The overall question of interest throughout this exam is: **Should miles per gallon be predicted based on weight alone, or on the linear combination of weight and displacement?**

Of course the later is correct, From physics we know that Energy consumed is linearly dependent of work done by Objects which is Combination of weight and displacement

1. Answer the following based on a *simple* linear regression, predicting *mpg* (y) with *weight* (x_1).

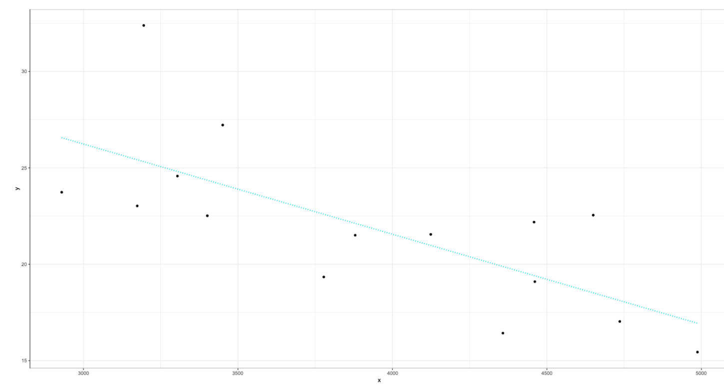
   (a) Fit the specified model. Write the model equation, including your estimates.

```{r}
x1 <- c(4124.129, 4736.041, 3777.898, 3174.024,
        4650.112, 3194.868, 3400.909, 4458.683,
        3879.585, 3450.74,  2929.358, 3304.248,
        4461.215, 4987.675, 4357.654)


y <- c(21.54716, 17.02911, 19.33781, 23.02399,
       22.54566, 32.38923, 22.5144,  22.18444,
       21.50476, 27.21958, 23.73371, 24.57349,
       19.09633, 15.44052, 16.42429)



summary(lm(y~x1))
```

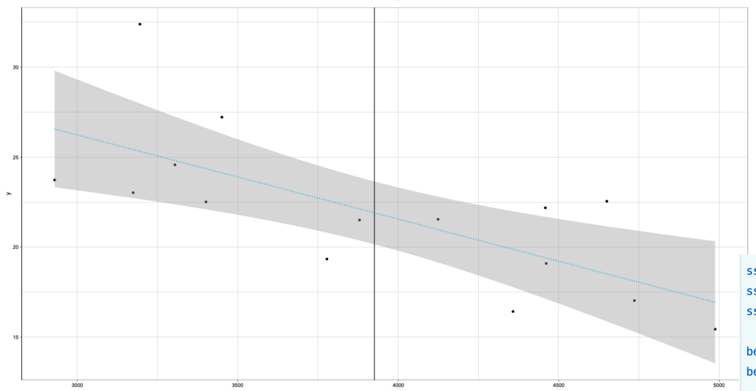
```
Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4600 -2.1210 -0.6158  1.6716  7.0659 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40.267655   5.038457   7.992 2.26e-06 ***
x1          -0.004678   0.001267  -3.692  0.00271 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.131 on 13 degrees of freedom
Multiple R-squared:  0.5119,	Adjusted R-squared:  0.4744 
F-statistic: 13.63 on 1 and 13 DF,  p-value: 0.002709
```

$$Y = \beta_1 \underset{\sim}{X} + \beta_0 \Rightarrow \boxed{\hat{Y} = -0.004678\, x_1 + 40.2677}$$

(b) Create a scatterplot of *mpg* and *weight*. Add a line representing the model, with 95% confidence bands. Does the model appear to fit the data?



```{r}
library(ggplot2)
library(tidyverse)
library(ggthemes)


x <- c(4124.129, 4736.041, 3777.898, 3174.024,
            4650.112, 3194.868, 3400.909, 4458.683,
            3879.585, 3450.74,  2929.358, 3304.248,
            4461.215, 4987.675, 4357.654)

y <- c(21.54716, 17.02911, 19.33781, 23.02399,
         22.54566, 32.38923, 22.5144,  22.18444,
         21.50476, 27.21958, 23.73371, 24.57349,
         19.09633, 15.44052, 16.42429)



summary(lm(y~x))



data <- tibble(x = x, y = y)

ggplot(data) + geom_point(aes(x,y)) + theme_bw() +
  geom_smooth(aes(x,y), method="lm",lty = 3, col="cyan3", se = F)
```



```r
ss_xy <- sum( (x-mean(x)) * (y-mean(y)))
ss_xx <- sum( (x-mean(x))^2)
ss_yy <- sum( (y-mean(y))^2)

beta_one_hat <-ss_xy/ss_xx
beta_zero_hat <- mean(y) - beta_one_hat * mean(x)

sse <- ss_yy - beta_one_hat * ss_xy
s <- sqrt(sse/(length(y) -2))
other <-sqrt(1/ss_xx)

t <- beta_one_hat/(s*other)

data1 <- data.frame(x = x,y = y)
temp_var <- predict(lm(y ~x),interval = "prediction")

data <- cbind(data,data.frame(temp_var))
names(data) <- c("x","y","fit","lwr","upr")

ggplot(data) + geom_point(aes(x,y)) + theme_linedraw() +
  geom_smooth(aes(x,y), method="lm",lty = 3, col="deepskyblue2", se = T, level = 0.95)+
  geom_vline(xintercept = mean(x))
```

The data apper to be lossely fitted w/ data which is Reasonable our R^2 is around 0.5

(c) Test the null hypothesis that the slope of x_1, β_1, is equal to zero. State the hypotheses, test statistic, rejection region(s), and p-value. **Do not** interpret the conclusion of this test.

$H_0: \mu = \beta_1 = 0 \qquad H_A: \mu = \beta_1 \neq 0$. (two-tailed Rejection Region)

$$\text{T-test} = \frac{\beta_i - \beta_{i0}}{s\sqrt{\frac{1}{S_{xx}}}}, \text{ where } S = \sqrt{SEE/(n-2)}$$

$$= -3.6925$$

Let $\alpha = 0.05$

Rejection Region: $(-\infty, -2.1609] \cup [2.1609, \infty)$

p-value: $P(>|t|) = 0.00271$ for 2-tail

```r
x <- c(4124.129, 4736.041, 3777.898, 3174.024,
           4650.112, 3194.868, 3400.909, 4458.683,
           3879.585, 3450.74,  2929.358, 3304.248,
           4461.215, 4987.675, 4357.654)

y <- c(21.54716, 17.02911, 19.33781, 23.02399,
         22.54566, 32.38923, 22.5144,  22.18444,
         21.50476, 27.21958, 23.73371, 24.57349,
         19.09633, 15.44052, 16.42429)

s_xy <- sum( (x-mean(x)) * (y-mean(y)))
s_xx <- sum( (x-mean(x))^2)
s_yy <- sum( (y-mean(y))^2)

beta_one_hat <-s_xy/s_xx
beta_zero_hat <- mean(y) - beta_one_hat * mean(x)

sse <- s_yy - beta_one_hat * s_xy
s <- sqrt(sse/(length(y) -2))
c_ii <-sqrt(1/s_xx)

t <- beta_one_hat/(s*c_ii)

t
```

```
[1] -3.692481
```

```
Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4600 -2.1210 -0.6158  1.6716  7.0659 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40.267655   5.038457   7.992 2.26e-06 ***
x1          -0.004678   0.001267  -3.692  0.00271 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.131 on 13 degrees of freedom
Multiple R-squared:  0.5119,	Adjusted R-squared:  0.4744 
F-statistic: 13.63 on 1 and 13 DF,  p-value: 0.002709
```

2. Answer the following based on a *multiple* linear regression, predicting *mpg* with *weight* (x_1) and *engine displacement* (x_2).

   (a) Fit the specified model. Write the model equation, including your estimates.

```{r}
x1 <- c(4124.129, 4736.041, 3777.898, 3174.024,
        4650.112, 3194.868, 3400.909, 4458.683,
        3879.585, 3450.74,  2929.358, 3304.248,
        4461.215, 4987.675, 4357.654)

x2 <- c(178.5575,236.0139,179.4107,190.2972,
         164.4554,114.4701,168.2990,208.4433,
         197.3525,137.7964,122.0215,142.4937,
         218.8619,302.1571,239.6896)

y <- c(21.54716, 17.02911, 19.33781, 23.02399,
       22.54566, 32.38923, 22.5144,  22.18444,
       21.50476, 27.21958, 23.73371, 24.57349,
       19.09633, 15.44052, 16.42429)



summary(lm(y~x1))
```

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1342 -0.9828 -0.6934  1.4039  5.0779 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 36.5095516  3.8852963   9.397 6.98e-07 ***
x1          -0.0003083  0.0015820  -0.195    0.849    
x2          -0.0717513  0.0209294  -3.428    0.005 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.316 on 12 degrees of freedom
Multiple R-squared:  0.7534,    Adjusted R-squared:  0.7123 
F-statistic: 18.33 on 2 and 12 DF,  p-value: 0.0002248
```

$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \Rightarrow \boxed{\hat{Y} = 36.51 - 0.0003083 X_1 - 0.07175 X_2}$

(b) Test the null hypothesis that the slope of x_1, β_1, is equal to zero. State the hypotheses, test statistic, rejection region(s), and p-value. Interpret the conclusion of this test at $\alpha = 0.05$.

$H_0 : \mu = \beta_1 = 0 \qquad H_A : \mu = \beta_1 \neq 0$

$$\text{T - test} : \frac{\beta_i - \beta_{i0}}{S\sqrt{\frac{1}{S_{xx}}}}$$

$$= -0.195$$

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1342 -0.9828 -0.6934  1.4039  5.0779 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 36.5095516  3.8852963   9.397 6.98e-07 ***
x1          -0.0003083  0.0015820  -0.195    0.849    
x2          -0.0717513  0.0209294  -3.428    0.005 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.316 on 12 degrees of freedom
Multiple R-squared:  0.7534,	Adjusted R-squared:  0.7123 
F-statistic: 18.33 on 2 and 12 DF,  p-value: 0.0002248
```

Df: $15 - 2 - 1 = 13$

Rejection Region:

$(-\infty, -2.1788] \cup [2.1788, \infty)$

p-value: $P(>|t|) = 0.849$

See code in Part (a) of this Question

```{r}
qt(p=0.05/2,df=12,lower.tail = F)
```

```
[1] 2.178813
```

Fail to Reject Ho, which is saying that β_1 can be dropped.

(c) Consider $x_1^* = 3000$ and $x_2^* = 150$. Calculate a 95% confidence interval for $E[Y|x_1 = x_1^*, x_2 = x_2^*]$. Calculate a 95% prediction interval for y_i, given $x_1 = x_1^*$ and $x_2 = x_2^*$. Interpret both of these intervals in context.

$$E[Y \mid X_1 = 3000,\ X_2 = 150] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = a'\beta$$

$$a' = \begin{bmatrix} 1 \\ X_1 = 3000 \\ X_2 = 150 \end{bmatrix} \qquad \boxed{T_{\alpha/2,\ df:12} = 2.179}$$

```{r}
x1 <- c(4124.129, 4736.041, 3777.898, 3174.024,
        4650.112, 3194.868, 3400.909, 4458.683,
        3879.585, 3450.74,  2929.358, 3304.248,
        4461.215, 4987.675, 4357.654)

x2 <- c(178.5575,236.0139,179.4107,190.2972,
        164.4554,114.4701,168.2990,208.4433,
        197.3525,137.7964,122.0215,142.4937,
        218.8619,302.1571,239.6896)
x0 <- c(1,1,1,1,
        1,1,1,1,
        1,1,1,1,
        1,1,1)

yy <- c(21.54716, 17.02911, 19.33781, 23.02399,
        22.54566, 32.38923, 22.5144,  22.18444,
        21.50476, 27.21958, 23.73371, 24.57349,
        19.09633, 15.44052, 16.42429)

xx <- cbind(x0,x1,x2)
x_trans = t(xx)
x_inverse = solve(x_trans%*%xx)
beta_hat=x_inverse %*%x_trans%*%yy
a <-c(1,3000,150)
y_trans <- t(yy)
sse <- y_trans%*%yy - t(beta_hat)%*%x_trans%*%yy
n <- 15
k = 2
s <- sqrt(sse/(n-k-1))
a_trans= t(a)
```

```
k = 2
s <- sqrt(sse/(n-k-1))
a_trans= t(a)
t_alpha_half = 2.179
to_be_sqrt = t(a)%*%x_inverse%*%a
first_part=a_trans%*%beta_hat
second_part = t_alpha_half*(s*sqrt(to_be_sqrt))
first_part + second_part
first_part - second_part
```

```
        [,1]
[1,] 27.28766
        [,1]
[1,] 22.35653
```

```
n <- 15
k = 2
s <- sqrt(sse/(n-k-1))
a_trans= t(a)
t_alpha_half = 2.179
to_be_sqrt = t(a)%*%x_inverse%*%a
first_part=a_trans%*%beta_hat
second_part = t_alpha_half*(s*sqrt(to_be_sqrt+1))
first_part + second_part
first_part - second_part
```

```
        [,1]
[1,] 30.43943
        [,1]
[1,] 19.20476
```

95% CI (22.357 , 27.288)

95% PI (19.205, 30.439)

For CI, it can be said that we can be 95% confident the popularation mean resides between (22.357, 27.288)

For PI, it can be said that we can be 95% confident the next observation will fall within (19.205, 30.439)

(d) Which model constitutes the "complete" model and which the "reduced" model? Can x_2 be dropped from the model without losing predictive information? Test at the $\alpha = 0.05$ significance level.

The Complete model is :

$$\hat{Y}_C = 36.51 - 0.0003083 X_1 - 0.07175 X_2$$

Use F-test

The Reduced model is :

$$\hat{Y}_R = -0.004678\, x_1 + 40.2677$$

$H_0: \beta_2 = 0, \quad H_a: \beta_2 \neq 0, \quad \alpha = 0.05$

$$F = \frac{(SSE_R - SSE_c)/(k-g)}{SSE_c/(n-(k+1))} = \frac{(127.445 - 64.385)/(2-1)}{64.385)/(15-2-1)}$$

$= 11.752$

$F_{\alpha = 0.05,\ \nu_1 = 1,\ \nu_2 = 12} = 4.747$ from table

Since $F > F_\alpha \Rightarrow H_0$ Rejected

We can't drop x_2 from the model b/c it's Significant

````
x1 <- c(4124.129, 4736.041, 3777.898, 3174.024,
        4650.112, 3194.868, 3400.909, 4458.683,
        3879.585, 3450.74,  2929.358, 3304.248,
        4461.215, 4987.675, 4357.654)

x2 <- c(178.5575,236.0139,179.4107,190.2972,
        164.4554,114.4701,168.2990,208.4433,
        197.3525,137.7964,122.0215,142.4937,
        218.8619,302.1571,239.6896)
x0 <- c(1,1,1,1,
        1,1,1,1,
        1,1,1,1,
        1,1,1)

yy <- c(21.54716, 17.02911, 19.33781, 23.02399,
        22.54566, 32.38923, 22.5144,  22.18444,
        21.50476, 27.21958, 23.73371, 24.57349,
        19.09633, 15.44052, 16.42429)

xx <- cbind(x0,x1,x2)
x_trans = t(xx)
x_inverse = solve(x_trans%*%xx)
beta_hat=x_inverse %*%x_trans%*%yy
a <-c(1,3000,150)
y_trans <- t(yy)
sse <- y_trans%*%yy - t(beta_hat)%*%x_trans%*%yy
sse

```

      [,1]
[1,] 64.38564
````

````
x <- c(4124.129, 4736.041, 3777.898, 3174.024,
           4650.112, 3194.868, 3400.909, 4458.683,
           3879.585, 3450.74,  2929.358, 3304.248,
           4461.215, 4987.675, 4357.654)

y <- c(21.54716, 17.02911, 19.33781, 23.02399,
        22.54566, 32.38923, 22.5144,  22.18444,
        21.50476, 27.21958, 23.73371, 24.57349,
        19.09633, 15.44052, 16.42429)

s_xy <- sum( (x-mean(x)) * (y-mean(y)))
s_xx <- sum( (x-mean(x))^2)
s_yy <- sum( (y-mean(y))^2)

beta_one_hat <-s_xy/s_xx
beta_zero_hat <- mean(y) - beta_one_hat * mean(x)

sse <- s_yy - beta_one_hat * s_xy
sse
s <- sqrt(sse/(length(y) -2))
c_ii <-sqrt(1/s_xx)

t <- beta_one_hat/(s*c_ii)

beta_hat_1 <- s_xy/s_xx
beta_hat_0 <- mean(y)-beta_hat_1*mean(x)
#beta_hat_0
#beta_hat_1

```

[1] 127.4454
````

3. Consider your answers to the previous questions, then answer the following.

Suppose that the true population relationship is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Further suppose that there is a relationship between x_1 and x_2, given by:

$$x_2 = \gamma_0 + \gamma_1 x_1 + \delta$$

where γ_1 and β_2 are non-zero.

(a) Find the expected values of β_0 and β_1 if the independent variable x_2 is omitted from the regression.

Plug in $x_2 = \gamma_0 + \gamma_1 x_1 + \delta$

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

$y = \beta_0 + \beta_1 x_1 + \beta_2 (\gamma_0 + \gamma_1 x_1 + \delta) + \varepsilon$

$y = (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) x_1 + (\delta + \varepsilon)$

$\boxed{[\beta_0] = \beta_0 + \beta_2 \gamma_0 \qquad [\beta_1] = \beta_1 + \beta_2 \gamma_1}$

$E[\hat{\beta}_1] = E\left[\frac{\Sigma (x_i - \bar{x}) y_i}{S_{xx}}\right]$ where $E[y_i] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

We know $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$

$\Rightarrow E[\hat{\beta}_1] = \frac{\Sigma (x_i - \bar{x})(\beta_0 + \gamma_0 \beta_2)}{S_{xx}} + \frac{\Sigma (x_i - \bar{x}) x_i}{S_{xx}} \cdot (\beta_1 + \beta_2 \gamma_1)$

$= 1 \cdot \boxed{(\beta_1 + \beta_2 \gamma_1)}$

$E[\hat{\beta}_0] = E(\bar{y}) - E(\hat{\beta}_1) \bar{x}_1$

$= (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) \bar{x}_1 - (\beta_1 + \beta_2 \gamma_1) \bar{x}_1$

$\boxed{= \beta_0 + \beta_2 \gamma_0}$

(b) Calculate the bias (if any) of β_0 and β_1 when x_2 is omitted.

$$E(\hat{\beta}_1) - \beta_1 = \text{bias}[\hat{\beta}]$$

$$\beta_1 + \beta_2 \gamma_1 - \beta_1 = \text{bias}[\hat{\beta}]$$

$$\text{bias}[\hat{\beta}_1] = \beta_2 \gamma_1$$

$$E[\hat{\beta}_0] - \beta_0 = \text{bias}(\hat{\beta}_0)$$

$$\boxed{\text{Bias}[\hat{\beta}_0] = \beta_2 \gamma_0}$$

(c) What values of γ_1 and β_2 would result in β_0 and β_1 remaining unbiased?

① $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

② $x_2 = \gamma_0 + \gamma_1 x_1 + \delta$

We know E(Y) is unbiased because we found it meet Condition 1-4.

Case 1

γ_1, β_2 are non-zero

Since

$$y = \beta_0 + \beta_1 x_1 + \beta_2(\gamma_0 + \gamma_1 x_1 + \delta) + \varepsilon$$

$$= \beta_0 + (\beta_1 + \beta_2\gamma_1)x_1 + \beta_2\gamma_0 + \beta_2\delta + \varepsilon$$

$\Rightarrow$ $\boxed{\beta_1 = -\beta_2\gamma_1}$ would Result in x_1 being omitted from ①,

Case 2

γ_1, β_2 are zero

$E(\hat{\beta}_1) = \beta_1 + \beta_2\gamma_1 \Rightarrow \beta_2\gamma_1 = 0 \Rightarrow \beta_2 = 0$ or $\gamma_1 = 0$

(d) In light of the above:

i. What assumption of linear regression is being violated in Question 1? Is this assumption met in Question 2?

ii. In Question 1, are the estimates of β_0 and β_1 BLUE? Why or why not?

i) Assumptions being violated in Q1 : $E(\varepsilon) = 0$
Yes it's been met in Q2 .

ii) No, β_0 and β_1 are not BLUE .
At least 1 Assumption is not met.