Assignment 2

```r
file <- "Ass2.txt"
D <- read.table(file, header = TRUE)
```

Question 1

(a) Compute and report the least-squares estimates of the vector β using MPG as reponse variable and engine size, weight and horse power as explanatory variables. Write down the estimated regression equation.

```r
n<-11
p<-3
y <- D$MPG
x1 <- D$Engine
x2 <- D$HP
x3 <-D$Weight
xvals <- c(x1, x2, x3)
(X<-matrix(c(rep(1,n),xvals),nrow=n,ncol=p+1))
```

```
##       [,1] [,2] [,3] [,4]
##  [1,]    1 3471  260 4420
##  [2,]    1 2979  225 4586
##  [3,]    1 4195  275 4787
##  [4,]    1 4701  235 4379
##  [5,]    1 3471  240 4439
##  [6,]    1 3960  195 3786
##  [7,]    1 4701  235 3786
##  [8,]    1 4701  265 3786
##  [9,]    1 3311  230 3860
## [10,]    1 4664  235 5390
## [11,]    1 4605  302 4834
```

```r
BetaHat <- solve(t(X)%*%X)%*%t(X)%*%y
yhat <- X%*%BetaHat
print(BetaHat)
```

```
##              [,1]
## [1,] 35.180503587
## [2,] -0.002567547
## [3,]  0.015388888
## [4,] -0.001843143
```

The estimated regression equation is
$\hat{Y}$ = 35.180503587 + (-0.002567547) x_1 + 0.015388888 x_2 + (-0.001843143)x_3

(b) Explain in context what the coefficient corresponding to horsepower means.

Increasing horsepower by 1 pound, the gas mileage increases by 0.015388888 mpg. 3/5

(c) Using the response variable(y), and the fitted value $\hat{y}$, compute the biased and unbiased estimates of the error variance σ^2.

```
ehat <- y - yhat
RSS <- t(ehat)%*%ehat
biased <- RSS/n
unbiased <- RSS/(n-p-1)
biased
```

```
##           [,1]
## [1,] 0.5180943
```

```
unbiased
```

```
##           [,1]
## [1,] 0.8141483
```

Thus, the biased estimate of error variance is 0.5180943 and the unbiased estimate of the error variance is 0.8141483.

0/10

(d) Compute the variance-covariance matrix of the estimated regression coefficients. Derive estimates of the variances and the covariance of the estimators of the regression coefficients associated with predictors engine size and horsepower?

```
cov <- solve(t(X)%*%X)
cov
```

s2<-SSR/(n-p-1)
Inv_tXX<-solve(t(X)%*%X)
varcov<-s2*Inv_tXX

```
##               [,1]          [,2]          [,3]          [,4]
## [1,] 11.8942309998 -6.060494e-04 -1.747300e-02 -1.156758e-03
## [2,] -0.0006060494  2.571567e-07 -1.977181e-06  1.017469e-08
## [3,] -0.0174729998 -1.977181e-06  1.560537e-04 -2.917137e-06
## [4,] -0.0011567581  1.017469e-08 -2.917137e-06  4.190466e-07
```

```
cov[1,1]
```

```
## [1] 11.89423
```

```
cov[2,2]
```

```
## [1] 2.571567e-07
```

```
cov[3,3]
```

```
## [1] 0.0001560537
```

```
cov[4,4]
```

```
## [1] 4.190466e-07
```

```
cov[2,3]
```

```
## [1] -1.977181e-06
```

$\text{Var}(\hat{\beta}_0)$ = 11.89423
$\text{Var}(\hat{\beta}_1)$ = 2.571567e-07
$\text{Var}(\hat{\beta}_2)$ = 1.560537e-04
$\text{Var}(\hat{\beta}_3)$ = 4.190466e-07
$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ = -1.977181e-06

Question 2

4/5

(a) Conduct the F-test for the overall fit of the regression. Comment on the results.

We want to test H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ against H_a: at least one of $\beta_i \neq 0$, where i = 1, 2, 3

```
fit <- lm(y~x1+x2+x3, data = D)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = D)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54163 -0.06518  0.18154  0.29778  0.89573
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.1805036  3.1118592  11.305 9.48e-06 ***
## x1          -0.0025675  0.0004576  -5.611 0.000806 ***
## x2           0.0153889  0.0112717   1.365 0.214421
## x3          -0.0018431  0.0005841  -3.156 0.016027 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9023 on 7 degrees of freedom
## Multiple R-squared:  0.8601, Adjusted R-squared:  0.8001
## F-statistic: 14.34 on 3 and 7 DF,  p-value: 0.002253
```

The output shows that F = 14.34 (p-value = 0.002253), indicating that we should reject the null hypothesis that the variables Engine, HP and Weight collectively have no effect on MPG. The results also show that Engine and Weight are significant, but HP is nonsignificant. In addition, the output also shows that $R^2 = 0.8601$ and $R^2_{adjusted} = 0.8001$.

(b) Test each of the individual regression coefficients. Do the results indicate that any of the explanatory variables should be removed from the model?

Here we use individual T-test.
1. Test: $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$

```
coef <- summary(fit)$coef
print(coef)
```

```
##                 Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) 35.180503587 3.1118591642 11.305301 9.477994e-06
## x1          -0.002567547 0.0004575627 -5.611354 8.063205e-04
## x2           0.015388888 0.0112716834  1.365270 2.144215e-01
## x3          -0.001843143 0.0005840943 -3.155558 1.602730e-02
```

```
beta1hat<-coef[2,1]
sebeta1hat<-coef[2,2]
t <-beta1hat/sebeta1hat
print(t)
```

```
## [1] -5.611354
```

```
alpha<-0.05
t0<-qt(1-alpha/2,n-p-1)
print(t0)
```

```
## [1] 2.364624
```

```r
pval1<-coef[2,4]
print(pval1)
```

```
## [1] 0.0008063205
```

So, T-statistic: $|T_1| = |\frac{\hat{\beta_1}}{se(\hat{\beta_1})}| = 5.611354 > t(1-\alpha/2, n-1) = 2.364624$
We reject H_0 and conclude that β_1 is significantly different from 0.
The p-value of the test is equal to 0.0008063205, which leads to the same conclusion.

2. Test: $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$

```r
beta2hat<-coef[3,1]
sebeta2hat<-coef[3,2]
t2 <-beta2hat/sebeta2hat
print(t2)
```

```
## [1] 1.36527
```

```r
pval2<-coef[3,4]
print(pval2)
```

```
## [1] 0.2144215
```

So, T-statistic: $|T_2| = |\frac{\hat{\beta_2}}{se(\hat{\beta_2})}| = 1.36527 < t(1-\alpha/2, n-1) = 2.364624$
We do not reject H_0 and conclude that β_2 is not significantly different from 0.
The p-value of the test is equal to 0.2144215, which leads to the same conclusion.

3. Test: $H_0 : \beta_3 = 0$ against $H_a : \beta_3 \neq 0$

```r
beta3hat<-coef[4,1]
sebeta3hat<-coef[4,2]
t3 <-beta3hat/sebeta3hat
print(t3)
```

```
## [1] -3.155558
```

```r
pval3<-coef[4,4]
print(pval3)
```

```
## [1] 0.0160273
```

So, T-statistic: $|T_3| = |\frac{\hat{\beta_3}}{se(\hat{\beta_3})}| = 3.155558 > t(1-\alpha/2, n-1) = 2.364624$
We reject H_0 and conclude that β_3 is significantly different from 0.
The p-value of the test is equal to 0.0160273, which leads to the same conclusion.

Therefore, the result indicates that the variable HP should be removed from the model because it fails to reject the null hypothesis. Observing the highest p-value and weakest evidence against the null hypothesis, we can conclude the coefficient of HP is not significantly different from 0.

0/5 **(c) Determine the regression model with the explanatory variable(s) identified in part (b) removed. Write down the estimated regression equation.**

New Regression Model: $E(Y_i|X = x_i) = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3}$
The estimated regression equation is $\hat{Y}$ = 35.180503587 + (-0.002567547) x_1 + (-0.001843143)x_3

fit2<- lm(MPG ~ Engine + Weight, data=mydata)
coef2<-coef(summary(fit2))
coef2

(d) Going back to the original model containing all three explanatory variables, construct a 99% confidence interval for the mean gas mileage for SUVs with Engine = 2000, HP = 250 and Weight = 4000

```
predict(fit, data.frame(x1=2000, x2=250, x3=4000), interval="confidence", level=.99)
```

```
##        fit      lwr      upr
## 1 26.52006 22.89565 30.14447
```

Thus, the confidence interval for mean MPG with Engine=2000, HP=250 and Weight=4000 is (22.89565, 30.14447).

(e) Construct a 99% prediction interval for the mileage of a particular SUV with Engine= 2000, HP = 250 and Weight = 4000.

```
predict(fit, data.frame(x1=2000, x2=250, x3=4000), interval="prediction", level=.99)
```

```
##        fit      lwr    upr
## 1 26.52006 21.71312 31.327
```

Thus, the prediction interval for y at $x_1 = 2000$, $x_2 = 250$ and $x_3 = 4000$ is (21.71312, 31.327).

(f) Now, we are interested in testing whether Horsepower and Weight are significant after taking Engine size into consideration. (i) Compute the residual sum of squares(RSS) of each of the above model

1. $E(Y_i|X = x_i) = \beta_0 + \beta_1 x_{i1}$

```
fit2 <- lm(y~x1)
sum(fit2$residuals^2)
```

```
## [1] 13.85416
```

The residual sum of squares in above model is 13.85416.

2. $E(Y_i|X = x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$

```
sum(fit$residuals^2)
```

```
## [1] 5.699038
```

The residual sum of squares in above model is 5.699038.

(ii) Compute the F test statistic for comparing these two models

```
anova(fit2, fit)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1
## Model 2: y ~ x1 + x2 + x3
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1      9 13.854
## 2      7  5.699  2    8.1551 5.0084 0.04465 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows the F test statistic for comparing these two models is 5.0084.

Since the pvalue, 0.0446452 is less than the significan level – = 0.05, we reject the null hypothesis.

It appears that the variables HP and Weight contribute significant information to the MPG once the variables Engine have been taken into consideration.

0/5 **(iii) At the 5% level of significant, what conclusions can you draw?**

Since the p-value is $0.04465 < 0.05$, we cannot reject the null hypothesis ($\beta_2 = \beta_3 = 0$). It appears that the variables HP and Weight do not contribute significant information to MPG once the variable Engine have been taken into consideration.

(iv) Compare the fit of these two models on the basis of R^2 and of $R^2_{adjusted}$. Comment on your result.

```r
summary(fit2)$r.squared
```

```
## [1] 0.6598309
```

```r
summary(fit)$r.squared
```

```
## [1] 0.8600683
```

R^2 of the reduced model is smaller than R^2 of the full model. However, adding irrelevant predictor variables to the regression equation often increases R^2. So not many meaningful conclusions can be drawn here when comparing the two models.

```r
summary(fit2)$adj.r.squared
```

```
## [1] 0.6220343
```

```r
summary(fit)$adj.r.squared
```

```
## [1] 0.8000975
```

About 62% of the variablility in the MPG can be explained by the reduced model, whereas approximately 80% of the variability in the MPG can be explained by the full model. The adjusted correlation of coefficient of the reduced model is smaller than the full model. This means the full model which contains all three predictors has more explanatory power of regression model. More proportions of the total sample variablity in the Y's can be explained by the full model. There is an improvement in the fit by adding this two predictors.