Test Exercise 2

In [4]:
```
data <- read.csv("./data/TestExer2-GPA-round2.csv")
names(data)[1] <- "Observation"
```

In [5]:
```
summary(data)
```

```
  Observation        FGPA            SATM            SATV             FEM
 Min.   :  1   Min.   :1.500   Min.   :4.000   Min.   :3.100   Min.   :0.0000
 1st Qu.:153   1st Qu.:2.485   1st Qu.:5.900   1st Qu.:5.100   1st Qu.:0.0000
 Median :305   Median :2.773   Median :6.300   Median :5.500   Median :0.0000
 Mean   :305   Mean   :2.793   Mean   :6.248   Mean   :5.565   Mean   :0.3875
 3rd Qu.:457   3rd Qu.:3.116   3rd Qu.:6.600   3rd Qu.:6.000   3rd Qu.:1.0000
 Max.   :609   Max.   :3.971   Max.   :7.900   Max.   :7.600   Max.   :1.0000
```

(a) (i) Regress FGPA on a constant and SATV. Report the coefficient of SATV and its standard error and p-value (give your answers with 3 decimals).

In [44]:
```
fit <- lm(FGPA ~ SATV, data=data)
sprintf("Regression line: FGPA = %.2f + %.2fSATV + e", coef(fit)[1], coef(fit)[2])
print(summary(fit)$coefficient)
```

'Regression line: FGPA = 2.44 + 0.06SATV + e'

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 2.44173246 0.15506207 15.746807 4.257353e-47
SATV        0.06308585 0.02766362  2.280462 2.292611e-02
```

(a) (ii) Determine a 95% confidence interval (with 3 decimals) for the effect on FGPA of an increase by 1 point in SATV.

In [51]:
```
se = coef(summary(fit))[2, "Std. Error"]
sprintf("Confidence interval at 95%%: (%.25, %.25), c oeef(fit)[2]-2*cse coef(fit)[2
```

'Confidence interval at 95%: (0.00776, 0.11841)'

(b) Answer questions (a-i) and (a-ii) also for the regression of FGPA on a constant, SATV, SATM, and FEM.

In [52]:
```
multi_fit <- lm(FGPA ~ SATV+SATM+FEM, data=data)
sprintf("Regression line: FGPA = %.2f + %.2fSATV + %.2fSATM + %.2fFEM + e", coef(mul
print(summary(multi_fit)$coefficients)
```

'Regression line: FGPA = 1.56 + 0.01SATV + 0.17SATM + 0.20FEM + e'

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 1.5570482 0.21609551 7.2053704 1.729863e-12
SATV        0.0141619 0.02792697 0.5071047 6.122662e-01
SATM        0.1727359 0.03192671 5.4103874 9.071480e-08
FEM         0.2002716 0.03738085 5.3575989 1.200266e-07
```

In [55]:
```
se_SATV = coef(summary(multi_fit))[2, "Std. Error"]
sprintf("Confidence interval afor SATV coefficient t 95%%: (%.5f, %.5f)",  coef(fmul

se_SATV =Mcoef(summary(fitmulti_))[2, 3Std. Error"]
sprintf("Confidence interval at 9for SATM coefficient 5%%: (%.5f, %.5f)",  coef(fit)
```

```
se_SATM_SA
se_SFEM= coef(summary(fmulti_it))[24 "Std. Error"]
sprintf("Confidence interval a for FEM coefficientt 95%%: (%.5f, %.5f)",  coef(fimul
```

'Confidence interval for SATV coefficient at 95%: (-0.04169, 0.07002)'

'Confidence interval for SATM coefficient at 95%: (0.10888, 0.23659)'

'Confidence interval for FEM coefficient at 95%: (0.12551, 0.27503)'

(c) Determine the (4 × 4) correlation matrix of FGPA, SATV, SATM, and FEM. Use these correlations to explain the differences between the outcomes in parts (a) and (b).

In part (a): $\beta_{SATV} =$ 0.063
In part (b): $\beta_{SATV} =$ 0.014

The value of $\beta_{SATV} =$ is affected by its relationship with other factors. We can look at the formula that governs the effect of towards the dependent variable.

$\text{Total Effect} = \text{Partial Effect} + \text{Indirect Effect}$
$\frac{dy}{dx_j} = \beta_j + \sum_{h=2,h\neq j}^{k} \beta_h \frac{\partial x_h}{\partial x_j}$

Where $\frac{\partial x_h}{\partial x_j}$ is proportional to the correlation between factor of interest (SATV in this case) with other factors.
The above equation means that as we add other factors, the effect of existing factors are explained by their relationship with other factors. Since the total effect for each factor remains the same, β of the factor of interest will change.

From the correlation matrix:
Correlation between SATV and SATM: 0.2878 > 0, $\beta_{SATM} =$ 0.173 > 0
Correlation between SATV and FEM: 0.0336 > 0, $\beta_{FEM} =$ 0.200 > 0
Since total effect remains the same, β_{SATV} has a lower value.

In [59]:
```
print(round(cor(data[, -1]), 4))
```

```
       FGPA    SATM   SATV     FEM
FGPA 1.0000  0.1950 0.0922  0.1765
SATM 0.1950  1.0000 0.2878 -0.1627
SATV 0.0922  0.2878 1.0000  0.0336
FEM  0.1765 -0.1627 0.0336  1.0000
```

(d) (i) Perform an F-test on the significance (at the 5% level) of the effect of SATV on FGPA, based on the regression in part (b) and another regression.

Note: Use the F-test in terms of SSR or R2 and use 6 decimals in your computations. The relevant critical value is 3.9.

$F = \frac{(R_1^2 - R_0^2)/g}{(1-R_1^2)/(n-k)}$
$H_0 : \beta_{SATV} = 0,\ H_1 : \beta_{SATV} \neq 0$

In [76]:
```
multi_fit2<<- lm(FGPA ~ SATM+FEM, data=data)
```

In [111…
```
printf("Regression l0 (another regression) FGPA = %.2f + %.2fSATM + %.2fFEM + e", co

rsprintf("Regression 1 (from part (b)): FGPA = %.2f + %.2fSATV + %.2fSATM + %.2fFEM

rsq_1 = as.numeric(summary(multi_fit2)$.squared)

g = 1F n = nrow(data)
=k = 4 ()rsq_rsq_1 - 0*()n-k/()1-sqrrsq_1/g
sprint()Ff""F-statistics value: %0.5f6 F
sprintf()""F-statistics value = -0.2%0.6f.9, therefore do not reject $H_$, F
```

'Regression 0 (another regression): FGPA = 1.61 + 0.18SATM + 0.20FEM + e'

'Regression 1 (from part (b)): FGPA = 1.56 + 0.01SATV + 0.17SATM + 0.20FEM + e'

'F-statistics value: 0.257155'

'F-statistics value = 0.257155 < 3.9, therefore do not reject H_0'

(d) (ii) Check numerically that F = t^2

$$t = \frac{b_j}{s\sqrt{a_{jj}}}$$

In [124…
```
b_j <- coef(multi_fit)[2]
se <- coef(summary(multi_fit))[2, 'Std. Error']
# s = summary(multi_fit)$sigma

# X <- data.matrix(data[, c(3:5)])
# XTX <- t(X) %*% (X)
# XTX_inv <- solve(XTX)
# a_jj = XTX_inv[2,2]
# tsq = b_j**2/((s**2)*a_jj)
tsq = (b_j/se)**2

sprintf("F = %0.6f", F)
sprintf("t-squared = %0.6f", tsq)
```

'F = 0.257155'

't-squared = 0.257155'

In [ ]: