HW 1.

Q1. $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}$

(a) $x^T A x$

$$= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \begin{bmatrix} x_1 a_{11} + x_2 a_{12} & x_1 a_{12} + x_2 a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \begin{bmatrix} x_1 (x_1 a_{11} + x_2 a_{12}) + x_2 (x_1 a_{12} + x_2 a_{22}) \end{bmatrix}$$

$$= \begin{bmatrix} x_1^2 a_{11} + x_1 x_2 a_{12} + x_1 x_2 a_{12} + x_2^2 a_{22} \end{bmatrix}$$

$$= \begin{bmatrix} x_1^2 a_{11} + 2 x_1 x_2 a_{12} + x_2^2 a_{22} \end{bmatrix}$$

(b) $$\frac{\partial x^T A x}{\partial x} = \begin{bmatrix} \dfrac{\partial x^T A x}{\partial x_1} \\ \dfrac{\partial x^T A x}{\partial x_2} \end{bmatrix}$$

$$= \begin{bmatrix} 2 a_{11} x_1 + 2 x_2 a_{12} \\ 2 x_1 a_{12} + 2 x_2 a_{22} \end{bmatrix}$$

$$= 2 \begin{bmatrix} a_{11} x_1 + x_2 a_{12} \\ x_1 a_{12} + x_2 a_{22} \end{bmatrix}$$

$$\because A \cdot X = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{12}x_1 + a_{22}x_2 \end{bmatrix}$$

$$\therefore \frac{\partial X^T A X}{\partial X} = 2 \cdot A \cdot X$$

Q2. (a) $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$

(b) $$X^T X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

(c) Yes, because $X^T X_{1,2} = X^T X_{2,1}$

(d) now $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & \cdots & x_{2p} \\ \vdots & \vdots & & \ddots & \\ 1 & x_{n1} & & \cdots & x_{np} \end{bmatrix}$ $n \times (p+1)$

$X^T = \begin{bmatrix} 1 & 1 & \cdots & \cdots & 1 \\ x_{11} & x_{21} & \cdots & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & & \\ x_{1p} & x_{2p} & & \cdots & x_{np} \end{bmatrix}$ $(p+1) \times n$

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & & \ddots & \\ x_{1p} & x_{2p} & & x_{np} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & & x_{2p} \\ \vdots & \vdots & & \ddots & \\ 1 & x_{n1} & & & x_{np} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \sum_i^n x_{i1} & \cdots & \sum_i^n x_{ip} \\ \sum_i^n x_{i1} & \ddots & & \sum_i^n x_{i1} \cdot x_{ip} \\ \vdots & & \ddots & \\ \sum_i^n x_{ip} & & & \sum_i^n (x_{ip})^2 \end{bmatrix} \quad (p+1)(1+p)$$

Let $A = X^T X$, $A_{ij} = A_{ji}$

Problem 3

(a)

If $\beta=0$, $y_i=\alpha+e_i$. It means that there's no association between y and x. The regression function will be plotted as a horizontal straight line which is $y_i=\alpha$

(b) $y_i=\alpha+e_i$

$\hat{y}_i=\hat{\alpha}$

$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\hat{\alpha})^2=\sum_{i=1}^{n}y_i^2-2\hat{\alpha}y_i+\hat{\alpha}^2$

$\frac{\partial SSE}{\partial \alpha}=\sum_{i=1}^{n}-2y_i+2\hat{\alpha}=0$

$\sum_{i=1}^{n}(-y_i+\hat{\alpha})=0.$

$n\hat{\alpha}=\sum_{i=1}^{n}y_i$

$\hat{\alpha}=\frac{1}{n}\sum_{i=1}^{n}y_i$

$\hat{\alpha}=\bar{y}$

(c)

To prove $\hat{\alpha}$ is an unbiased estimator of α, I'll show $E(\hat{\alpha})=\alpha$.

$y_i=\alpha+e_i$ $\quad e_i\sim NID(0,\sigma^2)$

$E(y_i)=E(\alpha+e_i)=\alpha+0=\alpha$

$Var(y_i)=Var(\alpha+e_i)=Var(e_i)=\sigma^2$

So y_i's have $N(\alpha,\sigma^2)$ distribution and are independent.

$y_i\sim NID(\alpha,\sigma^2)$

$E(\hat{\alpha})=E(\frac{1}{n}\sum_{i=1}^{n}y_i)$ ↙ y_i's are independent

$=\frac{1}{n}\sum_{i=1}^{n}E(y_i)$

$=\frac{1}{n}\cdot n\cdot\alpha$

$\therefore E(\hat{\alpha}) = \alpha$

$\therefore$ $\hat{\alpha}$ is an unbiased estimator of α.

(d)

$\hat{\alpha} = \bar{y}$

$$Var(\hat{\alpha}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

$$= \frac{1}{n^2} Var(y_1 + y_2 + \cdots\cdots + y_n)$$ y_i's are independent.

$$= \frac{1}{n^2}\left[Var(y_1) + Var(y_2) + \cdots\cdots + Var(y_n)\right]$$

$$= \frac{1}{n^2} \cdot n \cdot \sigma^2$$

$$= \frac{\sigma^2}{n}$$

$$\therefore Var(\hat{\alpha}) = \frac{\sigma^2}{n}.$$

(e)

$\hat{\alpha} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

In question (c), I showed $y_i \sim NID(\alpha, \sigma^2)$, i.e. y_i's are independent and have $N(\alpha, \sigma^2)$ distributions.

$\hat{\alpha}$ is the summation of n $y_i \sim NID(\alpha, \sigma^2)$ divided by n

Since the summation of independent random variables with normal distribution also has normal distribution, $\sum_{i=1}^{n} y_i \sim N(n\alpha, n\sigma^2)$.

The division of n doesn't influence its distribution.

So, $\hat{\alpha}$ has normal distribution.

(Note, even if y_i's don't have normal distribution, by Central Limit theorem, when n is large enough, $\hat{\alpha}$ will still have normal distribution.)

1f).

Suppose $\hat{\alpha} = \sum_{i=1}^{n} c_i y_i$. To make $\hat{\alpha}$ unbiased, $E(\hat{\alpha}) = \alpha$

$$E(\hat{\alpha}) = E\left(\sum_{i=1}^{n} c_i y_i\right)$$

y_i's are independent.

$$= \sum_{i=1}^{n} c_i E(y_i)$$

$$= \alpha \sum_{i=1}^{n} c_i$$

This implies $\sum_{i=1}^{n} c_i = 1$

$$c_i = d_i + \frac{1}{n}$$

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \left(d_i + \frac{1}{n}\right)$$

$$= d_1 + \frac{1}{n} + d_2 + \frac{1}{n} + \cdots\cdots + d_n + \frac{1}{n}$$

$$= 1 + \sum_{i=1}^{n} d_i$$

$\therefore \sum_{i=1}^{n} c_i = 1 \rightarrow 1 + \sum_{i=1}^{n} d_i = 1 \rightarrow \sum_{i=1}^{n} d_i = 0 \rightarrow \sum_{i=1}^{n} d_i / n = 0.$

$$Var(\hat{\alpha}) = Var\left(\sum_{i=1}^{n} c_i y_i\right)$$

y_i's are independent

$$= \sum_{i=1}^{n} Var(c_i y_i)$$

$$= c_1^2 Var(y_1) + c_2^2 Var(y_2) + \cdots\cdots + c_n^2 Var(y_n)$$

$$= \sigma^2 (c_1^2 + \cdots\cdots + c_n^2)$$

$$= \sigma^2 \left[\left(d_1 + \frac{1}{n}\right)^2 + \left(d_2 + \frac{1}{n}\right)^2 + \cdots\cdots + \left(d_n + \frac{1}{n}\right)^2\right]$$

$$\min Var(\hat{\alpha}) \Leftrightarrow \min \sigma^2 \left[\left(d_1 + \frac{1}{n}\right)^2 + \left(d_2 + \frac{1}{n}\right)^2 + \cdots\cdots + \left(d_n + \frac{1}{n}\right)^2\right]$$

$$\Leftrightarrow \min d_1^2 + \frac{2d_1}{n} + \frac{1}{n^2} + d_2^2 + \frac{2d_2}{n} + \frac{1}{n^2} + \cdots\cdots + d_n^2 + \frac{2d_n}{n} + \frac{1}{n^2}$$

$$\Leftrightarrow \min d_1^2 + d_2^2 + \cdots\cdots + d_n^2 + \frac{2}{n}(d_1 + d_2 + \cdots\cdots + d_n) + n \cdot \frac{1}{n^2}$$

$\because \sum d_i = 0.$

$$\Leftrightarrow \min d_1^2 + \cdots\cdots + d_n^2 + \frac{1}{n}.$$

The final optimization problem is:

$$\min \quad M = d_1^2 + \cdots\cdots + d_n^2 + \frac{1}{n}.$$

$$\text{s.t.} \quad \sum_{i=1}^{n} d_i = 0.$$

$$\frac{\partial M}{\partial d_i} = 2d_i = 0.$$

$$d_i = 0. \quad \text{for } i=1,\dots\dots,n$$

Therefore, to fuufill 2 conditions: unbiased and lowest variance, $d_i = 0$ $\forall i=1,\dots\dots,n$. $\hat{\alpha} = \bar{y}$ is the "BLUE".

Problem 4

L1.) $\overset{\overset{\text{cov}}{\downarrow}}{C}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$

$\Rightarrow C(aX, bY) = E[a(X - \mu_x)\, b(Y - \mu_y)]$

$= E[abXY - abX\overset{\mu}{}_y - abY\mu_x + ab\,\mu_x\mu_y]$

$= ab\,E[(X - \mu_x)(Y - \mu_y)] = ab\,C(X, Y)$

L2.) $C(X+Y, Z) = E[(X+Y - \mu_{X+Y})(Z - \mu_z)]$

$= E[(X + Y - \mu_x - \mu_Y)(Z - \mu_z)]$

$= E[XZ - X\mu_z + YZ - Y\mu_z - \mu_x Z + \mu_x\mu_z - \mu_y Z + \mu_x\mu_z]$

$= E[(XZ - X\mu_z - \mu_x Z + \mu_x\mu_z)(YZ - Y\mu_z - \mu_y Z + \mu_y\mu_z)]$

$= E[X(Z - \mu_z) - \mu_x(Z - \mu_z) + Y(Z - \mu_z) - \mu_y(Z - \mu_z)]$

$= E[(X - \mu_x)(Z - \mu_z) + (Y - \mu_y)(Z - \mu_z)]$

$= E[(X - \mu_x)(Z - \mu_z)] + E[(Y - \mu_y)(Z - \mu_z)]$

$= C(X, Z) + C(Y, Z)$

First, simplify $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i - \sum_{i=1}^{n} (x_i - \bar{x})\bar{y}}{S_{xx}}$$

$$= \frac{1}{S_{xx}} \left(\sum_{i=1}^{n} (x_i - \bar{x}) y_i - \bar{y} \sum_{i=1}^{n} x_i + n\bar{y}\bar{x} \right)$$

$$= \frac{1}{S_{xx}} \left(\sum_{i=1}^{n} (x_i - \bar{x}) y_i - n\bar{y}\bar{x} + n\bar{y}\bar{x} \right)$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{S_{xx}}$$

$$\mathrm{Cov}(\bar{y}, \hat{\beta}_1) = \mathrm{Cov}\left(\bar{y} , \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{S_{xx}} \right)$$

$$\big\downarrow \text{ lemma 1}$$

$$= \frac{1}{S_{xx}} \mathrm{Cov}\left(\bar{y}, \sum_{i=1}^{n} (x_i - \bar{x}) y_i\right)$$

$$\big\downarrow \text{ lemma 2}$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} \mathrm{Cov}(\bar{y}, (x_i - \bar{x}) y_i)$$

$$\big\downarrow \text{ lemma 1}$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \mathrm{Cov}(\bar{y}, y_i)$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \mathrm{Cov}\left(\frac{1}{n} \sum_{j=1}^{n} y_j , y_i\right)$$

$$\big\downarrow \text{ lemma 1 \& 2}$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot \frac{1}{n} \sum_{j=1}^{n} \mathrm{Cov}(y_j, y_i)$$

$$\big\downarrow \; y_i \perp\!\!\!\perp y_j \; \forall i \neq j$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot \frac{1}{n} \cdot \mathrm{Cov}(y_i, y_i)$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot \frac{1}{n} \mathrm{Var}(y_i)$$

$$= \frac{1}{S_{xx}} \cdot \frac{1}{n} \cdot \sigma^2 \cdot \sum_{i=1}^{n} (x_i - \bar{x})$$

$$= \frac{1}{S_{xx}} \cdot \frac{1}{n} \cdot \sigma^2 \cdot \left(\sum_{i=1}^{n} x_i - n\bar{X} \right)$$

$$= \frac{1}{S_{xx}} \cdot \frac{1}{n} \cdot \sigma^2 \cdot 0$$

$$= 0$$

Since $\mathrm{Cov}(\bar{y}, \hat{\beta}_1) = 0$, $\bar{y} \perp\!\!\!\perp \hat{\beta}_1$.

HW 01

Group 10

2022-09-30

Question 5

```r
## load the data
setwd("/Users/nuke2/Desktop/NW Work/MSiA 401/HW 01/HW 01 Parts")
auto <- read.csv("Auto.csv", na.strings = "?")
auto = na.omit(auto)
```

5(a) The estimated regression equation is $mpg = -0.158 * horsepower + 39.936$.

```r
simple_reg <- lm(mpg~horsepower, data = auto)
summary(simple_reg)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

5(b) The slope indicates a negative relationship between horsepower and mpg. As you increase horsepower the mpg drops.

5(d) To test the hypotheses (with Type I error rate 0.5) $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ P-value $= 2.2e^{-16} < 0.05$, so reject H_0 and conclude $\beta_1 \neq 0$. There is a significant relationship between mpg and horsepower.

5(e) The R-squared score is 0.606, meaning 60.6% of variation is explained by the model.The relationship is relatively strong.

5(f) The residual standard error is the standard error of estimate. It is used to measure how well a regression model fits the data set.

5(g) The predicted mpg of a car with 98 horsepower is 24.467.

```
predict(simple_reg, data.frame(horsepower = 98))
```

```
##        1
## 24.46708
```

5(h) The 95% PI is (14.809, 34.125).

```
predict(simple_reg, data.frame(horsepower = 98), interval = "prediction", level = 0.95 )
```

```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

5(i) The 99% CI is (23.817, 25.117).

```
predict(simple_reg, data.frame(horsepower = 98), interval = "confidence", level = .99)
```

```
##        fit      lwr      upr
## 1 24.46708 23.81669 25.11747
```

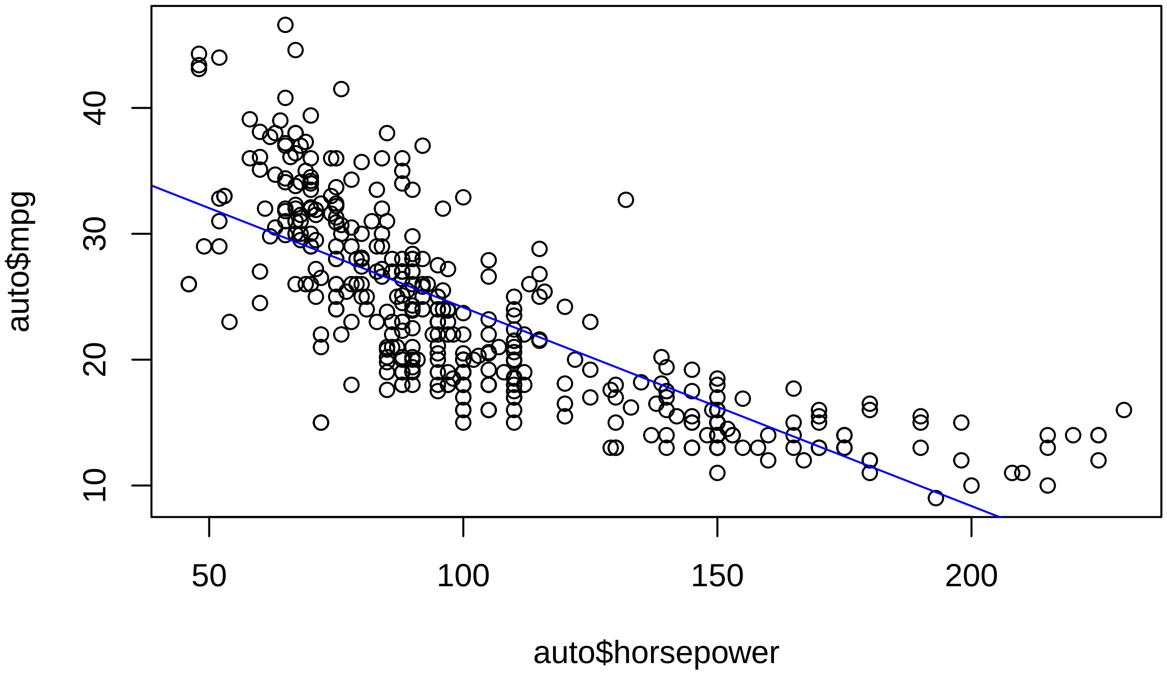
5(j) The 90% CI for the slope is (-0.168, -0.147).

```
confint(simple_reg, 'horsepower', level=0.90)
```

```
##                   5 %       95 %
## horsepower -0.1684719 -0.1472176
```

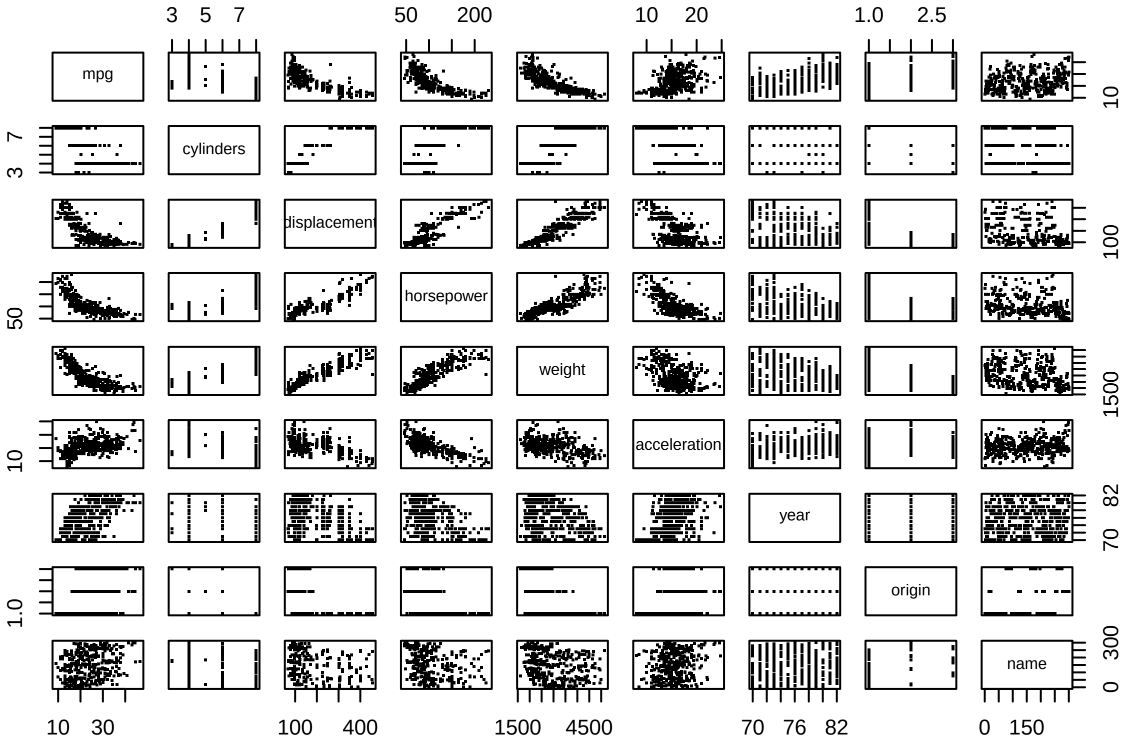
5(k) Our model assumes that there is a linear relationship between horsepower and mpg. However, instead of a straight line, the scatter plot is curve shaped, indicating potential non-linearity between horsepower and mpg.

```
plot(auto$horsepower, auto$mpg)
abline(simple_reg, col = "blue")
```



Question 6

```
plot(auto, pch=".") # part a
```



```
round(cor(auto[,1:7], use="pair"),4) # part b
```

```
##                  mpg cylinders displacement horsepower  weight acceleration
## mpg           1.0000   -0.7776      -0.8051    -0.7784 -0.8322       0.4233
## cylinders    -0.7776    1.0000       0.9508     0.8430  0.8975      -0.5047
## displacement -0.8051    0.9508       1.0000     0.8973  0.9330      -0.5438
## horsepower   -0.7784    0.8430       0.8973     1.0000  0.8645      -0.6892
## weight       -0.8322    0.8975       0.9330     0.8645  1.0000      -0.4168
## acceleration  0.4233   -0.5047      -0.5438    -0.6892 -0.4168       1.0000
## year          0.5805   -0.3456      -0.3699    -0.4164 -0.3091       0.2903
##                 year
## mpg           0.5805
## cylinders    -0.3456
## displacement -0.3699
## horsepower   -0.4164
## weight       -0.3091
## acceleration  0.2903
## year          1.0000
```

```
fit = lm(mpg~.-name, auto) # part c
summary(fit)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ . - name, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

6(a) The scatter plot indicates relatively strong relationship between mpg and displacement, mpg and horsepower, mpg and weight. There is weak relationship between mpg and acceleration, mpg and year. Lastly from the scatter plot, there is no significant linear relationship between mpg and year, and mpg and origin.

6(b) The correlation between mpg and displacement is -0.805, indicating a strong negative relationship, and mpg tends to decrease when displacement increases.

```
round(cor(auto[, 1:7], use = "pair"), 4)
```

```
##                  mpg cylinders displacement horsepower  weight acceleration
## mpg           1.0000   -0.7776      -0.8051    -0.7784 -0.8322       0.4233
## cylinders    -0.7776    1.0000       0.9508     0.8430  0.8975      -0.5047
## displacement -0.8051    0.9508       1.0000     0.8973  0.9330      -0.5438
## horsepower   -0.7784    0.8430       0.8973     1.0000  0.8645      -0.6892
## weight       -0.8322    0.8975       0.9330     0.8645  1.0000      -0.4168
## acceleration  0.4233   -0.5047      -0.5438    -0.6892 -0.4168       1.0000
## year          0.5805   -0.3456      -0.3699    -0.4164 -0.3091       0.2903
##                 year
## mpg           0.5805
## cylinders    -0.3456
## displacement -0.3699
## horsepower   -0.4164
## weight       -0.3091
## acceleration  0.2903
## year          1.0000
```

6(c) The null hypothesis in this case is that all of the regression coefficients are 0, and the alternative hypothesis states that at least one coefficient is nonzero. $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ versus

$H_1 : \textit{at least one } \beta_j \neq 0$ According to the model, the p-value is $2.2e^{-16} < 0.05$. Therefore we reject the null hypothesis and conclude that at least one predictor is significant.

```
fit = lm(mpg~.-name, auto)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

6(d) Weight, year, and origin appear to have a statistically significant relationship to mpg under 0.001 significant level. Displacement is significant under 0.01 significant level.

6(e) The slope coefficient for the year variable is around 0.75, suggesting a positive relationship between year and mpg. When model year increases by 1, mpg could increase by 0.75. Since the higher the mpg, the more economically friendly the car is, this coefficient suggests that newer cars are more economically friendly.

6(f) The slope coefficient for the displacement variable is around 0.02, suggesting a weak positive relationship between displacement and mpg. This discovery contradicts with findings from 6b (there is a negative relationship between displacement and mpg, the lower the displacement, the less fuel it consumes and thus higher mpg).

We think the findings from 6b is more plausible because there can be collinearity between displacement and other predictors in the multiple regression model. From the correlation matrix below, we can see that displacement is strongly correlated with cylinders, horsepower, and weight. In this case, further analysis such as variable selection and dimension reduction is needed.

```
auto_temp = auto[, c("mpg", "cylinders", "displacement", "horsepower", "weight", "acceleration", "year"
cor(auto_temp)
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```