DATA 303/473 Assignment 2

Due 1159pm Friday 31 March

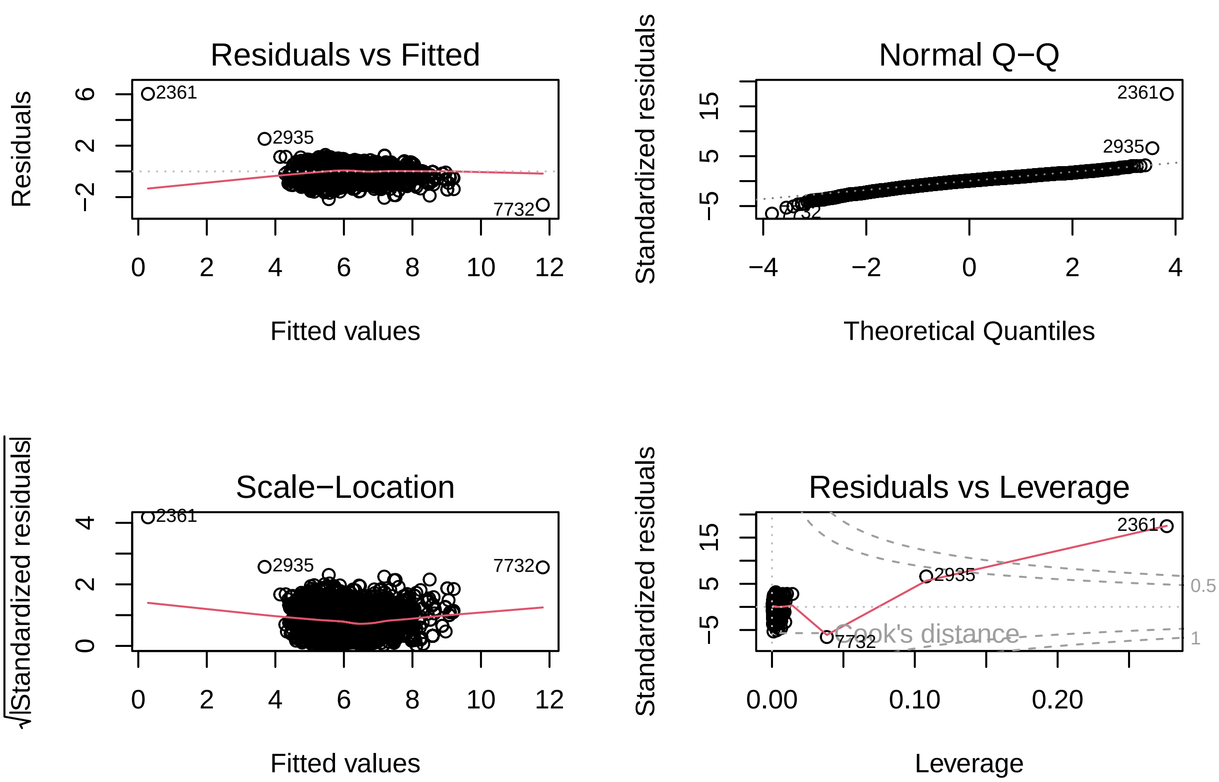
Assignment Questions

Q1.(20 marks) Car sales in India

a. **(3 marks)**

```
cars<-read.csv("cardekho2.csv", header=TRUE, stringsAsFactors = TRUE)

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



I would exclude observation 2361 as it's an influential observation. I would also consider excluding observations 2935 and 7735 as they have standardised residuals that are < -3 or > 3 and leverage values that are much higher than the threshold.

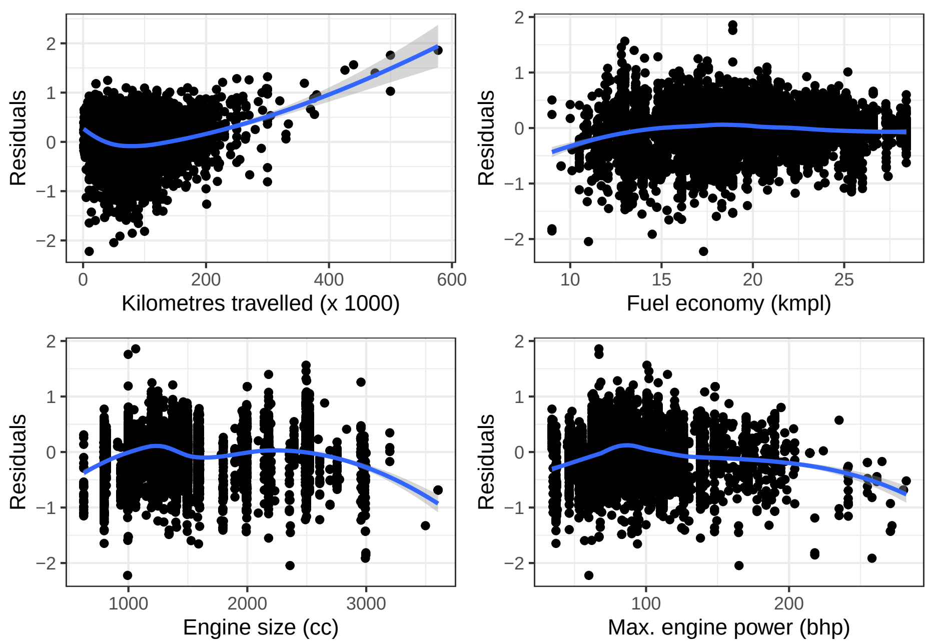
b. **(3 marks)**

```
cars3<-read.csv("cardekho3.csv", header=TRUE, stringsAsFactors = TRUE)

fit2<-lm(log(price) ~ make+ kms + fuel + seller + tx + owner + mileage + esize + power, data=cars3)
cars3$.resid<-fit2$residuals

library(ggplot2)
a<-ggplot(cars3,aes(x=kms, y=.resid))+
  geom_point()+
  geom_smooth(method='loess')+
  labs(x="Kilometres travelled (x 1000)", y="Residuals")+
  theme_bw()
b<-ggplot(cars3,aes(x=mileage, y=.resid))+
  geom_point()+
  geom_smooth(method='loess')+
  labs(x="Fuel economy (kmpl)", y="Residuals")+
  theme_bw()
c<-ggplot(cars3,aes(x=esize, y=.resid))+
  geom_point()+
  geom_smooth(method='loess')+
  labs(x="Engine size (cc)", y="Residuals")+
  theme_bw()
d<-ggplot(cars3,aes(x=power, y=.resid))+
  geom_point()+
  geom_smooth(method='loess')+
  labs(x="Max. engine power (bhp)", y="Residuals")+
  theme_bw()

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



Some non-linearity indicated for `kms`, `esize` and `power`.

c. **(3 marks)**

```
summary(fit2)
```

```
##
## Call:
## lm(formula = log(price) ~ make + kms + fuel + seller + tx + owner +
##     mileage + esize + power, data = cars3)
##
## 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
```

Difference in price for a petrol car vs diesel car, holding all other predictors constant is: $e^{\hat{\beta}_9} - 1 = e^{-0.1906} - 1 = -0.1735$.

We expect price to reduce by a factor of 0.174 for a petrol car relative to a diesel car, holding all other predictors constant.

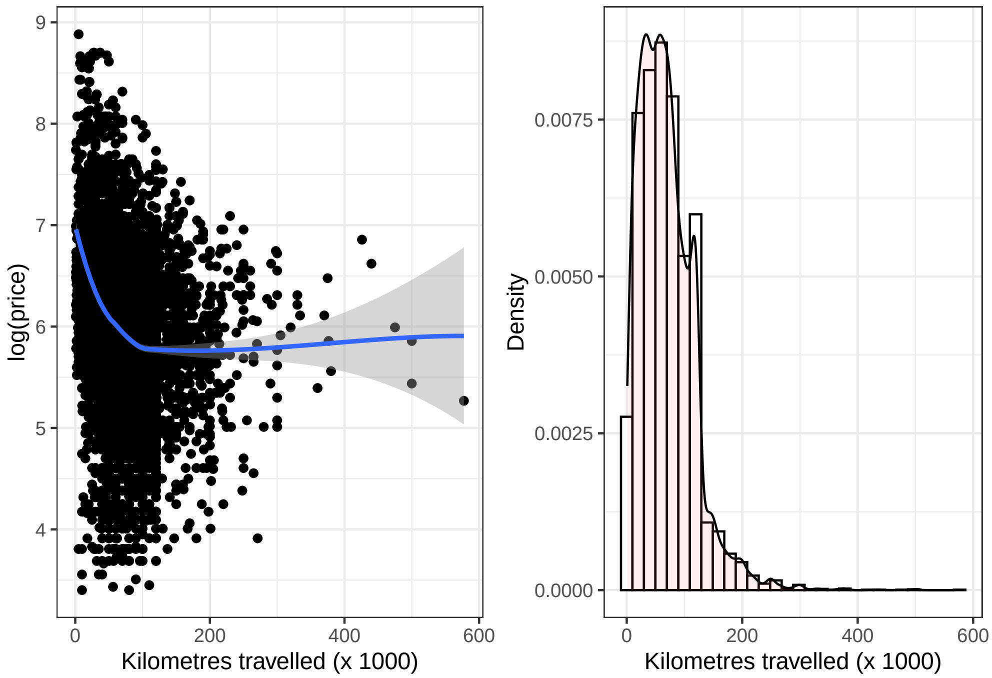
d. **(4 marks)**

```
a<-ggplot(cars3,aes(x=kms, y=log(price)))+
  geom_point()+
  geom_smooth(method='loess')+
  labs(x="Kilometres travelled (x 1000)", y="log(price)")+
  theme_bw()

b<-ggplot(cars3, aes(x=kms)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white")+
```

```
  geom_density(alpha=.1, fill="#FF6666" )+
  labs(x="Kilometres travelled (x 1000)", y="Density")+
  theme_bw()


library(gridExtra)
grid.arrange(a,b, nrow=1)
```



Plot of `log(price)` against `kms` shows a monotonic shape and histogram shows a skewed distribution for `kms`.

e. **(3 marks)**

```
library(MASS)
stepAIC(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 = cars3)
##
## 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
```

I would not exclude any predictors from the model. Excluding any of the predictors results in an increase in AIC.

f. **(4 marks)**

```
fit2.i<-lm(log(price) ~ make+ log(kms) + fuel + seller + tx + owner + mileage + esize +
             power + mileage:tx, data=cars3)
summary(fit2.i)
```

```
##
## Call:
## lm(formula = log(price) ~ make + log(kms) + fuel + seller + tx +
##     owner + mileage + esize + power + mileage:tx, data = cars3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.41213 -0.21387  0.03041  0.25367  1.27416
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.441e+00  8.778e-02  50.596  < 2e-16 ***
## makeHonda              6.774e-02  2.680e-02   2.527   0.0115 *
## makeHyundai            1.149e-01  2.226e-02   5.161 2.52e-07 ***
## makeMahindra           1.899e-01  2.569e-02   7.394 1.58e-13 ***
## makeMaruti             9.740e-02  2.150e-02   4.530 5.98e-06 ***
## makeOther              1.573e-02  2.235e-02   0.704   0.4814
## makeTata              -3.088e-01  2.407e-02 -12.828  < 2e-16 ***
## makeToyota             4.261e-01  2.894e-02  14.724  < 2e-16 ***
## log(kms)              -2.569e-01  6.114e-03 -42.022  < 2e-16 ***
## fuelPetrol            -2.199e-01  1.344e-02 -16.364  < 2e-16 ***
## sellerIndividual      -6.960e-02  1.379e-02  -5.048 4.56e-07 ***
## sellerTrustmark Dealer -3.202e-02  2.932e-02  -1.092   0.2747
## txManual               1.270e-01  6.734e-02   1.886   0.0593 .
## ownerSecond           -2.512e-01  1.084e-02 -23.165  < 2e-16 ***
## ownerThird or above   -4.443e-01  1.678e-02 -26.470  < 2e-16 ***
## mileage                6.927e-02  3.500e-03  19.790  < 2e-16 ***
## esize                  3.333e-04  2.075e-05  16.067  < 2e-16 ***
## power                  1.286e-02  2.230e-04  57.683  < 2e-16 ***
## txManual:mileage      -1.574e-02  3.560e-03  -4.423 9.89e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3788 on 7775 degrees of freedom
## Multiple R-squared:   0.79,  Adjusted R-squared:  0.7895
## F-statistic:  1625 on 18 and 7775 DF,  p-value: < 2.2e-16
```

(i) Automatic: $\hat{\beta}_{15} + \hat{\beta}_{18} \times 0 = 6.927 \times 10^{-2}$

(ii) Manual: $\hat{\beta}_{15} + \hat{\beta}_{18} \times 1 = 6.927 \times 10^{-2} - 1.574 \times 10^{-2} = 5.353 \times 10^{-2} = 0.0535$

Q2.(20 marks) Cruise ship data

a. **[8 marks]**

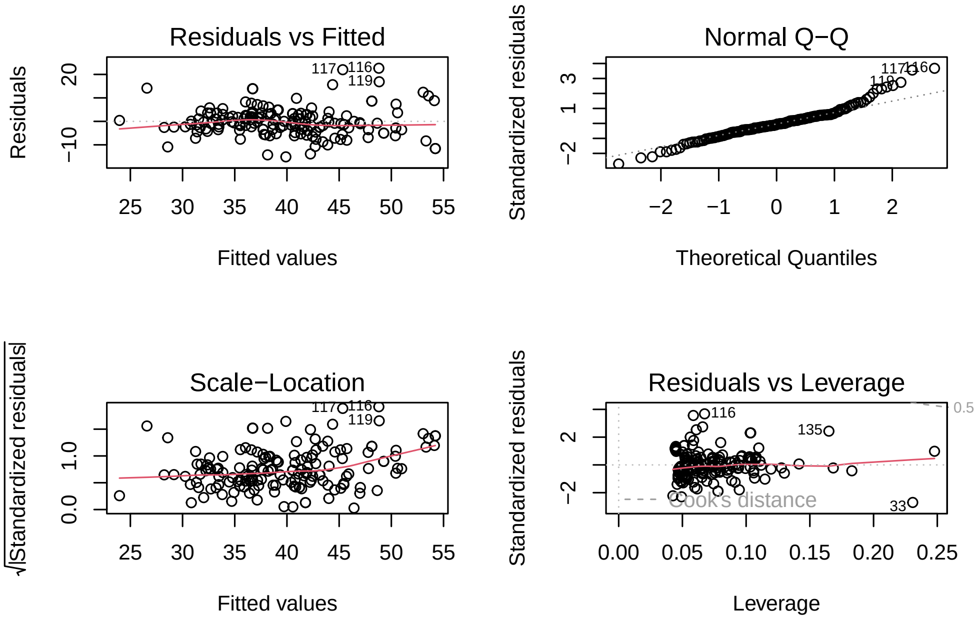
```
fit2<-lm(pass.density~line_grp+ age.2013 + passengers.100 + length, data=cru)
pander(summary(fit2), caption="Summary of fitted model")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	39.9	4.678	8.529	1.717e-14
line_grpCelebrity	-0.7113	2.497	-0.2849	0.7761
line_grpCosta	0.2718	2.374	0.1145	0.909
line_grpHolland American	2.674	2.306	1.16	0.248
line_grpNorwegian	-2.474	2.264	-1.093	0.2764
line_grpOther	7.111	2.214	3.212	0.001621
line_grpP&O group	1.913	2.004	0.9547	0.3413
line_grpPrincess	2.784	2.079	1.339	0.1827
line_grpRoyal Caribbean	2.253	1.942	1.16	0.2478
age.2013	-0.5588	0.0823	-6.79	2.635e-10
passengers.100	-0.8542	0.1312	-6.511	1.126e-09
length	2.775	0.6402	4.334	2.711e-05

Table 2: Summary of fitted model

Observations	Residual Std. Error	R^2	Adjusted R^2
158	6.371	0.4942	0.4561

```
par(mfrow=c(2,2))
plot(fit2)
```




```
ks.test(fit2$residuals, "pnorm")
```

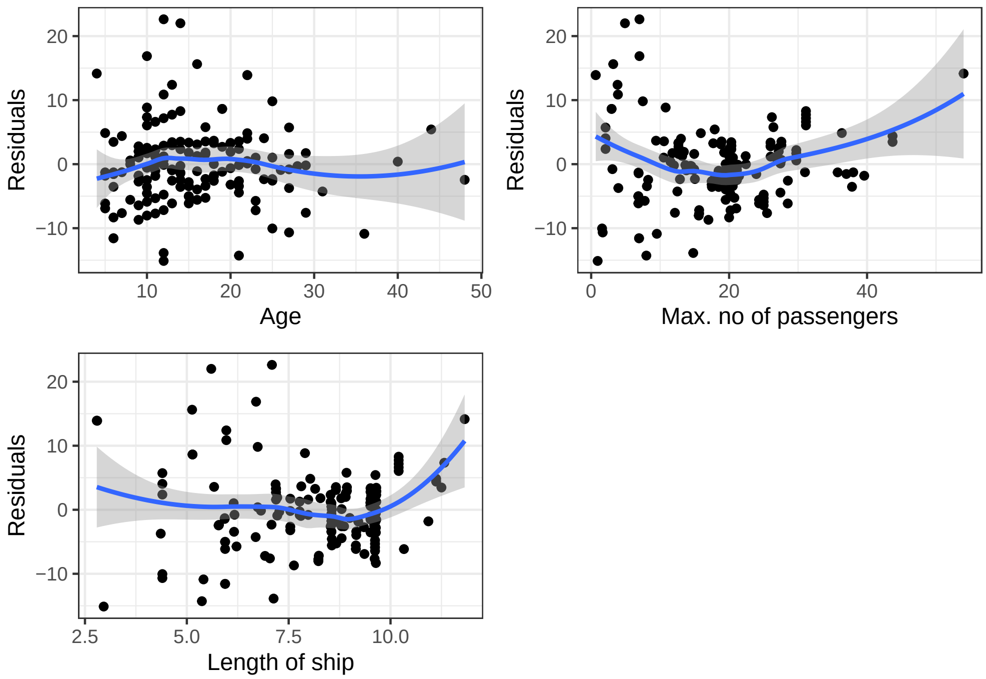
```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  fit2$residuals
## D = 0.34594, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
library(lmtest)
bptest(fit2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit2
## BP = 63.496, df = 11, p-value = 2.065e-09
```

```
library(ggplot2) ;library(gridExtra)
cru$.resid<-fit2$residuals

a<-ggplot(cru,aes(x=age.2013, y=.resid))+
  geom_point()+ geom_smooth(method='loess')+
  labs(x="Age", y="Residuals")+ theme_bw()
b<-ggplot(cru,aes(x=passengers.100, y=.resid))+
  geom_point()+ geom_smooth(method='loess')+
  labs(x="Max. no of passengers", y="Residuals")+ theme_bw()
c<-ggplot(cru,aes(x=length, y=.resid))+
  geom_point()+ geom_smooth(method='loess')+
  labs(x="Length of ship", y="Residuals")+ theme_bw()
grid.arrange(a,b,c, nrow=2)
```



- The Residual vs fitted plot indicates potential non-linear relationships of some predictors with `pass.density`.
- Plots of residuals against each of the numerical predictors indicates non-linearity is present for each of them. Therefore transformations of all numerical predictors will be required.
- The Q-Q plots suggests that the assumption of normal errors does not hold. This is confirmed by the K-S test.
- The scale-location plot suggests the assumption of constant variance does not hold. This is confirmed by the B-P test.
- A log-transformation of the response variable will be required to address non-constant variance and non-normality.

b. [**3 marks**]

```
fit3<-lm(log(pass.density) ~ line_grp+ age.2013 + passengers.100 + length, data=cru)
step(fit3, 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
```

I would exclude `line_grp`. Including it in a model with the other three predictors results in an increase in BIC.

c. **[3 marks]**

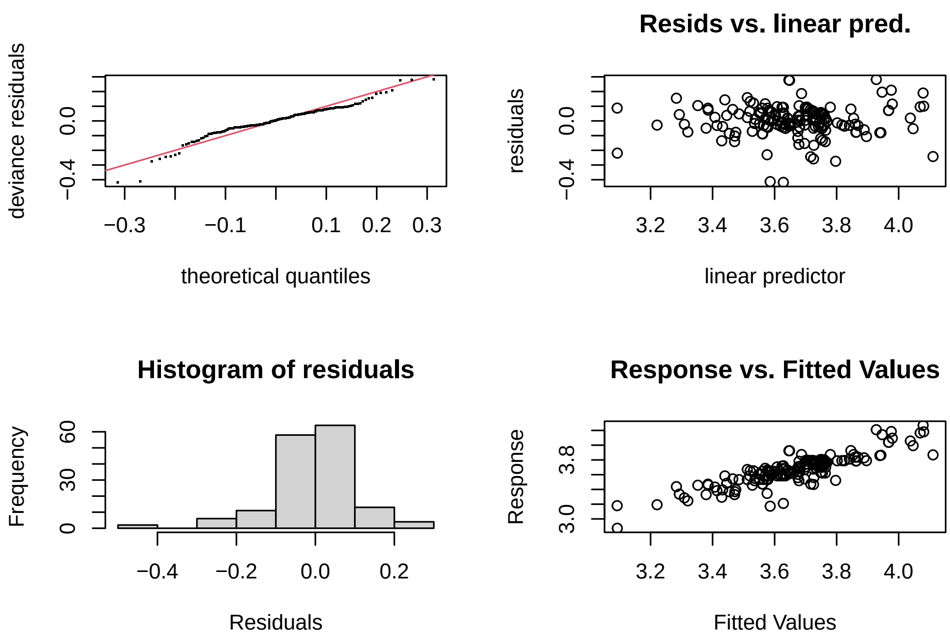
```
library(mgcv)
fit4<-gam(log(pass.density) ~ line_grp + s(age.2013) + s(passengers.100) + s(length),
          data=cru, method="REML")
summary(fit4)
```

```
##
## 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
```

All smooth terms are non-linear and significant.

d. **[2 marks]**

```
par(mfrow=c(2,2))
gam.check(fit4, 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   0.803
## 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
```

No evidence that more smooth terms are required since `edf` is less than `k'` for all predictors.

e. **[3 marks]**

```
fit.lm<-gam(log(pass.density) ~ line_grp + age.2013 + passengers.100 + length,
            data=cru,  method="REML")
fit.gam<-gam(log(pass.density) ~ line_grp + s(age.2013) + s(passengers.100) + s(length),
             data=cru, method="REML")

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

Table 3: BIC values

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

Model with smooth terms is preferred as it has a lower BIC value.

f. **[1 mark]** The models use the same response variable (`log(pass.density)`) and the same estimation method (REML) is used in both cases

Assignment total: 40 marks