms,kms_log, nrow=1)
```



The plot of `log(Price)` against `kms` shows a non-linear and monotonic relationship, therefore a transformation of `kms` should be considered. In addition, the values for `kms` are also right-skewed, so a log-transformation for `kms` would be implemented.

e. **(3 marks)** Apply stepwise regression based on the AIC criterion for the model in part (b). Are there any predictors you would exclude from the model? Explain your answer briefly.

```
step(fit2, 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 = cardekho3)
##
## 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
```

We have:

- B: the initial model with all predictors. BIC = -14513
- A: the model which excludes sellers. BIC = -14484

BIC(A) - BIC(B) = -14484 - (-14513) = 29

Excluding any of the predictors results in an increase in AIC of more than 2.5. Therefore, none of the predictors should be excluded from the model.

f. **(4 marks)** Fit a model you would use to investigate whether the effect of `mileage` on `log(price)` depends on the value of `tx`. Based on your model, give the change in $E(log(price))$ associated with a unit increase in `mileage` for a car with:

i. Automatic transmission

ii. Manual transmission.

```
fit3 <- lm(log(price) ~ make + log(kms) + fuel + seller + esize + power + owner + tx + mileage +
mileage:tx, data = cardekho3)
pander(summary(fit3), caption = "")
```

	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
esize	0.0003333	2.075e-05	16.07	3.614e-57
power	0.01286	0.000223	57.68	0
ownerSecond	-0.2512	0.01084	-23.16	7.476e-115
ownerThird or above	-0.4443	0.01678	-26.47	6.663e-148
txManual	0.127	0.06734	1.886	0.05933
mileage	0.06927	0.0035	19.79	4.387e-85
txManual:mileage	-0.01574	0.00356	-4.423	9.888e-06

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

```
mileage_coef <- summary(fit3)$coef["mileage",1]
txManual_mileage_coef <- summary(fit3)$coef["txManual:mileage",1]
auto_change <- mileage_coef + txManual_mileage_coef * 0
manual_change <- mileage_coef + txManual_mileage_coef * 1
```

i. Automatic transmission:

a unit increase in `mileage` for a car with Automatic transmission results in change in $E(log(price))$ = 0.0692721, holding all other variable constant

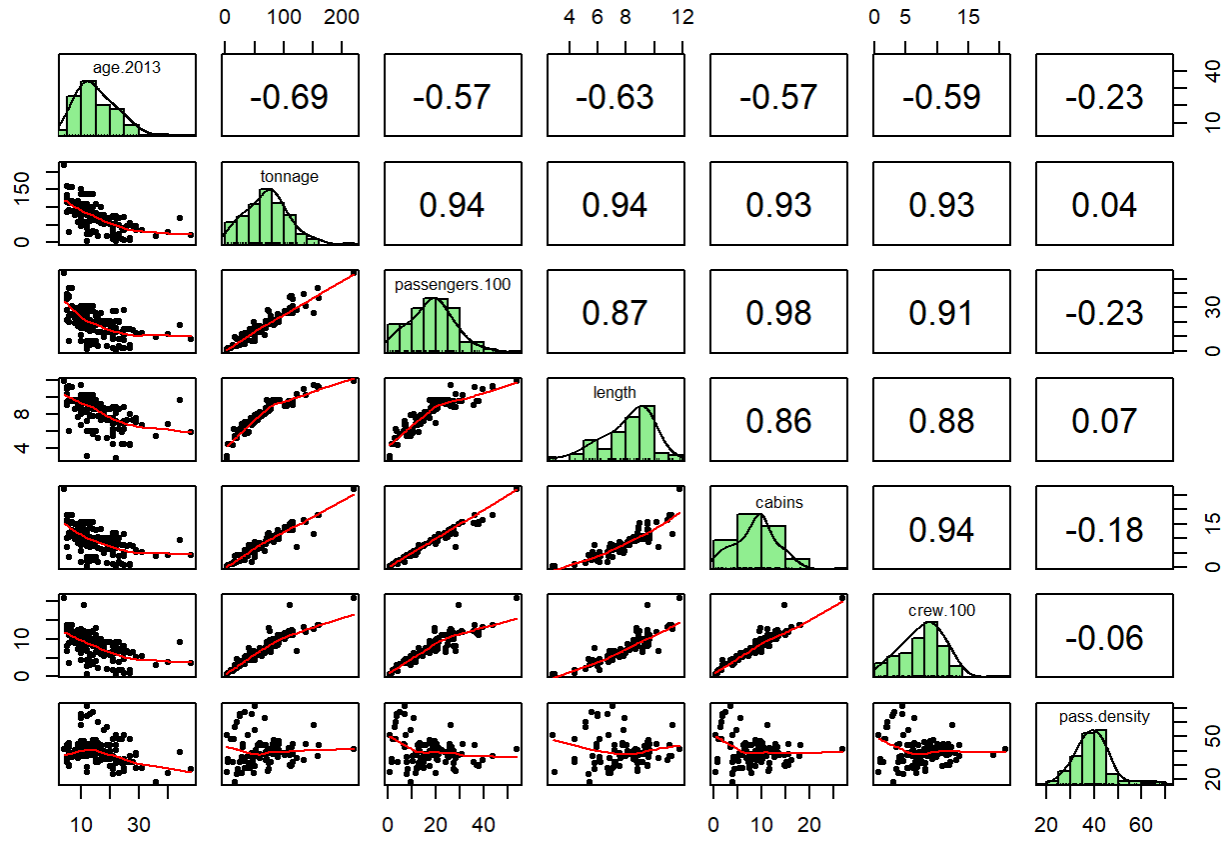
ii. Manual transmission:

a unit increase in `mileage` for a car with Manual transmission results in change in $E(log(price))$ = 0.0535272, holding all other variable constant

Q2.(20 marks) Data were collected on 158 cruise ships in operation around the world in 2013. Complaints had been raised by customers about overcrowding on cruises and there was interest in investigating whether there was a trend of overcrowding on certain types of ships. As part of the investigation, a regression analysis was carried out to explore the connection between passenger density (no. of passengers per unit area) and ship characteristics. The variables in the dataset were:

- `name` : Ship Name
- `line` : Cruise Line
- `line_grp` : Cruise Line grouped
- `age.2013` : Age (as of 2013)
- `tonnage` : Weight of ship (1000s of tonnes)
- `passengers.100` : Maximum no. of passengers (100s)
- `length` : Length of ship (100s of feet)
- `cabins` : No. of passenger cabins (100s)
- `pass.density` : Passenger density (no. of passengers per square foot)
- `crew.100` : No. of crew member (100s)

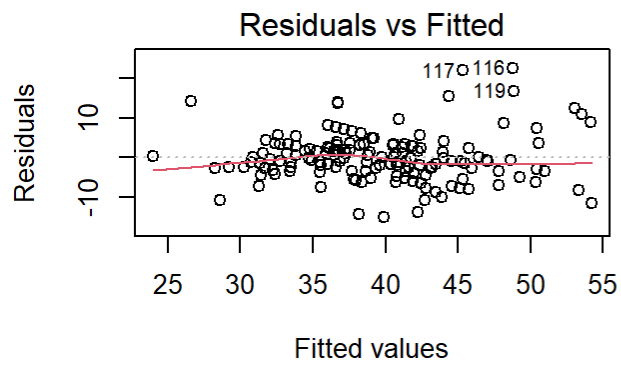
The data are available in the file `cruise_ship.csv` . The dataset was imported into R and the scatterplot matrix below was obtained.

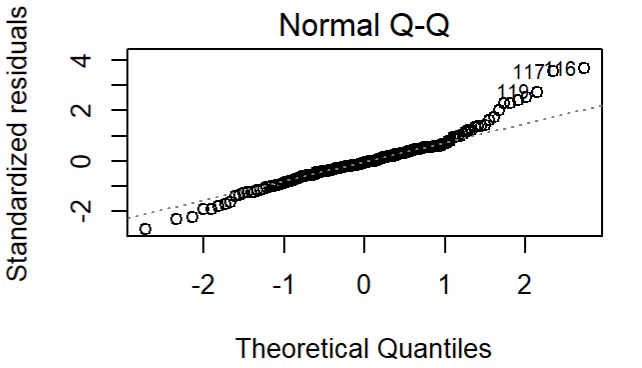


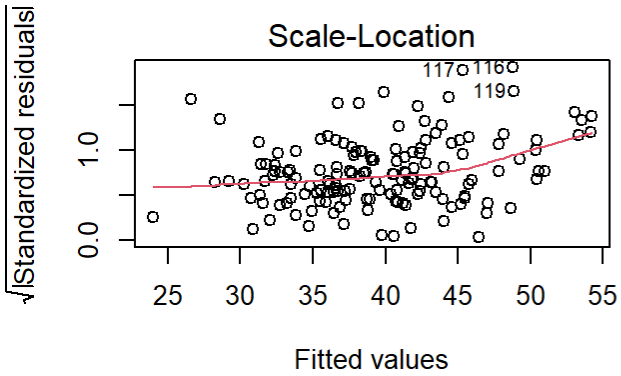
The scatterplot matrix indicates severe multicollinearity among the predictors `tonnage`, `passengers.100`, `length`, and `crews.100`. These four predictors all relate to the size of a ship, so only a subset will be used.

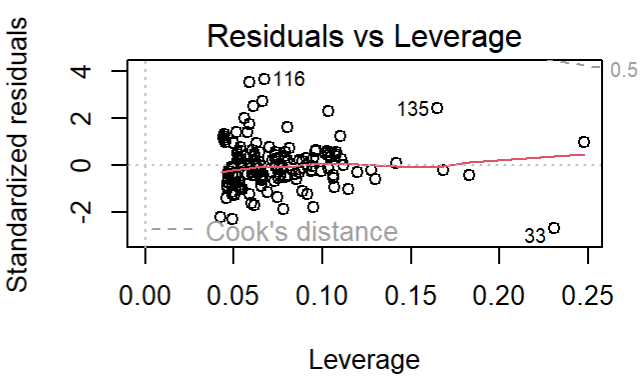
a. **[8 marks]** Fit a model for `pass.density` using the predictors `line_grp`, `age.2013`, `passengers.100` and `length`. Using residual diagnostic checks, determine whether any transformations of the predictors or response variable are necessary. Explain your answer, including identification of which predictors you may need to transform. Provide output of any graphical checks or hypothesis tests you perform.

```
#fit the model
fit4 <- lm(pass.density ~ line_grp + age.2013 + passengers.100 + length, data = cru)
par(mfrow=c(2,2))
# Residual diagnostics plot
plot(fit4)
```









-Linearity - residuals vs. fitted plot: there is no curved pattern in the residuals which indicates a non-linear relationship that was not captured by the model and also does not show up in the residuals. No transformation for the predictors are required

-Normality - normal Q-Q plot: There is slight deviation from the straight line, so there is evidence of potential non-normality. The transformation of the response variable is suggested.

-Equal variance (homoscedasticity) - Scale-Location or Spread-Location plot: there is evidence of non-constant variance. There is a need of transformation of the response variable to address this problem.

-Residuals vs. leverage plot: there are no highly influential observations. Cook's distance lines (a red dashed line) are barely visible because all observations are inside of the Cook's distance thresholds. Therefore, no observation is needed to be excluded.

For the rest of the question use `log(pass.density)` as the response variable.

b. **[3 marks]** Fit a model with `log(pass.density)` as the response variable including all the predictors in part (a) without any transformations. Apply stepwise regression based on the BIC criterion. Are there any predictors you would exclude from the model? Explain your answer briefly.

```
fit5 <- lm(log(pass.density) ~ line_grp + age.2013 + passengers.100 + length, data = cru)
step(fit5, direction = "both", k=log(nrow(cru)))
```

```
## 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 = cru)
##
## Coefficients:
##    (Intercept)        age.2013  passengers.100          length
##        3.69873        -0.01524        -0.02462         0.08102
```

We have:

- B: the model which excludes line_grp. BIC = -560.15
- A: the initial model with all predictors. BIC = -543.42

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

Applying BIC rules of thumb for BIC means there is very strong preference for the model with a smaller BIC value. Therefore, the preferred model is model B, which excludes line_grp and includes three predictors: age.2013, passengers.100, and length.

c. **[3 marks]** Fit a GAM for `log(pass.density)` and smooth terms for each of the predictors `age.2013`, `passengers.100` and `length`. Comment on the non-linearity and significance of smooth terms.

```
library(mgcv)
fit.gam <- gam(log(pass.density) ~ line_grp + s(age.2013) + s(passengers.100) + s(length), data
= cru, method = "REML")
summary.gam <- summary(fit.gam)
pander(summary.gam$s.table, digits=4)
```

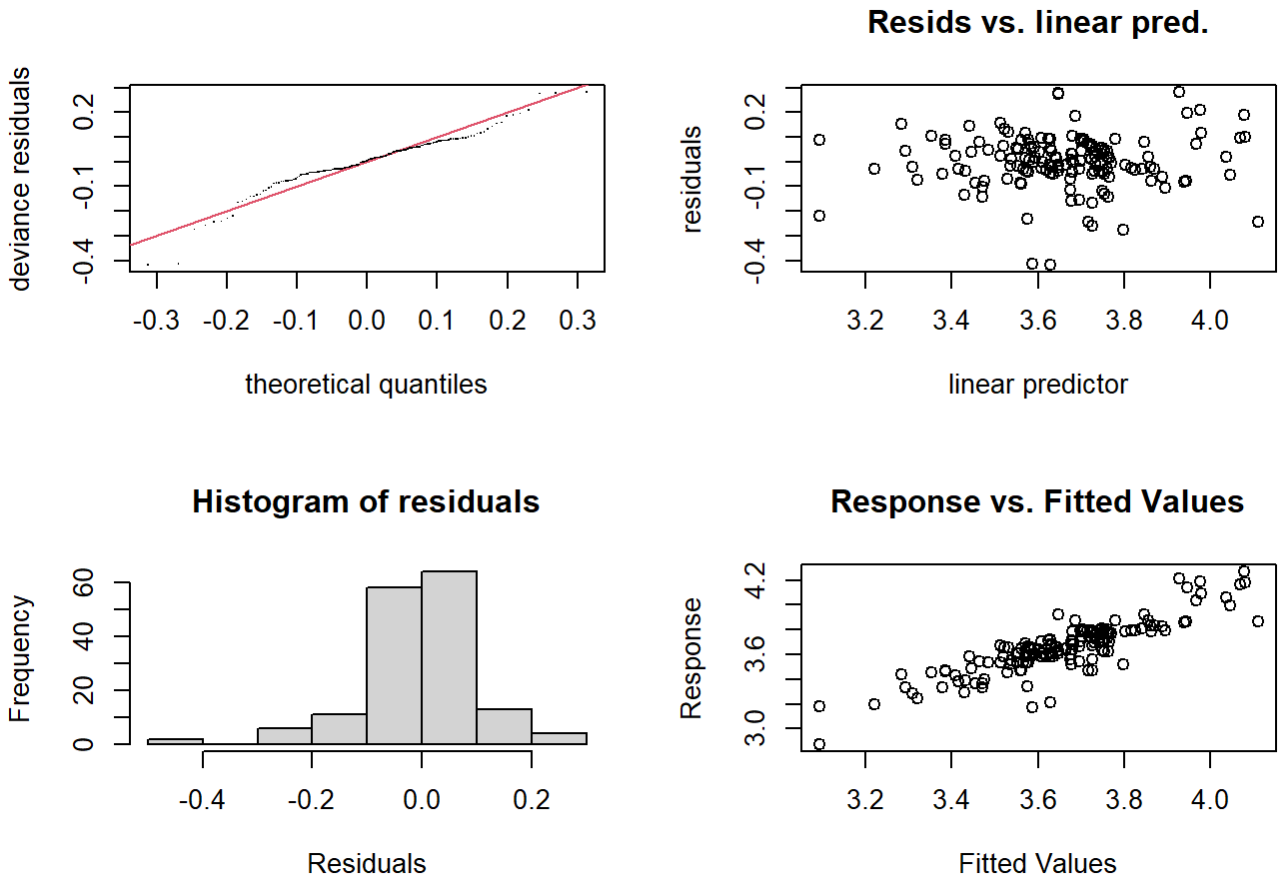
	edf	Ref.df	F	p-value
s(age.2013)	4.736	5.785	12.71	0
s(passengers.100)	5.939	7.047	17.93	0
s(length)	1.886	2.408	48.13	0

In this model:

- age.2013, passengers.100 are both non-linear and significant
- length is linear and significant

d. **[2 marks]** Is there evidence that more basis functions are required for any of the smooth terms? Explain your answer briefly.

```
par(mfrow=c(2,2))
gam.check(fit.gam, k.rep = 1000)
```



```
##
## 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   0.810
## s(passengers.100) 9.00 5.94    0.69  <2e-16 ***
## s(length)         9.00 1.89    0.83   0.017 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If p-value is low, k-index < 1, and edf ≈ k' => more basis functions are needed

In our model:

The p-values for `passengers.100` and `length` are relatively low. However, in all cases, edf is much lower than k, so we likely have enough basis functions.

e. **[3 marks]** Use the `gam()` function to fit a model for `log(pass.density)` with linear terms for all 4 predictors. Calculate BIC for this model and for the model with smooth terms in part (c). Print the results in a table and state which of the models is preferred. Explain your answer briefly.

```
fit.lm <- gam(log(pass.density) ~ line_grp + age.2013 + passengers.100 + length, data = cru, met
hod = "REML")
step(fit5, direction = "both", k=log(nrow(cru)))
```

```
## 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 = cru)
##
## Coefficients:
##    (Intercept)        age.2013  passengers.100          length
##        3.69873        -0.01524        -0.02462         0.08102
```

```
pander(BIC(fit.lm, fit.gam), caption = "")
```

	df	BIC
fit.lm	13	-89.97
fit.gam	25.24	-131.2

Based on the BIC, `fit.gam` model is preferred as it is has lower BIC (-131.2), compared to that of `fit.lm` model (89.97).

However, in order to choose the final preferred model, we also need to consider the AIC and the adjusted R^2.

g. **[1 mark]** Explain briefly why it is valid to make the comparison in part (f) using BIC.

It is valid to make the comparison because both `fit.lm` and `fit.gam` models use the same estimation method, namely 'REML'.