assignment 2

Keiryn Hart, 300428418

08/05/2020

```
library(ggplot2)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.6.3
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.3
```

```
## Loading required package: carData
```

```
library(knitr)
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 3.6.3
```

```
library(mgcv)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-28. For overview type 'help("mgcv-package")'.
```

```
library(plyr)
library(broom)
```

```
hybrid <- read.csv("hybrid_reg.csv", header = TRUE)
str(hybrid)
```

```
## 'data.frame':    153 obs. of  9 variables:
##  $ carid      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ vehicle    : Factor w/ 109 levels "3008","A5 BSG",..: 84 103 85 59 31 59 59 10 59 30 ...
##  $ year       : int  1997 2000 2000 2000 2001 2001 2002 2003 2003 2003 ...
##  $ msrp       : num  24510 35355 26832 18936 25833 ...
##  $ accelrate  : num  7.46 8.2 7.97 9.52 7.04 9.52 9.71 8.33 9.52 8.62 ...
##  $ mpg        : num  41.3 54.1 45.2 53 47 ...
##  $ mpgmpge    : num  41.3 54.1 45.2 53 47 ...
##  $ carclass   : Factor w/ 7 levels "C","L","M","MV",..: 1 1 1 7 1 7 7 4 7 1 ...
##  $ carclass_id: int  1 1 1 7 1 7 7 4 7 1 ...
```

Question 1:

a)

```
hybrid$yr_group <- cut(hybrid$year, c(1996,2004,2008,2011,2013))
hybrid$yr_group <- revalue(hybrid$yr_group, c("(1996,2004]"="1997-2004","(2004,2008]"="2005-2008", "(20
hybrid$msrp.1000 <- (hybrid$msrp/1000)
str(hybrid)
```

```
## 'data.frame':    153 obs. of  11 variables:
##  $ carid      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ vehicle    : Factor w/ 109 levels "3008","A5 BSG",..: 84 103 85 59 31 59 59 10 59 30 ...
##  $ year       : int  1997 2000 2000 2000 2001 2001 2002 2003 2003 2003 ...
##  $ msrp       : num  24510 35355 26832 18936 25833 ...
##  $ accelrate  : num  7.46 8.2 7.97 9.52 7.04 9.52 9.71 8.33 9.52 8.62 ...
##  $ mpg        : num  41.3 54.1 45.2 53 47 ...
##  $ mpgmpge    : num  41.3 54.1 45.2 53 47 ...
##  $ carclass   : Factor w/ 7 levels "C","L","M","MV",..: 1 1 1 7 1 7 7 4 7 1 ...
##  $ carclass_id: int  1 1 1 7 1 7 7 4 7 1 ...
##  $ yr_group   : Factor w/ 4 levels "1997-2004","2005-2008",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ msrp.1000  : num  24.5 35.4 26.8 18.9 25.8 ...
```

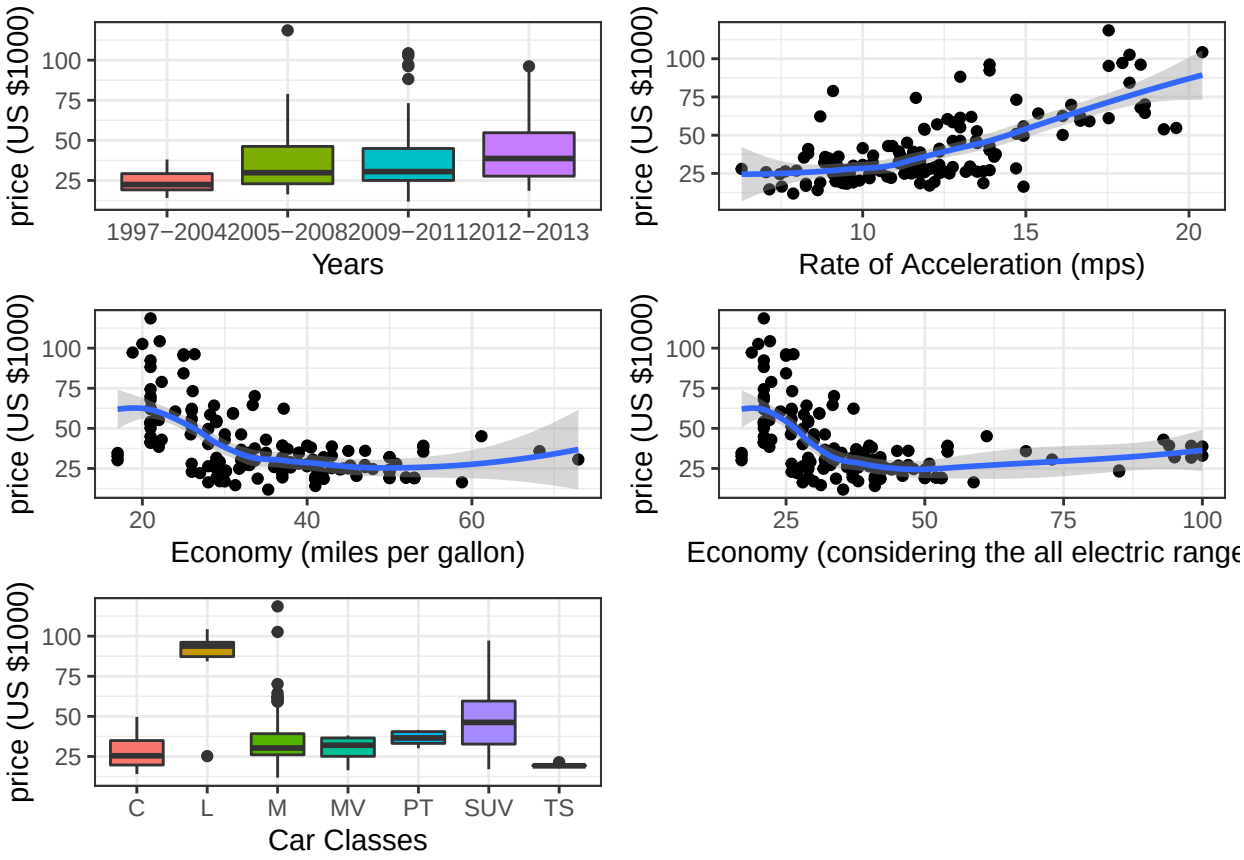
```
table(hybrid$yr_group)
```

```
##
## 1997-2004 2005-2008 2009-2011 2012-2013
##        14        25        57        57
```

b)

```
a<-ggplot(hybrid,aes(x=yr_group, y=msrp.1000))+
  geom_boxplot(aes(fill=yr_group), show.legend=FALSE) +
  labs(x="Years", y="price (US $1000)")+
  theme_bw()

b<-ggplot(hybrid,aes(x=accelrate, y=msrp.1000))+
  geom_point() +
  geom_smooth(method='loess') +
  labs(x="Rate of Acceleration (mps)", y="price (US $1000)")+
  theme_bw()

c<-ggplot(hybrid,aes(x=mpg, y=msrp.1000))+
  geom_point() +
  geom_smooth(method='loess') +
  labs(x="Economy (miles per gallon)", y="price (US $1000)")+
  theme_bw()

d<-ggplot(hybrid,aes(x=mpgmpge, y=msrp.1000))+
  geom_point() +
  geom_smooth(method='loess') +
  labs(x="Economy (considering the all electric range)", y="price (US $1000)")+
  theme_bw()

e<-ggplot(hybrid,aes(x=carclass, y=msrp.1000))+
  geom_boxplot(aes(fill=carclass), show.legend=FALSE) +
  labs(x="Car Classes", y="price (US $1000)")+
```

```
  theme_bw()

grid.arrange(a,b,c,d,e)
```

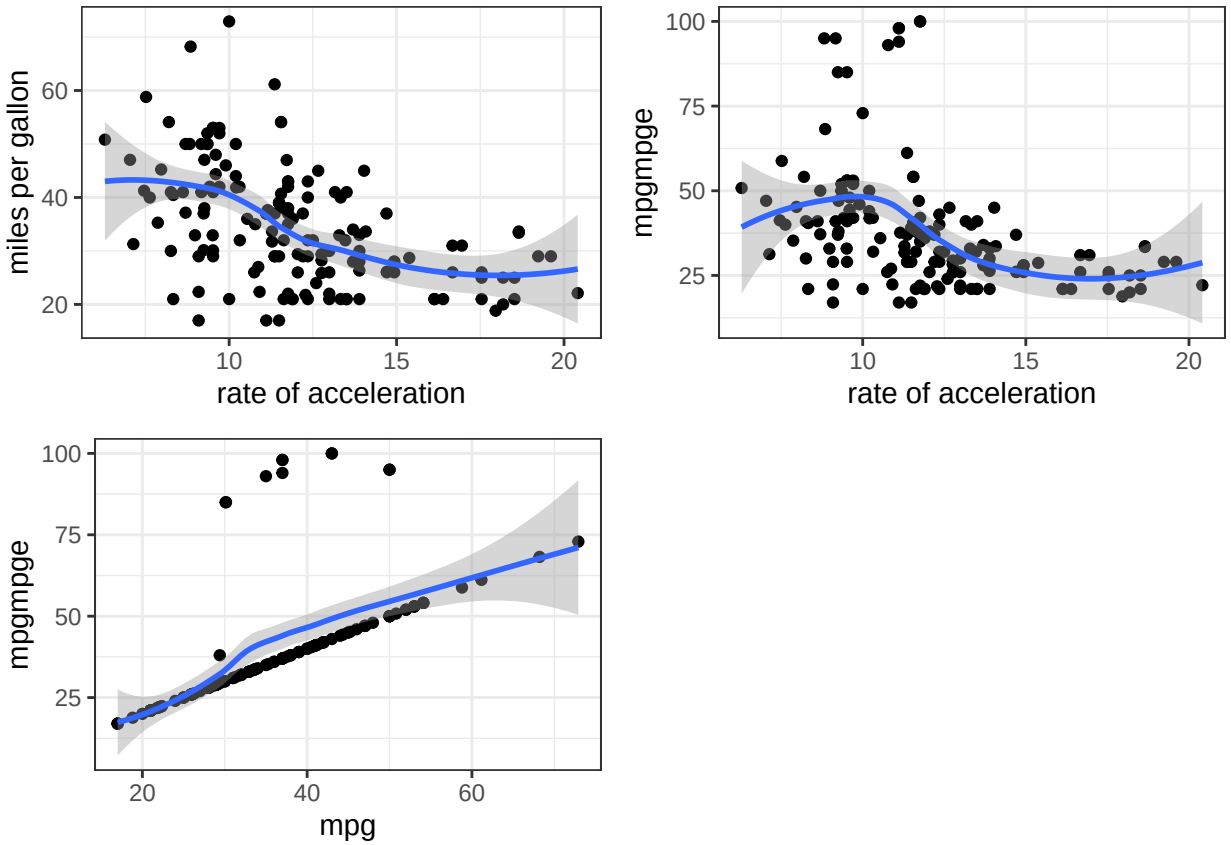


there are indicators of non-linear relationships between the numerical predictors and msrp.1000 most notably in Miles per gallon (mpg) and mpgmpge (miles per gallon when considering electric range aswell)

c)

```
pair1 <-ggplot(hybrid,aes(x=accelrate, y=mpg))+
  geom_point() +
  geom_smooth(method='loess')+
  labs(x="rate of acceleration", y="miles per gallon")+
  theme_bw()

pair2 <-ggplot(hybrid,aes(x=accelrate, y=mpgmpge))+
  geom_point() +
  geom_smooth(method='loess')+
  labs(x="rate of acceleration", y="mpgmpge")+
  theme_bw()

pair3 <-ggplot(hybrid,aes(x=mpg, y=mpgmpge))+
  geom_point() +
  geom_smooth(method='loess')+
  labs(x="mpg", y="mpgmpge")+
  theme_bw()
```

```
grid.arrange(pair1, pair2, pair3, nrow = 2)
```



there is evidence of multicolinearity between mpg and mpgmpge and it shows a linear relationship between the two predictors, possibly due to the fact that both predictors represent similar things and as a result one may not be able to increase without the other.

D)

```
fit1 <- lm(msrp.1000 ~ yr_group + accelrate + mpg + mpgmpge + carclass, data = hybrid)
VIFMODEL <- 1/(1-0.6417)
VIFMODEL
```

```
## [1] 2.790957
```

```
kable(vif(fit1), digits = 2, caption = "VIF Values")%>%
  kable_styling()
```

looking at the VIF values for this model it is interesting to note that none of the values for GVIF^(1/(2*Df)) exceed or even come close to 10 and none of them exceed the VIF model value of 2.790957. these results are quite surprising given the fact that the previous pairwise scatter plots indicated cases of multicolinearity between some of the predictors.

E)

Table 1: VIF Values

	GVIF	Df	GVIF^(1/(2*Df))
yr_group	1.71	3	1.09
accelrate	1.91	1	1.38
mpg	3.16	1	1.78
mpgmpge	1.98	1	1.41
carclass	3.76	6	1.12

```
fit.gam<-gam(msrp.1000 ~ yr_group + s(accelrate) + s(mpg) + s(mpgmpge) + carclass , data=hybrid, method=

summ.gam <- summary(fit.gam)
RSE <- summ.gam$scale
adjRsq <- summ.gam$r.sq
Rsq <- summ.gam$dev.expl
statistics <- c("RSE", "AdjRsq", "Rsq")
values <- c(RSE, adjRsq, Rsq)
stats <- data.frame(statistics, values)
kable(stats, booktabs = T, digits = 3)
```

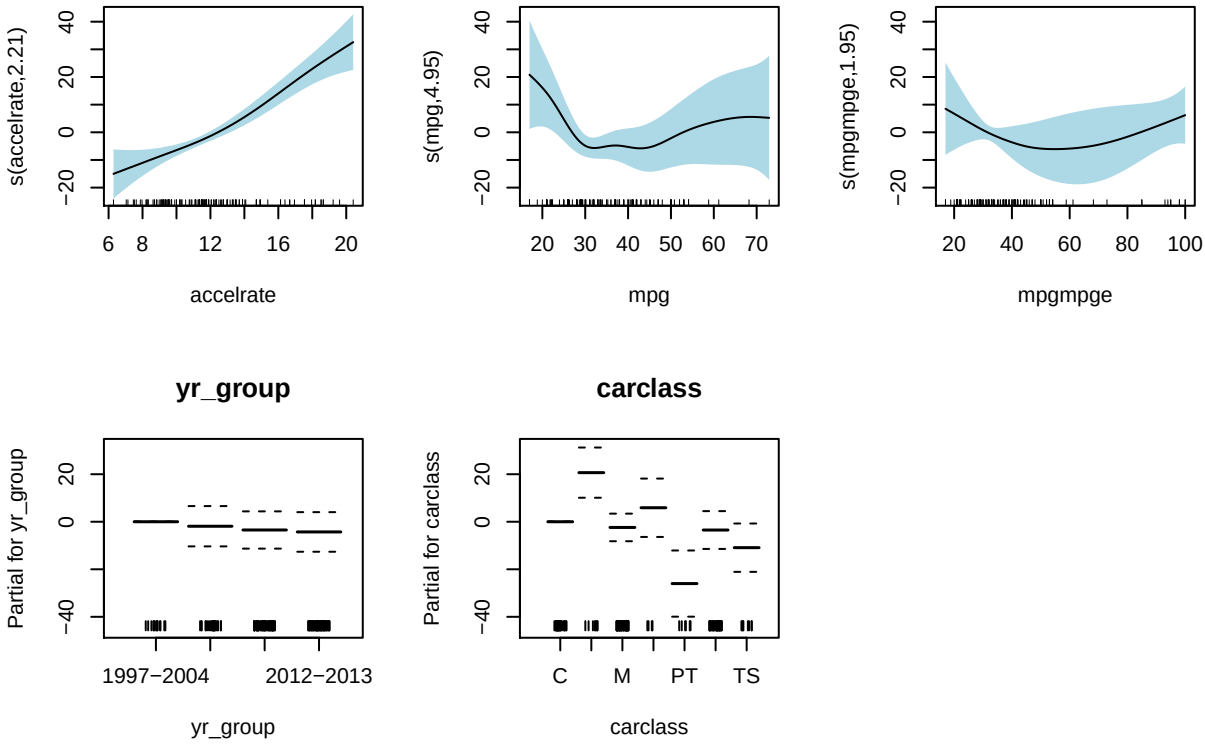
statistics	values
RSE	118.668
AdjRsq	0.741
Rsq	0.772

f)

```
summ.gam <- summary(fit.gam)
kable(summ.gam$s.table, booktabs="T", digits = 3)%>%
  kable_styling()
```

	edf	Ref.df	F	p-value
s(accelrate)	2.209	2.803	24.474	0.000
s(mpg)	4.946	6.027	2.700	0.019
s(mpgmpge)	1.950	2.324	1.115	0.361

```
plot(fit.gam, all.terms = TRUE,rug=TRUE,residuals=FALSE,
     pch=19, cex=0.65, scheme = 1, shade.col="lightblue", page=1)
```



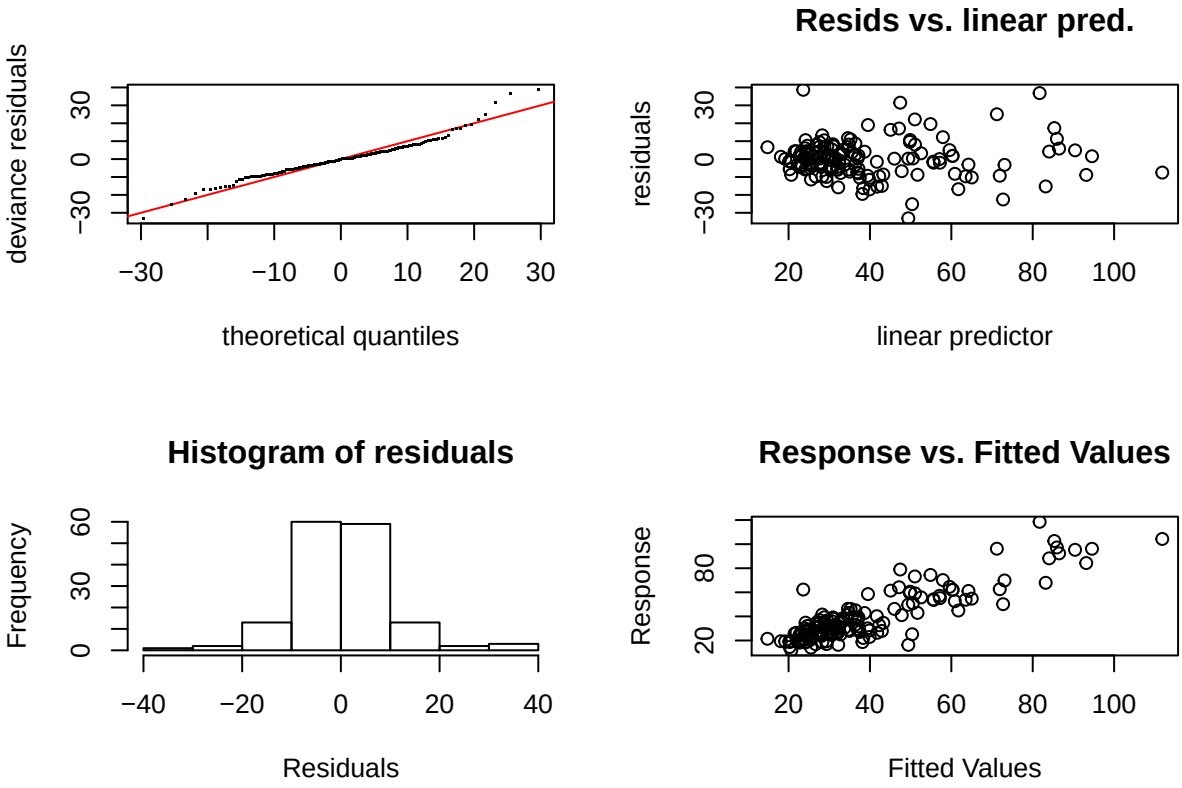
we can see that accelrate and mpg are non-linear and significant and mpgmpge is non-linear and non significant. accelrate has an efd value of 2.209 which indicates it is a quadratic curve with an f-statisti of 24.474 and an extremely low p-value which indicates it is significant.

mpg has an edf value of 4.946 which indicates that its curve is quite wiggly, it has a f-statistic of 2.700 and a p-value of 0.019 which also indicates its significance.

mpgmpge has an edf value of 1.950 which is the lowest of the numerical predictors, it has an fstatistic of 1.115 and a p-value of 0.361, given its large p-value we can be relatively confident in saying that it does not have a significant non-linear effect on msrp.1000.

g)

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




```
##
## Method: REML   Optimizer: outer newton
## full convergence after 5 iterations.
## Gradient range [-6.260482e-08,6.430341e-08]
## (score 559.64 & scale 118.6676).
## Hessian positive definite, eigenvalue range [0.212487,70.0645].
## Model rank =  37 / 37
##
## 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(accelrate) 9.00 2.21    1.06   0.729
## s(mpg)       9.00 4.95    0.81   0.006 **
## s(mpgmpge)   9.00 1.95    0.80   0.005 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

looking at the table of basis dimensions we can see that the maximum number of basis functions for this model is 9 throughout, for accelrate we can see that the k index isvery close to 1 and the p-value is quite large which indicates that we may have enough basis functions for this predictor.

for the other 2 predictors mpg and mpgmpge we can see that these predictors have k-index values which are quite far away from one being 0.81 and 0.80 repsectively, their p-values are also very low being 0.008 and 0.006, this is an indication that there may not be enough basis functions for these two variables.

Looking at the Q-Q plot for this model we can see that the residuals do not follow a straight line all the way through which indicates that there is non-normality.

when looking at the residual vs linear predictor plot it is quite easy to see that this plot indicates non linearity within the model, this could be due to potential outliers.

the histogram of residuals has a relatively symmetrical bell shape which is expected.

the response against fitted values plot forms a relatively straight line with variatio obviously as it is not a perfect model.

h)

```
model2 <-gam(msrp.1000 ~ yr_group + s(accelrate) + s(mpgmpge) + carclass , data=hybrid, method="REML")

model3 <-gam(msrp.1000 ~ yr_group + s(accelrate) + s(mpg) + carclass , data=hybrid, method="REML")

model4<-gam(msrp.1000 ~ yr_group + s(accelrate) + carclass , data=hybrid, method="REML")

likelihood1 <- anova(fit.gam, model2, test = 'F')
likelihood2 <- anova(fit.gam, model3, test = 'F')
likelihood3 <- anova(fit.gam, model4, test = 'F')

likelihood1
```

```
## Analysis of Deviance Table
##
## Model 1: msrp.1000 ~ yr_group + s(accelrate) + s(mpg) + s(mpgmpge) + carclass
## Model 2: msrp.1000 ~ yr_group + s(accelrate) + s(mpgmpge) + carclass
##   Resid. Df Resid. Dev      Df Deviance      F Pr(>F)
## 1     129.8      15889
## 2     132.7      16503 -2.9043  -613.95 1.7814 0.1557
```

```
likelihood2
```

```
## Analysis of Deviance Table
##
## Model 1: msrp.1000 ~ yr_group + s(accelrate) + s(mpg) + s(mpgmpge) + carclass
## Model 2: msrp.1000 ~ yr_group + s(accelrate) + s(mpg) + carclass
##   Resid. Df Resid. Dev     Df Deviance      F Pr(>F)
## 1    129.80      15889
## 2    131.84      16384 -2.046  -494.61 2.0371 0.1334
```

```
likelihood3
```

```
## Analysis of Deviance Table
##
## Model 1: msrp.1000 ~ yr_group + s(accelrate) + s(mpg) + s(mpgmpge) + carclass
## Model 2: msrp.1000 ~ yr_group + s(accelrate) + carclass
##   Resid. Df Resid. Dev      Df Deviance      F    Pr(>F)
## 1     129.8      15889
## 2     137.8      24070 -8.0037  -8180.5 8.6131 2.245e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

when looking at the liklihood ratio test for models one and two the p-value is quite high which indicates that we cannot reject the null hypothesis and say that these models do not differ significantly, but when looking at the Residual deviance we can see that the value for model 1 is slightly less than the value for model 2 indicating that, model 1 (fit.gam) is the better fit out of the two.

looking at the second liklihood ratio test for models 1 and 3 the p-value is also quite large which indicates we cannot reject the null hypothesis. when looking at the residual deviance we can see that again similar to the first test the value for model 1 is less than the value for model 3 by around 500 indicating again that model 1 is the better fit.

looking at the third and final test between models 1 and 4 the p-value is very small which indicates that we can reject the null hypothesis and say that these two model differ significantly. The residual deviance value for model 4 is roughly 8000 more than that of model 1 which again indicates that model 1 is the beter fit.

i)——-

The results in part (h) indicates that when both mpg and mpge are included in the model, the model has a better fit and explains more variance. when just one or the other variable is included the difference between the fit of the models is worse than the original although this is not by much.this highlihgts the pitfall of multicolinearity.

j)——-

the results in part h and i are relatively surprising given the findings in part (d) purely because section (d) shows the original model that includes both the variables mpg and mpgmpge clearly has indications of multicolinearity between these two variable. But when the liklihood ratio tests are run it shows that model 1 is still the better model out of various others that do and do not include one or the other of these variables, this is strange.

k)——-

When considering the AIC values you would choose model 2 but it is hard to say because the differences between the AIc values in the first 3 models is hardly indifferent with model 2 being slightly smaller. But when considering the BIC values model 2 becomes the obvious choice as it is smaller than all the other models and the difference in every case is larger than 2 meaning there is a positive difference favouring model 2.

```
aic.model1 <- AIC(fit.gam)
aic.model2 <- AIC(model2)
aic.model3 <- AIC(model3)
aic.model4 <- AIC(model4)

bic.model1 <- BIC(fit.gam)
bic.model2 <- BIC(model2)
bic.model3 <- BIC(model3)
bic.model4 <- BIC(model4)


n <- c("AIC", "BIC")
m1 <- c(aic.model1, bic.model1)
m2 <- c(aic.model2, bic.model2)
m3 <- c(aic.model3, bic.model3)
m4 <- c(aic.model4, bic.model4)

df <- data.frame(n, m1, m2, m3, m4)
df
```

```
##     n       m1       m2       m3       m4
## 1 AIC 1188.873 1189.562 1190.066 1238.712
```

```
## 2 BIC 1256.008 1248.953 1251.902 1285.080
```