DATA 303/473 Assignment 2

Nicholas Tran, 300296259

Due: 31 March 2022

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

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

```
library(mgcv)
```

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

```
## This is mgcv 1.8-31. For overview type 'help("mgcv-package")'.
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:nlme':
##
##     collapse
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.0.5
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
library(pander)
```

```
## Warning: package 'pander' was built under R version 4.0.5
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

Q1. (35 marks)

a. (3 marks)

```r
hybrid <- read.csv("hybrid_reg.csv")
hybrid$year <- as.factor(hybrid$year)
hybrid$carclass <- as.factor(hybrid$carclass)
hybrid$carclass_id <- as.factor(hybrid$carclass_id)
hybrid <- hybrid %>% mutate(msrp.1000 = msrp/1000,
                            yr_group = case_when(
                              year %in% 1997:2004 ~ "1997-2004",
                              year %in% 2005:2008 ~ "2005-2008",
                              year %in% 2009:2011 ~ "2009-2011",
                              year %in% 2012:2013 ~ "2012-2013"
                            ))
hybrid$yr_group <- as.factor(hybrid$yr_group)
head(hybrid)
```

```
##   carid          vehicle year     msrp accelrate   mpg mpgmpge carclass
## 1     1 Prius (1st Gen) 1997 24509.74      7.46 41.26   41.26        C
## 2     2            Tino 2000 35354.97      8.20 54.10   54.10        C
## 3     3 Prius (2nd Gen) 2000 26832.25      7.97 45.23   45.23        C
## 4     4         Insight 2000 18936.41      9.52 53.00   53.00       TS
## 5     5 Civic (1st Gen) 2001 25833.38      7.04 47.04   47.04        C
## 6     6         Insight 2001 19036.71      9.52 53.00   53.00       TS
##   carclass_id msrp.1000  yr_group
## 1           1  24.50974 1997-2004
## 2           1  35.35497 1997-2004
## 3           1  26.83225 1997-2004
## 4           7  18.93641 1997-2004
## 5           1  25.83338 1997-2004
## 6           7  19.03671 1997-2004
```

```r
addmargins(table(hybrid$yr_group))
```

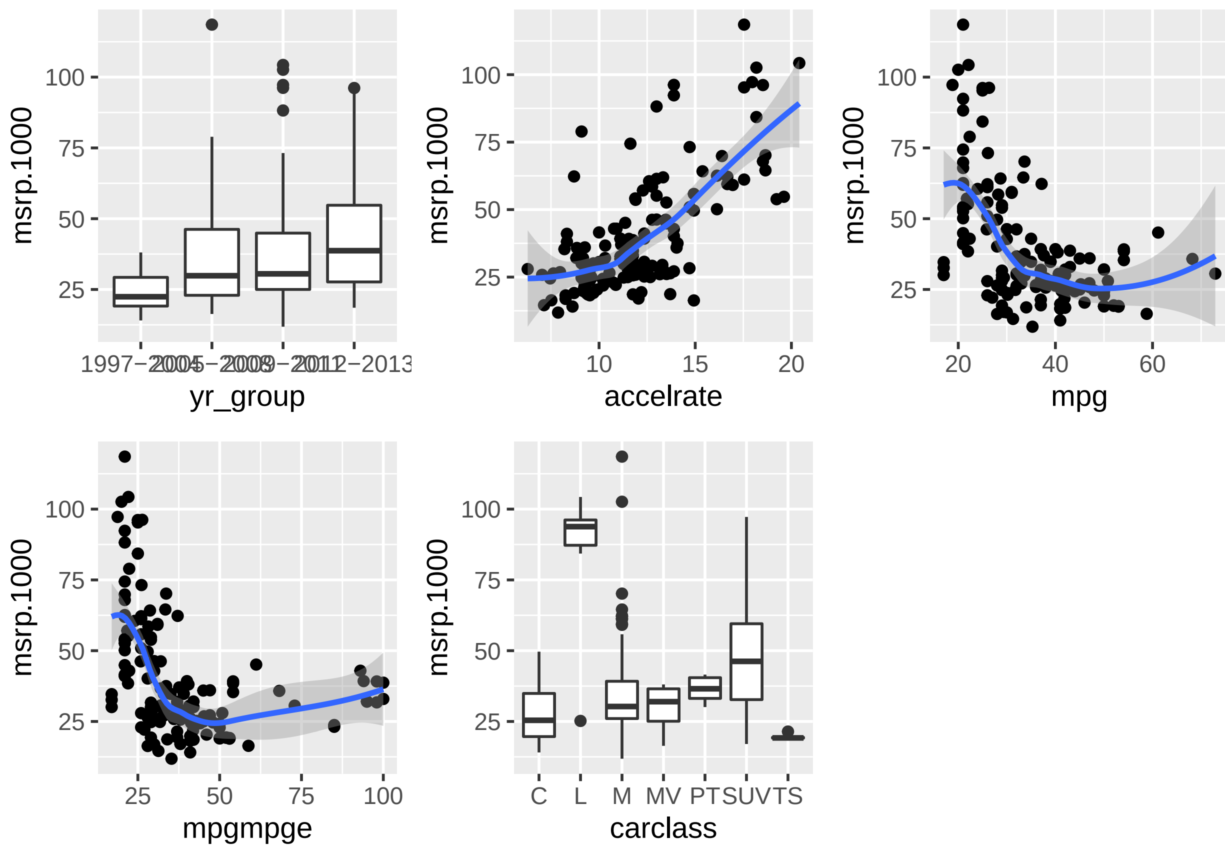
```
##
## 1997-2004 2005-2008 2009-2011 2012-2013       Sum
```

```
##          14         25         57         57        153
```

b. (3 marks)

```r
a <- ggplot(data=hybrid, aes(x=yr_group, y=msrp.1000))+geom_boxplot()+
  labs(x="yr_group", y="msrp.1000")

b <- ggplot(data=hybrid, aes(x=accelrate, y=msrp.1000))+geom_point()+
  geom_smooth(method='loess')+labs(x="accelrate", y="msrp.1000")

c <- ggplot(data=hybrid, aes(x=mpg, y=msrp.1000))+geom_point()+
  geom_smooth(method='loess')+labs(x="mpg", y="msrp.1000")

d <- ggplot(data=hybrid, aes(x=mpgmpge, y=msrp.1000))+geom_point()+
  geom_smooth(method='loess')+labs(x="mpgmpge", y="msrp.1000")

e <- ggplot(data=hybrid, aes(x=carclass, y=msrp.1000))+geom_boxplot()+
  labs(x="carclass", y="msrp.1000")

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

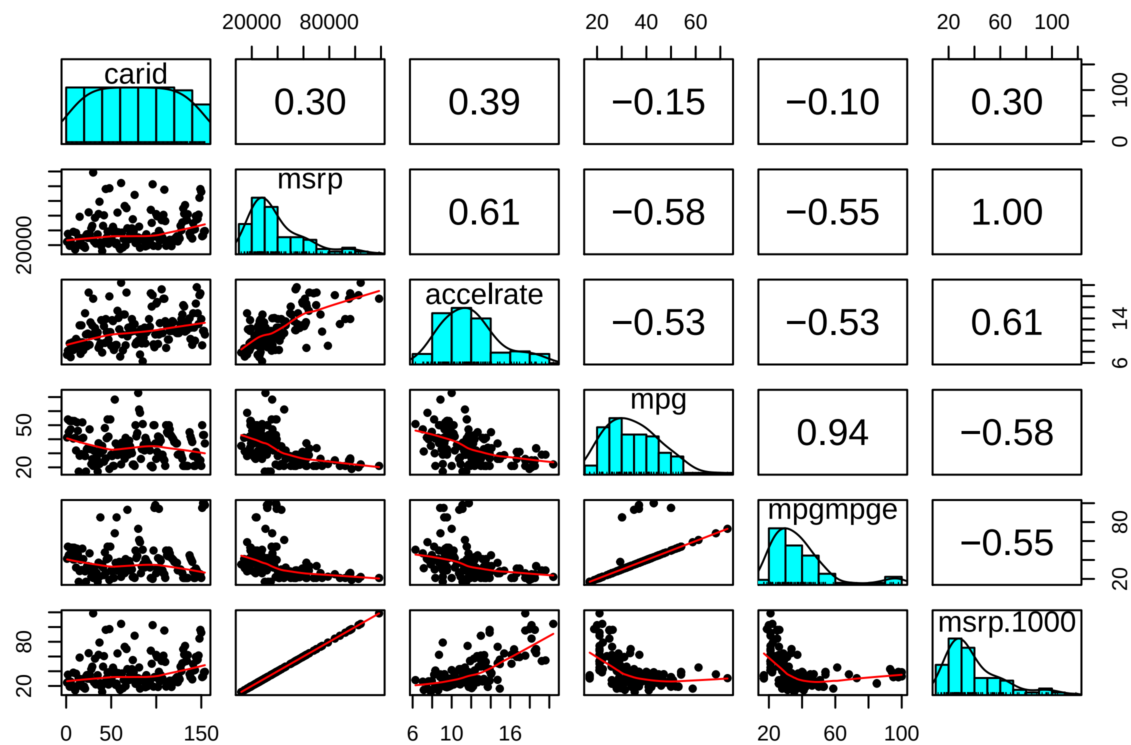
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



From the plots above there's evidence that the relationships between `msrp.1000` + `mpg` and `msrp.1000` + `mpgmpge` are non-linear.

c. (3 marks)

```
hybrid %>%
  select(where(is.numeric)) %>%
  pairs.panels(method = "spearman", density = TRUE, ellipses = FALSE)
```



There's very strong evidence of multicollinearity between `mpg` and `mpgmpge`, the correlation coefficient between them is 0.94 which is incredibly high. This makes sense as
`mpgmpge` is the max of `mge` and `mpge`. There's no strong evidence of multicollinearity between the other predictors.

d. (4 marks)

```
fit1 <- lm(msrp.1000~yr_group+accelrate+mpg+mpgmpge+carclass, data=hybrid)
pander(summary(fit1))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.83	10.6	1.116	0.2664
yr_group2005-2008	-4.145	5.029	-0.8242	0.4112
yr_group2009-2011	-5.174	4.564	-1.134	0.2588
yr_group2012-2013	-6.22	4.833	-1.287	0.2002
accelrate	3.958	0.5088	7.778	1.453e-12
mpg	-0.5168	0.1757	-2.941	0.003825
mpgmpge	0.08512	0.08311	1.024	0.3075
carclassL	27.29	6.221	4.387	2.247e-05
carclassM	-4.049	3.331	-1.215	0.2263
carclassMV	11.61	7.232	1.606	0.1106

	Estimate	Std. Error	t value	Pr(>|t|)
carclassPT	-6.218	7.032	-0.8842	0.3781
carclassSUV	0.9127	4.204	0.2171	0.8284
carclassTS	-8.479	5.771	-1.469	0.144

Table 2: Fitting linear model: msrp.1000 ~ yr_group + accelrate + mpg + mpgmpge + carclass

Observations	Residual Std. Error	R^2	Adjusted R^2
153	13.36	0.6417	0.611

```
pander(vif(fit1), digits=2, caption="VIF values")
```

Table 3: VIF values

	GVIF	Df	GVIF^(1/(2*Df))
yr_group	1.7	3	1.1
accelrate	1.9	1	1.4
mpg	3.2	1	1.8
mpgmpge	2	1	1.4
carclass	3.8	6	1.1

Using the threshold given by: $VIF_{model} = \frac{1}{1-R^2_{model}} = \frac{1}{1-0.6417} = 2.79$. We see that the $GVIF^{(1/2*DF)}$ all predictors are less than $VIF_{model} = 2.79$ which means there's no evidence of severe multicollinearity. It's surprising as `mpg` was identified possible multicollinearity from the pairwise plots.

e. (3 marks)

```
fit.gam <- gam(msrp.1000~yr_group+s(accelrate)+s(mpg)+s(mpgmpge)+carclass,
               data=hybrid, method="REML")
r_squared <- summary(fit.gam)$dev.expl
adj_r_squared <- summary(fit.gam)$r.sq
RSE <- summary(fit.gam)$scale

titles <- c("R-squared", "Adj. R-squared", "RSE")
vals <- c(r_squared, adj_r_squared, RSE)

tabl <- data.frame(titles, vals)
pander(tabl, digits=5, caption="GAM")
```

Table 4: GAM

titles	vals
R-squared	0.77219
Adj. R-squared	0.74139
RSE	118.67

f. (3 marks)

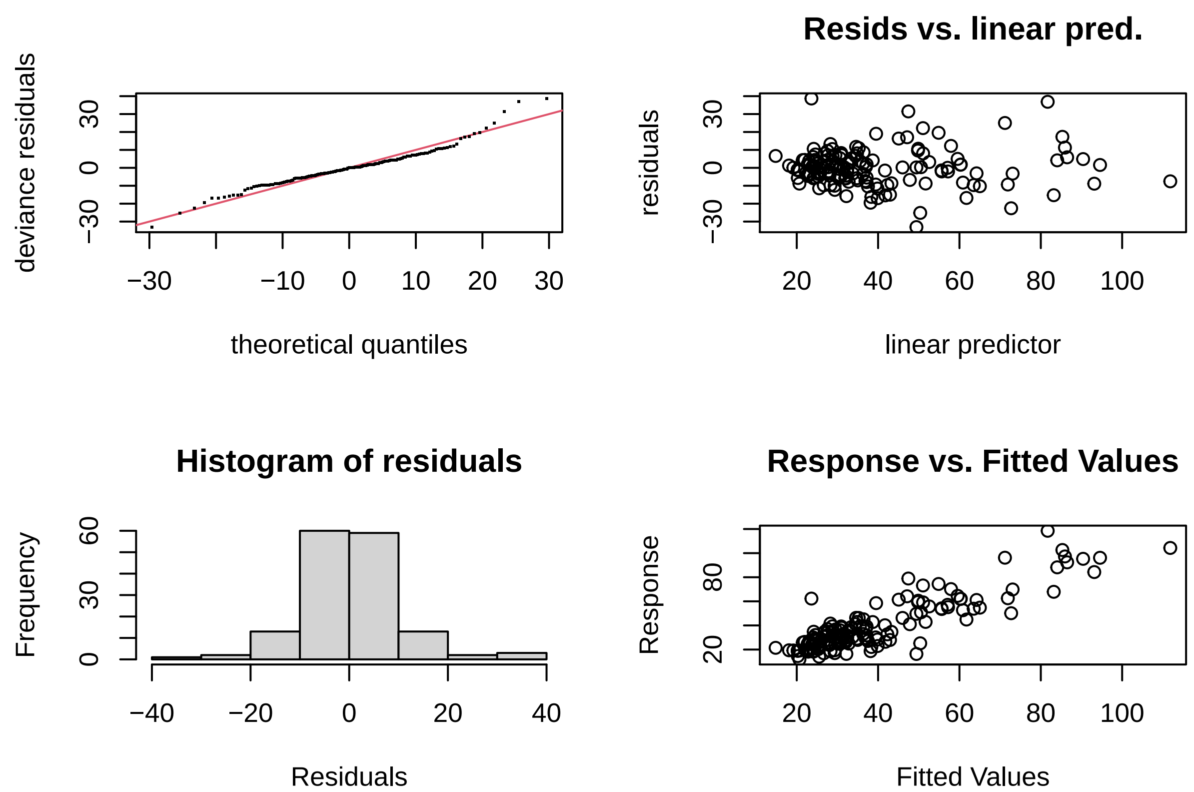
```r
pander(summary(fit.gam)$s.table, digits=3)
```

	edf	Ref.df	F	p-value
s(accelrate)	2.21	2.8	24.5	2.33e-12
s(mpg)	4.95	6.03	2.7	0.0195
s(mpgmpge)	1.95	2.32	1.11	0.361

From the result of the GAM we see that the relationship between `mpg` and `mrsp.1000` is significantly non-linear (high edf [1 is linear, >2 is non-linear] and low p-value). The relationship between `accelrate` and `mrsp.1000` is also significantly non-linear. The relationship between `mpgmpge` and `mrsp.1000` however appears to be insignificant.

g. (4 marks)

```r
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   0.762
## s(mpg)       9.00 4.95    0.81   0.006 **
## s(mpgmpge)   9.00 1.95    0.80   0.011 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Convergence: The report reports full convergence after 5 iterations which indicates that an optimal solution has been found. If no convergence then the results are unreliable.

Basis functions: the p-values for `mpg` and `mpgmpge` are really small indicating that the residuals are not randomly distributed so there might not be enough basis functions. The k-index values for both these predictors are less than 1 further providing evidence that more basis functions might be needed, however the edf is not $edf \approx k'$, it's not near the max amount of basis functions so more basis functions may not be needed.

Q-Q Plot: There's major deviations from the line which suggest non-normality of errors is present.

Residuals vs. Fitted: There's a fanning, as y increases the variance of the residuals increases so there's non-constant variance in the errors.

Histogram of residuals: The histogram of residuals is normally distributed.

Response vs. Fitted: Not a perfect fit but roughly linear.

h. (4 marks)

```
fit.gam2 <- gam(msrp.1000~yr_group+s(accelrate)+s(mpgmpge)+carclass,
               data=hybrid, method="REML")

fit.gam3 <- gam(msrp.1000~yr_group+s(accelrate)+s(mpg)+carclass,
               data=hybrid, method="REML")

fit.gam4 <- gam(msrp.1000~yr_group+s(accelrate)+carclass,
               data=hybrid, method="REML")

aic.gam <- AIC(fit.gam)
aic.gam.2 <- AIC(fit.gam2)
aic.gam.3 <- AIC(fit.gam3)
aic.gam.4 <- AIC(fit.gam4)

modname <- c("All predictors", "-mpg", "-mpgmpge", "-mpg, -mpgmpge")
aicval <- c(aic.gam, aic.gam.2, aic.gam.3, aic.gam.4)

mod.compare <- data.frame(modname, aicval)
names(mod.compare) <- c("Model", "AIC")
pander(mod.compare, digits=3, align='c')
```

Model	AIC
All predictors	1189
-mpg	1190
-mpgmpge	1190
-mpg, -mpgmpge	1239

i. (3 marks)

The model with the lowest AIC is the model with all predictors however the difference between the AIC of the model with all predictors and the model without `mpg` and the one without `mpgmpge` are both less than 2.5 so we apply the rule of parsimony and choose the simpler model, either the model without `mpg` or the one without `mpgmpge` but not excluding both. This points to multicollinearity where both predictors are important but only one is neccesary.

j. (2 marks)

It's surprising since the resulting $GVIFs^{(1/2*DF)}$ didn't give any evidence of severe multicollinearity. However from the AIC model selection we see that both `mpg` and `mpgmpge` together doesn't have a great effect on the response however removing both of them increases the AIC greatly which indicate at least one of them is an important predictor for the response.

k. (4 marks)

```r
bic.gam <- BIC(fit.gam)
bic.gam.2 <- BIC(fit.gam2)
bic.gam.3 <- BIC(fit.gam3)
bic.gam.4 <- BIC(fit.gam4)

modname <- c("All predictors", "-mpg", "-mpgmpge", "-mpg, -mpgmpge")
bicval <- c(bic.gam, bic.gam.2, bic.gam.3, bic.gam.4)

mod.compare <- data.frame(modname, bicval)
names(mod.compare) <- c("Model", "BIC")
pander(mod.compare, digits=3, align='c')
```

Model	BIC
All predictors	1256
-mpg	1249
-mpgmpge	1252
-mpg, -mpgmpge	1285

The model with the lowest BIC is the one that excludes `mpg`, the next lowest BIC model is the one that excludes `mpgmpge` however the difference is greater than 2.0 so we pick the model without `mpg`.

Q2. (5 marks)

a. (1 marks)

$$\widehat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 (X_1 \times X_2) + \hat{\beta}_5 (X_1 \times X_3)$$

$$\widehat{Salary} = \hat{\beta}_0 + \hat{\beta}_1 GPA + \hat{\beta}_2 IQ + \hat{\beta}_3 GenderMale + \hat{\beta}_4 (GPA \times IQ) + \hat{\beta}_5 (GPA \times GenderMale)$$

$$\widehat{Salary} = 5 + 8GPA + 0.2IQ + 10GenderMale + 0.05(GPA \times IQ) + 2(GPA \times GenderMale)$$

$$\widehat{Salary} = \hat{\beta}_0 + \hat{\beta}_2 IQ + \hat{\beta}_3 GenderMale + (\hat{\beta}_1 + \hat{\beta}_4 IQ + \hat{\beta}_5 GenderMale) \times GPA$$

$$\widehat{Salary} = 5 + 0.2IQ + 10GenderMale + (8 + 0.05IQ + 2GenderMale) \times GPA$$

b. (3 marks)

Keeping GPA and IQ the same we're left with: $\hat{\beta}_3 = 10$ and $\hat{\beta}_5 = 2$, $\hat{\beta}_3 = 10$ means that males earn mroe than females and $\hat{\beta}_5 = 2$ means that the interaction between males and GPA means that an increase in both results in an increase of salary so males earn higher than females regardless of GPA, it also means that the higher the GPA the wider the gap between salaries between Males and Females. For IQ $\hat{\beta}_2 = 0.2$ and $\hat{\beta}_4 = 0.05$, an increase in IQ results in an increase of salary.

i. **False.**
ii. **False.**

iii. **True.**
iv. **False.**

c. (1 marks)

False: Just because the interaction coefficient is small does not mean there's little evidence of an interaction effect. Coefficient of the interaction term does not equal statistical significance, statistical significance testing is usually done via a hypothesis test.