DATA 303/473 Assignment 2 Solutions

Q1.

a. **(3 marks)**

```r
hb<-read.csv("hybrid_reg.csv", header=T)
str(hb)
```

```
## 'data.frame':    153 obs. of  9 variables:
##  $ carid      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ vehicle    : chr  "Prius (1st Gen)" "Tino" "Prius (2nd Gen)" "Insight" ...
##  $ 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   : chr  "C" "C" "C" "TS" ...
##  $ carclass_id: int  1 1 1 7 1 7 7 4 7 1 ...
```

```r
library(dplyr)
library(memisc)
hb<-hb%>%
  mutate(yr_group=memisc::recode(year,"1997-2004"<-range(min,2004),
                                "2005-2008"<-2005:2008,
                                "2009-2011"<-2009:2011,
                                "2012-2013"<-2012:2013),
         msrp.1000=msrp/1000)
library(pander)
pander(table(hb$yr_group), caption = "No. of observations in each year group")
```

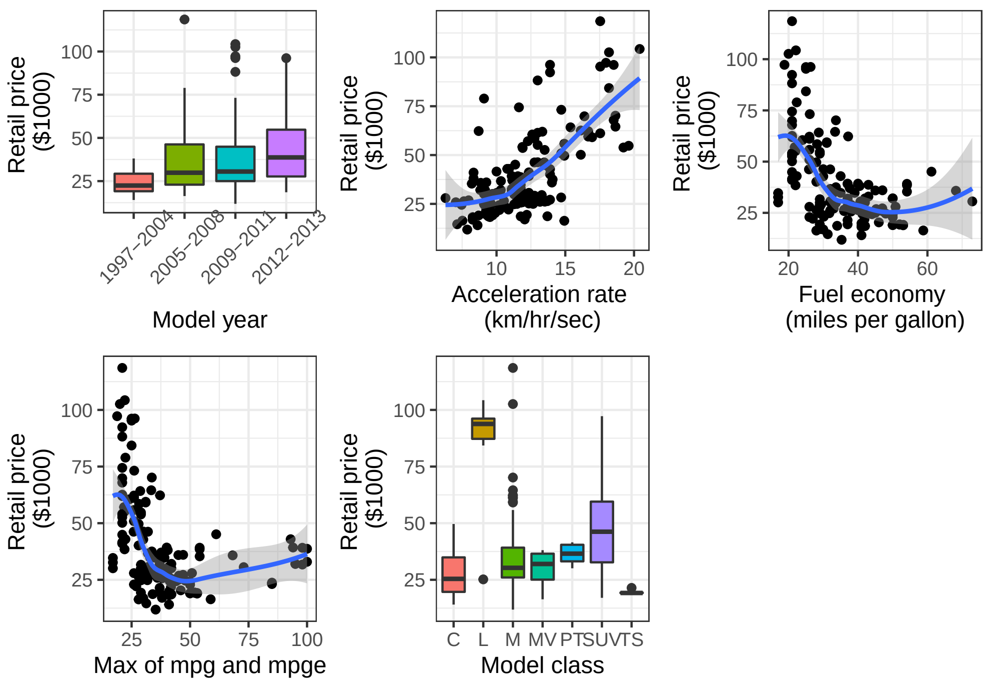
Table 1: No. of observations in each year group

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

b. **(3 marks)**

```r
library(ggplot2)
a<-ggplot(hb,aes(x=yr_group, y=msrp.1000))+
  geom_boxplot(aes(fill=yr_group), show.legend=FALSE) +
  labs(x="Model year", y="Retail price \n($1000)")+
  theme_bw()+
  theme(axis.text.x = element_text(angle = 45, vjust=0.75))
b<-ggplot(hb,aes(x=accelrate, y=msrp.1000))+
  geom_point()+ geom_smooth(method='loess')+
  labs(x="Acceleration rate \n(km/hr/sec)", y="Retail price \n($1000)")+
  theme_bw()
c<-ggplot(hb,aes(x=mpg, y=msrp.1000))+
  geom_point()+ geom_smooth(method='loess')+
```

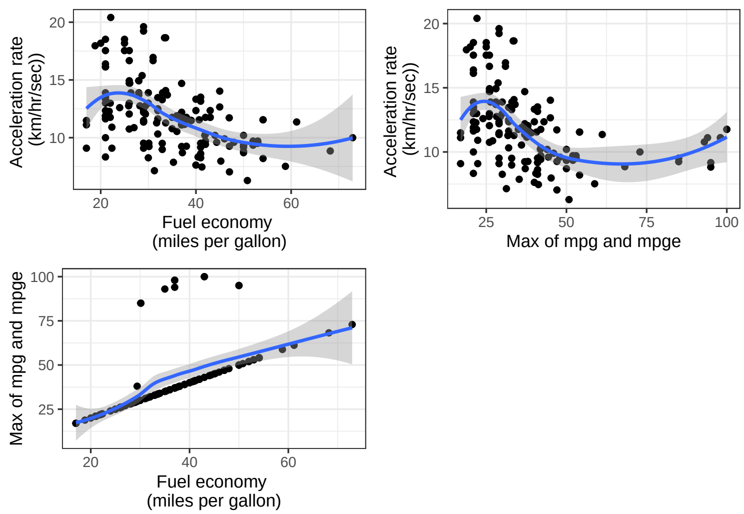
```r
  labs(x="Fuel economy \n(miles per gallon)", y="Retail price \n($1000)")+
  theme_bw()
d<-ggplot(hb,aes(x=mpgmpge, y=msrp.1000))+
  geom_point()+ geom_smooth(method='loess')+
  labs(x="Max of mpg and mpge", y="Retail price \n($1000)")+
  theme_bw()
e<-ggplot(hb,aes(x=carclass, y=msrp.1000))+
  geom_boxplot(aes(fill=carclass), show.legend=FALSE) +
  labs(x="Model class", y="Retail price \n($1000)")+
  theme_bw()
library(gridExtra)
grid.arrange(a,b,c,d,e, nrow=2)
```



Non-linear relationship with `mpg` and `mpgmpgme`.

c. **(3 marks)**

```r
library(ggplot2)
library(gridExtra)
a<-ggplot(hb,aes(x=mpg, y=accelrate))+
  geom_point()+ geom_smooth(method='loess')+
  labs(x="Fuel economy \n(miles per gallon)", y="Acceleration rate \n(km/hr/sec))")+
  theme_bw()
b<-ggplot(hb,aes(x=mpgmpge, y=accelrate))+
  geom_point()+ geom_smooth(method='loess')+
  labs(x="Max of mpg and mpge", y="Acceleration rate \n(km/hr/sec))")+
  theme_bw()
c<-ggplot(hb,aes(x=mpg, y=mpgmpge))+
  geom_point()+ geom_smooth(method='loess')+
  labs(x="Fuel economy \n(miles per gallon)", y="Max of mpg and mpge")+
  theme_bw()
grid.arrange(a,b,c, nrow=2)
```



There is evidence of potential multicollinearity among all pairs of predictors, particularly between `mpg` and `mpgmpge`.

d. **(4 marks)**

```
fit1<-lm(msrp.1000 ~ yr_group + accelrate + mpg + mpgmpge + carclass, data=hb)
library(car)
pander(vif(fit1), caption="VIF values")
```

Table 2: VIF values

	GVIF	Df	GVIF^(1/(2*Df))
yr_group	1.706	3	1.093
accelrate	1.906	1	1.38
mpg	3.164	1	1.779
mpgmpge	1.983	1	1.408
carclass	3.756	6	1.117

```
vif.model<-1/(1-summary(fit1)$r.squared); vif.model
```

```
## [1] 2.790885
```

The $GVIF^{(1/(2\times Df))}$ values are all less than VIF_{model}, therefore no evidence of severe multicollinearity. This is surprising as the plot suggested a strong relationship between `mpg` and `mpgmpge` and I expected at least one of these variables to show severe multicollinearity.

e. **(3 marks)**

```
library(mgcv)
gam1<-gam(msrp.1000 ~ yr_group + s(accelrate) +s(mpg) + s(mpgmpge) + carclass, method="REML", data=hb)
summ.gam<-summary(gam1)
RSE.gam<-sqrt(summ.gam$scale)
Rsq.gam<-summ.gam$dev.expl
AdjRsq.gam<-summ.gam$r.sq
mod.summs<-data.frame(Statistic= c("RSE", "R-squared","Adj. R-squared"),
                      GAM=c(RSE.gam, Rsq.gam, AdjRsq.gam))
pander(mod.summs, caption="Model fit assessment measures")
```

Table 3: Model fit assessment measures

Statistic	GAM
RSE	10.89
R-squared	0.7722
Adj. R-squared	0.7414

f. **(3 marks)**

```
library(pander)
pander(summ.gam$s.table, caption="Summary of smooth terms", keep.trailing.zeros=TRUE)
```

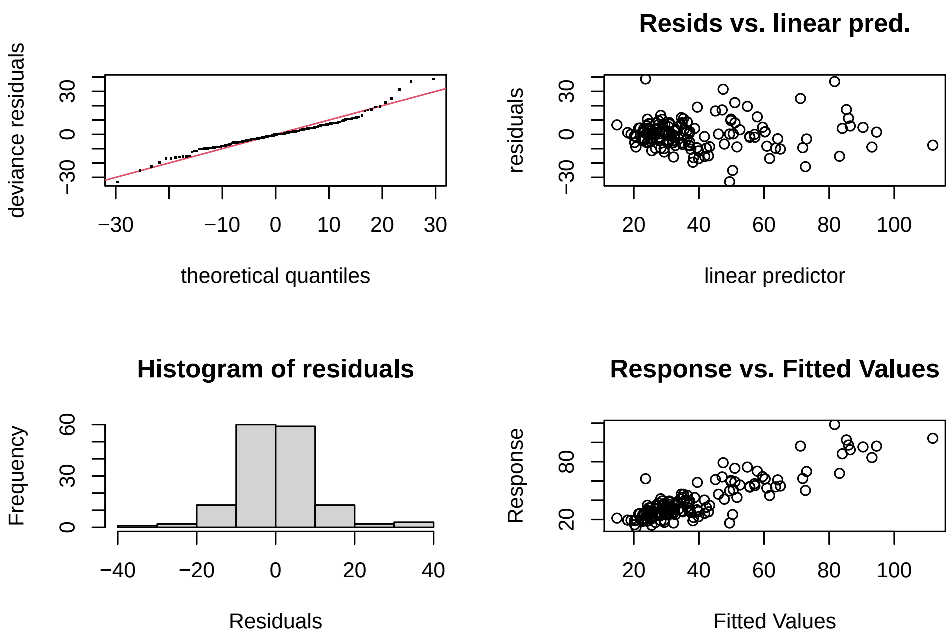
Table 4: Summary of smooth terms

	edf	Ref.df	F	p-value
s(accelrate)	2.209	2.803	24.474	0.00000
s(mpg)	4.946	6.027	2.700	0.01956
s(mpgmpge)	1.950	2.324	1.115	0.36222

The predictors `accelrate` and `mpg` have a significant non-linear effect on `msrp.1000` since they both have `edf`>1 and small p-values. `mpgmpge` has a non-linear, but non-significant effect on `msrp.1000` as `edf`>1 and the p-value is large.

g. **(4 marks)**

```
par(mfrow=c(2,2))
gam.check(gam1)
```




```
##
## 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   0.695
## s(mpg)       9.00 4.95    0.81   0.005 **
## s(mpgmpge)   9.00 1.95    0.80   0.010 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostic plots:

- Histogram and QQ-plot show normality assumption is likely to be met, though there is evidence of a heavy right tail.
- Residual vs linear predictor plot shows no evidence of non-linearity, but does show evidence of non-constant variance.
- Response vs fitted values shows a general linear pattern indicating a well fitting model.

Adequacy of basis functions: The low p-values with k-index <1 are of concern for `mgp` and `mpgmpge`. However, `edf` is much less than `k'` in both cases, giving no evidence of inadequate basis functions.

h. **(4 marks)**

```
gam2<-gam(msrp.1000~yr_group  + carclass +
            s(accelrate)+s(mpgmpge), method="REML", data=hb)
gam3<-gam(msrp.1000~yr_group  + carclass +
            s(accelrate)+s(mpg), method="REML", data=hb)
gam4<-gam(msrp.1000~yr_group  + carclass +
            s(accelrate), method="REML", data=hb)
aicvals<-c(AIC(gam1), AIC(gam2), AIC(gam3), AIC(gam4))
mod.summs<-data.frame(model=c("All predictors", "-mpg", "-mpgmpge", "-mpg,-mpgmpge"),
  aicvals)
colnames(mod.summs)=c("Model","AIC")
pander(mod.summs, caption="AIC values")
```

Table 5: AIC values

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

i. **(3 marks)**

- The AIC statistic indicates that the first three models are all equivalent (difference between any pair is less than 2.5). This suggests that a model that excludes either one of `mpg` or `mpgmpge` has equivalent fit to a model that includes both predictors.
- Excluding both predictors results in a model with a large increase in AIC, indicating significantly poorer fit. This indicates that either of these two predictors can be included.
- This points to a collinearity issue between `mpg` ānd `mpgmpge`.

j. **(2 marks)** Yes I am surprised by the indication of multi-collinearity in part (i), given that the VIF statistic in part (d) suggested there was no severe multicollinearity.

NOTE: This is an example of a situation where the VIF statistic failed to detect multicollinearity that is present, and illustrates the need to use various approaches to investigate regression pitfalls.

k. **(3 marks)**

```
bicvals<-c(BIC(gam1), BIC(gam2), BIC(gam3), BIC(gam4))
mod.summs<-data.frame(model=c("All predictors", "-mpg", "-mpgmpge", "-mpg,-mpgmpge"),
  bicvals)
colnames(mod.summs)=c("Model","BIC")
pander(mod.summs, caption="BIC values")
```

Table 6: BIC values

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

- My preferred model is Model 2 (excludes `mpg` but includes `mpgmpge`).
- The model has the lowest BIC value, and compared to Model 3 (excludes `mpgmpge` but includes `mpg`) which has the next highest BIC value, the difference in BIC is 6, which indicates strong preference for Model 2.

2. **Q2. (5 marks)**

a. **(1 mark)**

$$\hat{Y} = 5 + 8X_1 + 0.2X_2 + 10X_3 + 0.05X_1X_2 + 2X_1X_3$$

b. **(3 marks)**

The model equation can be re-written as:

$$\begin{aligned}
\hat{Y} &= 5 + 8X_1 + 0.2X_2 + 10X_3 + 0.05X_1X_2 + 2X_1X_3 \\
&= 5 + 8GPA + 0.2IQ + 10Gender_{male} + 0.05GPA \times IQ + 2GPA \times Gender_{male} \\
&= 5 + \underbrace{(10 + 2GPA)}_{\hat{\beta}_{Gender_{male}}} Gender_{male} + \underbrace{(0.2 + 0.05GPA)}_{\hat{\beta}_{IQ}} IQ + 8GPA
\end{aligned}$$

This suggests:

(i) FALSE. $\hat{\beta}_{Gender_{male}}$ is positive for all GPA values, therefore expected starting salary for males is higher than that for females for all GPA values.
(ii) FALSE. For the reason given in (i), expected starting salary for females is lower than for males for all GPA values.
(iii) TRUE. For each additional GPA point, the difference in expected starting salary between males and females increases by $2000.
(iv) FALSE. $\hat{\beta}_{IQ}$ is positive for all GPA values, therefore an increase in IQ by one point is associated with an increase in expected starting salary for all GPA values.

c. **(1 mark)** False. We can't know that without carrying out a hypothesis test.

Assignment total: 40 marks