STAT30040 Homework 6

Sarah Adilijiang

Problem 1

$MSE(\hat{\theta}) = E(\hat{\theta}-\theta)^2 = E(\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2) = [(E(\hat{\theta}))^2 - 2\theta E(\hat{\theta}) + \theta^2] + [E(\hat{\theta}^2) - (E(\hat{\theta}))^2] = bias^2(\hat{\theta}) + Var(\hat{\theta})$

Since X_i i.i.d. $\sim N(\mu, \sigma^2)$, we have: $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \;\Rightarrow\; \frac{\sum\limits_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2_{n-1} \;\Rightarrow\; \sum\limits_{i=1}^{n}(X_i-\bar{X})^2 \sim \sigma^2\chi^2_{n-1}$

So $E(\sum\limits_{i=1}^{n}(X_i-\bar{X})^2) = \sigma^2(n-1), \quad Var(\sum\limits_{i=1}^{n}(X_i-\bar{X})^2) = 2\sigma^4(n-1)$

$$\Rightarrow\; E(s^2) = \frac{1}{n-1}E(\sum_{i=1}^{n}(X_i-\bar{X})^2) = \frac{n-1}{n-1}\sigma^2 = \sigma^2 \;\Rightarrow\; bias(s^2) = E(s^2) - \sigma^2 = 0$$

$$E(\hat{\sigma}^2) = \frac{1}{n}E(\sum_{i=1}^{n}(X_i-\bar{X})^2) = \frac{n-1}{n}\sigma^2 \;\Rightarrow\; bias(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\sigma^2/n$$

$$\Rightarrow\; Var(s^2) = \frac{1}{(n-1)^2}Var(\sum_{i=1}^{n}(X_i-\bar{X})^2) = \frac{2(n-1)}{(n-1)^2}\sigma^4 = \frac{2}{n-1}\sigma^4$$

$$Var(\hat{\sigma}^2) = \frac{1}{n^2}Var(\sum_{i=1}^{n}(X_i-\bar{X})^2) = \frac{2(n-1)}{n^2}\sigma^4$$

$$\Rightarrow\; MSE(s^2) = bias^2(s^2) + Var(s^2) = 0 + \frac{2}{n-1}\sigma^4 = \frac{2}{n-1}\sigma^4$$

$$MSE(\hat{\sigma}^2) = bias^2(\hat{\sigma}^2) + Var(\hat{\sigma}^2) = \frac{1}{n^2}\sigma^4 + \frac{2(n-1)}{n^2}\sigma^4 = \frac{2n-1}{n^2}\sigma^4$$

Comparing these two values, since $n \geq 1$, we have:

$$\frac{MSE(s^2)}{MSE(\hat{\sigma}^2)} = \frac{2n^2}{(n-1)(2n-1)} = \frac{2n^2}{2n^2-3n+1} > 1, \textit{ and } \lim_{n\to\infty}\frac{MSE(s^2)}{MSE(\hat{\sigma}^2)} = \lim_{n\to\infty}\frac{2n^2}{2n^2-3n+1} = 1$$

Thus, we get that: $MSE(s^2) > MSE(\hat{\sigma}^2)$, and when n is large, $MSE(s^2) \approx MSE(\hat{\sigma}^2)$.

Problem 2

(a)

```
asthma = read.table("./data/asthma.txt", header=TRUE)
x = asthma$height
y = asthma$resistance
SXX = sum( (x-mean(x))^2 )
beta_1 = sum( (x-mean(x))*y ) / SXX; beta_1
```

```
## [1] -0.1360581
```

```
beta_0 = mean(y) - beta_1*mean(x);  beta_0
```

```
## [1] 27.5157
```

Answer:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \approx -0.1361, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \approx 27.5157$$

So the fitted least squares line is: $y = \hat{\beta}_0 + \hat{\beta}_1 x = 27.5157 - 0.1361x$, where y is the respiratory resistance, and x is the height (cm).

(b)

```
y_fit = beta_0 + beta_1*x
resids = y - y_fit
RSS = sum(resids^2);  RSS
```

```
## [1] 796.0553
```

```
n = nrow(asthma)
se = sqrt(RSS/(n-2));  se
```

```
## [1] 4.461097
```

```
beta_1_se = sqrt(se^2/SXX);  beta_1_se
```

```
## [1] 0.0530971
```

```
T_stat = beta_1 / beta_1_se;  T_stat
```

```
## [1] -2.56244
```

```
p_val = 2 * pt(T_stat, n-2);  p_val
```

```
## [1] 0.01426396
```

Answer:

Since $RSS = \sum_{i=1}^{n}(y-\hat{y})^2 \approx 796.0553$, so $se = \hat{\sigma} = \sqrt{\frac{RSS}{n-2}} \approx 4.4611$

Under $H_0 : \beta_1 = 0$, the test statistic is:

$$T = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \approx \frac{-0.1361}{0.0531} \approx -2.5624$$

And under null hypothesis: $T \sim t(n-2)$, so p-value is: $p = 2 \times P(T < -2.5624) \approx 0.0143$

Since p-value < 0.05, so we reject H_0 at 5% significance level, so the slope is significantly different from zero.

(c)

```
t = qt(0.975, n-2)

# 95% CI for slope
c(beta_1-t*beta_1_se, beta_1+t*beta_1_se)
```

```
## [1] -0.24337134 -0.02874487
```

```r
# 95% CI for intercept
beta_0_se = sqrt( (se^2/SXX) * sum(x^2)/n );  beta_0_se
```

```
## [1] 6.636272
```

```r
c(beta_0-t*beta_0_se, beta_0+t*beta_0_se)
```

```
## [1] 14.10329 40.92811
```

Answer:

In question (b), we have obtained that: $se(\hat{\beta}_1) = \sqrt{V\hat{a}r(\hat{\beta}_1)} \approx 0.0531$, so a 95% confidence interval for the slope β_1 is:

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-2) \times se(\hat{\beta}_1) \approx (-0.2434, -0.0287)$$

Since $se(\hat{\beta}_0) = \sqrt{V\hat{a}r(\hat{\beta}_0)} = \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2/n}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \approx 6.6363$, so a 95% confidence interval for the intercept β_0 is:

$$\hat{\beta}_0 \pm t_{\alpha/2}(n-2) \times se(\hat{\beta}_0) \approx (14.1033, 40.9281)$$

(d)

```r
SST = sum( (y-mean(y))^2 )
R2 = 1 - RSS/SST;   R2
```

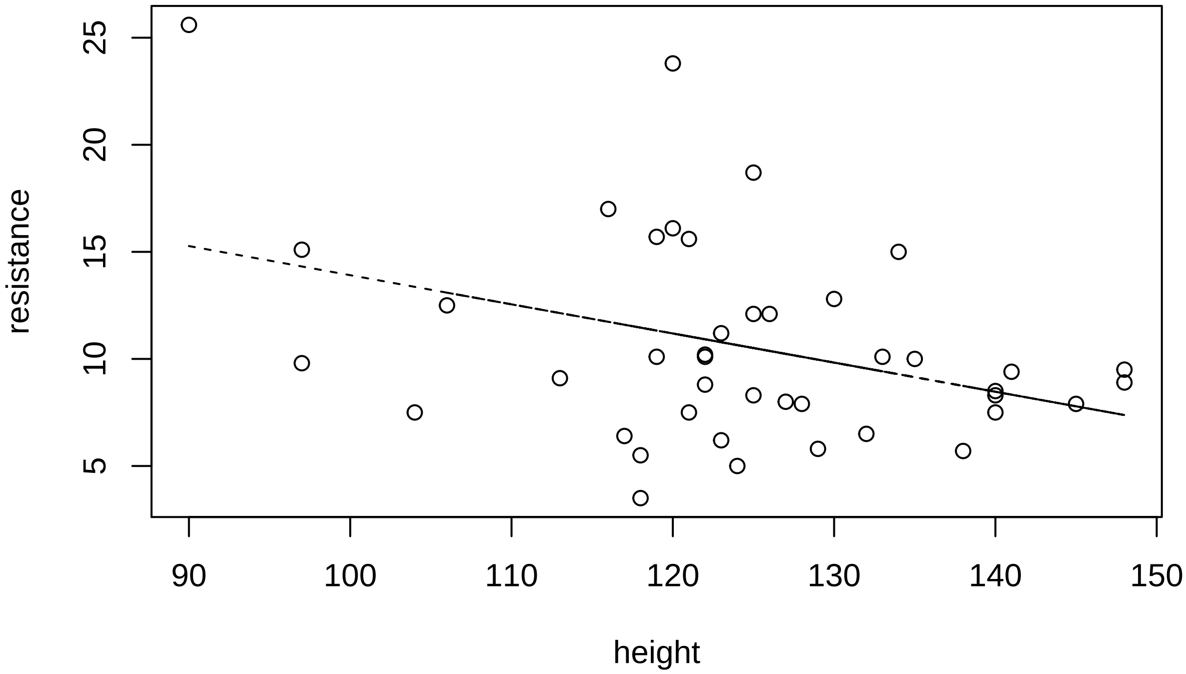
```
## [1] 0.141006
```

Answer:

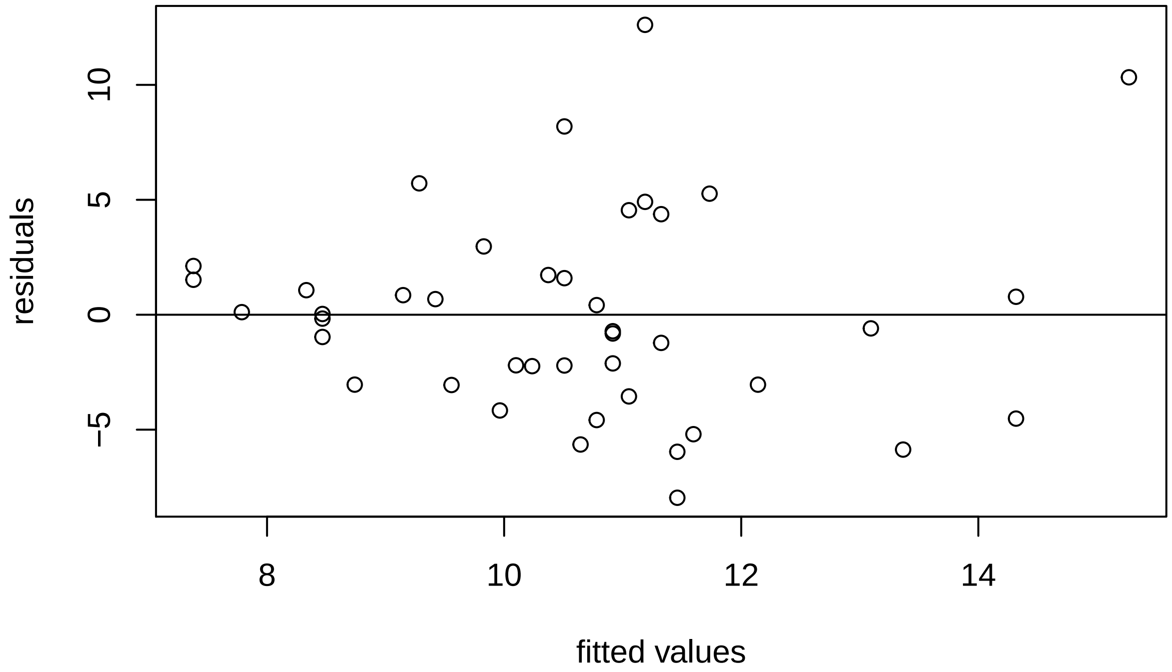
$R^2 = 1 - \frac{RSS}{SS_{Total}} \approx 0.1410$, so about 14.10% of the variation in the data is explained by the simple linear regression model.

(e)

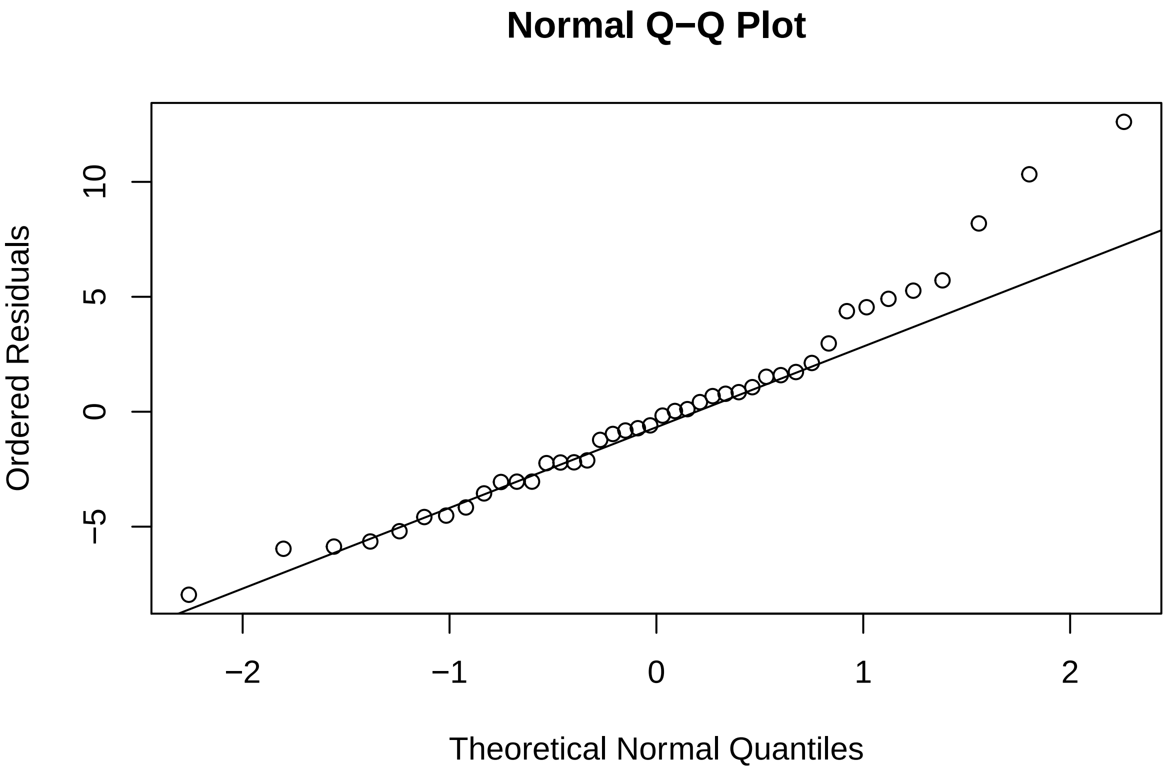
```r
# scatter plot
plot(y~x, xlab="height", ylab="resistance")
lines(y_fit~x, lty=2)
```



```
# residual plot
plot(resids~y_fit, xlab="fitted values", ylab="residuals")
abline(h=0)
```



```
# normal probability plot for the residuals
qqnorm(resids, ylab="Ordered Residuals", xlab="Theoretical Normal Quantiles")
qqline(resids)
```



(f)

Discussion:

In the scatter-plot, the data does not seem to fit in the line very well, and the linear relationship between the response and the predictor is not very clear.

In the residual plot, the variance of errors seems to increase as the fitted value increases, indicating a nonconstant variance problem in this dataset, which violates the model assumptions.

In the normal probability plot for the residuals, i.e. QQ-plot, right tail of the points does not lie in the line of theoretical normal quantiles and there is an increasing trend above the line. Thus the error terms do not follow a good sysmetric normal distribution but having a heavy right tail, which also somewhat violates the model assumptions.

Problem 3

(a)

For least squares estimation, we want to minimize: $S(\mu) = \sum\limits_{i=1}^{n} \epsilon_i^2 = \sum\limits_{i=1}^{n} (y_i - \mu)^2$

So we set the following derivative to zero: $\frac{\partial S}{\partial \mu} = -2 \sum\limits_{i=1}^{n} (y_i - \mu) = 0$

And since $\frac{\partial^2 S}{\partial \mu^2} = 2n > 0$

Thus we obtain the minimizer $\hat{\mu} = \frac{\sum\limits_{i=1}^{n} y_i}{n} = \bar{y}$

(b)

For least squares estimation, we want to minimize: $S(\beta) = \sum\limits_{i=1}^{n} \epsilon_i^2 = \sum\limits_{i=1}^{n} (y_i - \beta x_i)^2$

So we set the following derivative to zero: $\frac{\partial S}{\partial \beta} = -2 \sum\limits_{i=1}^{n} x_i(y_i - \beta x_i) = 0$

And since $\frac{\partial^2 S}{\partial \beta^2} = 2 \sum\limits_{i=1}^{n} x_i^2 \geq 0$

Thus we obtain the minimizer $\hat{\beta} = \frac{\sum\limits_{i=1}^{n} x_i y_i}{\sum\limits_{i=1}^{n} x_i^2}$

(c)

$$Var(\hat{\beta_1}) = \frac{\hat{\sigma}^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{RSS/(n-2)}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y_i})^2/(n-2)}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$$

In this expression, $RSS = \sum\limits_{i=1}^{n}(y_i - \hat{y_i})^2$ is basically determined by the correlation between $y_i's$ and $x_i's$, not the values of $x_i's$. So if we want to minimize $Var(\hat{\beta_1})$, we should maximize $\sum\limits_{i=1}^{n}(x_i - \bar{x})^2$. Since $x_i's \in [-1, 1]$, and:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Thus: $\max\{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2\} = \max\{\sum\limits_{i=1}^{n} x_i^2\} - n\min\{\bar{x}^2\}$

If n is an even number, $\min\{Var(\hat{\beta_1})\}$ is obtained when $\frac{n}{2}$ number of $x_i's$ are -1 and $\frac{n}{2}$ number of $x_i's$ are 1.

If n is an odd number, $\min\{Var(\hat{\beta_1})\}$ is obtained when $\frac{n-1}{2}$ number of $x_i's$ are -1 and $\frac{n+1}{2}$ number of $x_i's$ are 1; or when $\frac{n+1}{2}$ number of $x_i's$ are -1 and $\frac{n-1}{2}$ number of $x_i's$ are 1.

(d)

Set $K = \begin{bmatrix} 1 & & \\ & k & \\ & & 1 \end{bmatrix}$, so the new X^* is:

$$X^* = \begin{bmatrix} 1 & kx_{11} & x_{21} \\ 1 & kx_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & kx_{1n} & x_{2n} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix} \begin{bmatrix} 1 & & \\ & k & \\ & & 1 \end{bmatrix} = XK$$

Since in the matrix form: $\hat{\beta} = (X^T X)^{-1} X^T Y, \quad \hat{Y} = X\hat{\beta}$, so for new X^* we can derive that:

$$\begin{bmatrix} \hat{\beta_0^*} \\ \hat{\beta_1^*} \\ \hat{\beta_2^*} \end{bmatrix} = \hat{\beta^*} = ({X^*}^T X^*)^{-1} {X^*}^T Y = ((XK)^T(XK))^{-1}(XK)^T Y = (K^T X^T X K)^{-1} K^T X^T Y = K^{-1}(X^T X)^{-1}(K^{-1})^T K^T X^T Y$$

$$= K^{-1}(X^TX)^{-1}(KK^{-1})^TX^TY = K^{-1}(X^TX)^{-1}X^TY = K^{-1}\hat{\beta} = \begin{bmatrix} 1 & & \\ & 1/k & \\ & & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1/k \\ \hat{\beta}_2 \end{bmatrix}$$

Thus, we have: $\hat{\beta}_1^* = \hat{\beta}_1/k$, $\hat{\beta}_2^* = \hat{\beta}_2$, and $\hat{Y^*} = X^*\hat{\beta}^* = XKK^{-1}\hat{\beta} = X\hat{\beta} = \hat{Y}$

To sum up, $\hat{Y}$ and $\hat{\beta}_2$ both do not change, but $\hat{\beta}_1$ changes and its new estimate is: $\hat{\beta}_1^* = \hat{\beta}_1/k$

(e)

(i)

Since X_i are independent random variables with variance σ^2, so: $Var(Y_i) = Var(X_i - X_{i-1}) = Var(X_i) + Var(X_{i-1}) = 2\sigma^2$

And for $i \neq j$: $Cov(Y_i, Y_j) = Cov(X_i - X_{i-1}, X_j - X_{j-1}) = -Cov(X_i, X_{j-1}) - Cov(X_{i-1}, X_j) = -\sigma^2$ when $j = i + 1$ or $j = i - 1$, and zero otherwise.

Therefore, the covariance matrix of Y is:

$$Cov(Y) = \begin{bmatrix} 2\sigma^2 & -\sigma^2 & & & & \\ -\sigma^2 & 2\sigma^2 & -\sigma^2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\sigma^2 & 2\sigma^2 & -\sigma^2 \\ & & & & -\sigma^2 & 2\sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}$$

(ii)

$X_1 = Y_1$, $X_2 = Y_2 + X_1 = Y_2 + Y_1 = \sum\limits_{i=1}^{2} Y_i$, set $X_k = \sum\limits_{i=1}^{k} Y_i$, then $X_{k+1} = Y_k + X_k = Y_k + \sum\limits_{i=1}^{k} Y_i = \sum\limits_{i=1}^{k+1} Y_i$, thus we have derived that: $X_k = \sum\limits_{i=1}^{k} Y_i \quad \forall k = 1, ..., n$

Since Y_i are independent random variables with variance σ^2, so: $Var(X_k) = Var(\sum\limits_{i=1}^{k} Y_i) = \sum\limits_{i=1}^{k} Var(Y_i) = k\sigma^2$

And for $k \neq l$: $Cov(X_k, X_l) = Cov(\sum\limits_{i=1}^{k} Y_i, \sum\limits_{j=1}^{l} Y_j) = \min\{k, l\}\, \sigma^2$

Therefore, the covariance matrix of X is:

$$Cov(X) = \begin{bmatrix} \sigma^2 & \sigma^2 & \sigma^2 & \sigma^2 & \cdots & \sigma^2 \\ \sigma^2 & 2\sigma^2 & 2\sigma^2 & 2\sigma^2 & \cdots & 2\sigma^2 \\ \sigma^2 & 2\sigma^2 & 3\sigma^2 & 3\sigma^2 & \cdots & 3\sigma^2 \\ \sigma^2 & 2\sigma^2 & 3\sigma^2 & 4\sigma^2 & \cdots & 4\sigma^2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2 & 2\sigma^2 & 3\sigma^2 & 4\sigma^2 & \cdots & n\sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2 & 2 & \cdots & 2 \\ 1 & 2 & 3 & 3 & \cdots & 3 \\ 1 & 2 & 3 & 4 & \cdots & 4 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & 4 & \cdots & n \end{bmatrix}$$

Problem 4

(a)

The model $Y_i = I_F(i)\beta_F + I_M(i)\beta_M + \beta_1 x_i + e_i$ can be written in the matrix form: $Y = X\beta + e$, where the design matrix X and the parameter matrix β are:

$$X = \begin{bmatrix} I_F(1) & I_M(1) & x_1 \\ I_F(2) & I_M(2) & x_2 \\ \vdots & \vdots & \vdots \\ I_F(n) & I_M(n) & x_n \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_F \\ \beta_M \\ \beta_1 \end{bmatrix}$$

(b)

$$X^T X = \begin{bmatrix} I_F(1) & I_F(2) & \cdots & I_F(n) \\ I_M(1) & I_M(2) & \cdots & I_M(n) \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} I_F(1) & I_M(1) & x_1 \\ I_F(2) & I_M(2) & x_2 \\ \vdots & \vdots & \vdots \\ I_F(n) & I_M(n) & x_n \end{bmatrix} = \begin{bmatrix} \sum\limits_{i=1}^n I_F(i) & 0 & \sum\limits_{i=1}^n I_F(i)x_i \\ 0 & \sum\limits_{i=1}^n I_M(i) & \sum\limits_{i=1}^n I_M(i)x_i \\ \sum\limits_{i=1}^n I_F(i)x_i & \sum\limits_{i=1}^n I_M(i)x_i & \sum\limits_{i=1}^n x_i^2 \end{bmatrix}$$

(c)

For easier expression and computation, let's set the number of male students is m within the n total students. So we have: $\sum\limits_{i=1}^n I_F(i) = n - m$ and $\sum\limits_{i=1}^n I_M(i) = m$, thus:

$$X^T X = \begin{bmatrix} n-m & 0 & \sum\limits_{i=1}^n I_F(i)x_i \\ 0 & m & \sum\limits_{i=1}^n I_M(i)x_i \\ \sum\limits_{i=1}^n I_F(i)x_i & \sum\limits_{i=1}^n I_M(i)x_i & \sum\limits_{i=1}^n x_i^2 \end{bmatrix}$$

$$\Rightarrow \ |X^T X| = (n-m)\left(m\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n I_M(i)x_i\right)^2\right) + \left(\sum_{i=1}^n I_F(i)x_i\right)\left(0 - m\sum_{i=1}^n I_F(i)x_i\right)$$

$$= (n-m)m\sum_{i=1}^n x_i^2 - (n-m)\left(\sum_{i=1}^n I_M(i)x_i\right)^2 - m\left(\sum_{i=1}^n I_F(i)x_i\right)^2$$

$$\Rightarrow \ (X^T X)^{-1} = \frac{1}{|X^T X|}\begin{bmatrix} m\sum\limits_{i=1}^n x_i^2 - \left(\sum\limits_{i=1}^n I_M(i)x_i\right)^2 & \sum\limits_{i=1}^n I_M(i)x_i \sum\limits_{i=1}^n I_F(i)x_i & -m\sum\limits_{i=1}^n I_F(i)x_i \\ \sum\limits_{i=1}^n I_M(i)x_i \sum\limits_{i=1}^n I_F(i)x_i & (n-m)\sum\limits_{i=1}^n x_i^2 - \left(\sum\limits_{i=1}^n I_F(i)x_i\right)^2 & -(n-m)\sum\limits_{i=1}^n I_M(i)x_i \\ -m\sum\limits_{i=1}^n I_F(i)x_i & -(n-m)\sum\limits_{i=1}^n I_M(i)x_i & (n-m)m \end{bmatrix}$$

where $m = \sum\limits_{i=1}^n I_M(i)$, $n - m = \sum\limits_{i=1}^n I_F(i)$,

and $|X^T X| = (n-m)m\sum\limits_{i=1}^n x_i^2 - (n-m)\left(\sum\limits_{i=1}^n I_M(i)x_i\right)^2 - m\left(\sum\limits_{i=1}^n I_F(i)x_i\right)^2$

(d)

The least squares estimate of β is: $\hat{\beta} = (X^TX)^{-1}X^TY$, thus we can compute:

$$\begin{bmatrix}\hat{\beta_F}\\ \hat{\beta_M}\\ \hat{\beta_1}\end{bmatrix} = \hat{\beta} = (X^TX)^{-1}\begin{bmatrix} I_F(1) & I_F(2) & \cdots & I_F(n)\\ I_M(1) & I_M(2) & \cdots & I_M(n)\\ x_1 & x_2 & \cdots & x_n\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}$$

$$= \frac{1}{|X^TX|}\begin{bmatrix} m\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} I_M(i)x_i\right)^2 & \sum_{i=1}^{n} I_M(i)x_i \sum_{i=1}^{n} I_F(i)x_i & -m\sum_{i=1}^{n} I_F(i)x_i\\ \sum_{i=1}^{n} I_M(i)x_i \sum_{i=1}^{n} I_F(i)x_i & (n-m)\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} I_F(i)x_i\right)^2 & -(n-m)\sum_{i=1}^{n} I_M(i)x_i\\ -m\sum_{i=1}^{n} I_F(i)x_i & -(n-m)\sum_{i=1}^{n} I_M(i)x_i & (n-m)m \end{bmatrix}\begin{bmatrix}\sum_{i=1}^{n} I_F(i)y_i\\ \sum_{i=1}^{n} I_M(i)y_i\\ \sum_{i=1}^{n} x_iy_i\end{bmatrix}$$

$$= \frac{1}{|X^TX|}\left[\begin{matrix} m\left(\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} I_F(i)y_i - \sum_{i=1}^{n} x_iy_i \sum_{i=1}^{n} I_F(i)x_i\right) - \sum_{i=1}^{n} I_M(i)x_i \left(\sum_{i=1}^{n} I_M(i)x_i \sum_{i=1}^{n} I_F(i)y_i - \sum_{i=1}^{n} I_F(i)x_i \sum_{i=1}^{n} I_M(i)y_i\right)\\ (n-m)\left(\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} I_M(i)y_i - \sum_{i=1}^{n} x_iy_i \sum_{i=1}^{n} I_M(i)x_i\right) + \sum_{i=1}^{n} I_F(i)x_i \left(\sum_{i=1}^{n} I_M(i)x_i \sum_{i=1}^{n} I_F(i)y_i - \sum_{i=1}^{n} I_F(i)x_i \sum_{i=1}^{n} I_M(i)y_i\right)\\ (n-m)m\sum_{i=1}^{n} x_iy_i - m\sum_{i=1}^{n} I_F(i)x_i \sum_{i=1}^{n} I_F(i)y_i - (n-m)\sum_{i=1}^{n} I_M(i)x_i \sum_{i=1}^{n} I_M(i)y_i \end{matrix}\right.$$

Therefore, we obtain the explicit expressions for the least squares estimates of the following parameters:

$$\hat{\beta_F} = \frac{m\left(\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} I_F(i)y_i - \sum_{i=1}^{n} x_iy_i \sum_{i=1}^{n} I_F(i)x_i\right) - \sum_{i=1}^{n} I_M(i)x_i \left(\sum_{i=1}^{n} I_M(i)x_i \sum_{i=1}^{n} I_F(i)y_i - \sum_{i=1}^{n} I_F(i)x_i \sum_{i=1}^{n} I_M(i)y_i\right)}{(n-m)m\sum_{i=1}^{n} x_i^2 - (n-m)\left(\sum_{i=1}^{n} I_M(i)x_i\right)^2 - m\left(\sum_{i=1}^{n} I_F(i)x_i\right)^2}$$

$$\hat{\beta_M} = \frac{(n-m)\left(\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} I_M(i)y_i - \sum_{i=1}^{n} x_iy_i \sum_{i=1}^{n} I_M(i)x_i\right) + \sum_{i=1}^{n} I_F(i)x_i \left(\sum_{i=1}^{n} I_M(i)x_i \sum_{i=1}^{n} I_F(i)y_i - \sum_{i=1}^{n} I_F(i)x_i \sum_{i=1}^{n} I_M(i)y_i\right)}{(n-m)m\sum_{i=1}^{n} x_i^2 - (n-m)\left(\sum_{i=1}^{n} I_M(i)x_i\right)^2 - m\left(\sum_{i=1}^{n} I_F(i)x_i\right)^2}$$

$$\hat{\beta_1} = \frac{(n-m)m\sum_{i=1}^{n} x_iy_i - m\sum_{i=1}^{n} I_F(i)x_i \sum_{i=1}^{n} I_F(i)y_i - (n-m)\sum_{i=1}^{n} I_M(i)x_i \sum_{i=1}^{n} I_M(i)y_i}{(n-m)m\sum_{i=1}^{n} x_i^2 - (n-m)\left(\sum_{i=1}^{n} I_M(i)x_i\right)^2 - m\left(\sum_{i=1}^{n} I_F(i)x_i\right)^2}$$

where $m = \sum_{i=1}^{n} I_M(i)$, $n-m = \sum_{i=1}^{n} I_F(i)$

Problem 5

(a)

The least squares estimates of β_1 by fitting model (2) is:

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Thus its expected value when in fact model (1) is true is:

$$E(\hat{\beta_1}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(y_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(I_F(i)\beta_F + I_M(i)\beta_M)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(b)

Since there are roughly equal number of male and female students in the sample from this college, so: $\sum_{i=1}^{n} I_F(i) \approx \sum_{i=1}^{n} I_M(i) \approx \frac{n}{2}$

Plus, since girls at this college tend to have higher GPAs than boys in both high school and freshman year of college, and there are roughly equal number of male and female students in the sample, so: $\sum_{i=1}^{n} I_F(i)x_i > \sum_{i=1}^{n} I_M(i)x_i$ and $\beta_F > \beta_M$

Thus:

$$E(\hat{\beta_1}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(I_F(i)\beta_F + I_M(i)\beta_M)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\beta_F \sum_{i=1}^{n} I_F(i)x_i + \beta_M \sum_{i=1}^{n} I_M(i)x_i - \bar{x}(\beta_F \sum_{i=1}^{n} I_F(i) + \beta_M \sum_{i=1}^{n} I_M(i))}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\approx \frac{\beta_F \sum_{i=1}^{n} I_F(i)x_i + \beta_M \sum_{i=1}^{n} I_M(i)x_i - \frac{n\bar{x}}{2}(\beta_F + \beta_M)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{2\beta_F \sum_{i=1}^{n} I_F(i)x_i + 2\beta_M \sum_{i=1}^{n} I_M(i)x_i - (\beta_F + \beta_M)\sum_{i=1}^{n} x_i}{2\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{2\beta_F \sum_{i=1}^{n} I_F(i)x_i + 2\beta_M \sum_{i=1}^{n} I_M(i)x_i - (\beta_F + \beta_M)(\sum_{i=1}^{n} I_F(i)x_i + \sum_{i=1}^{n} I_M(i)x_i)}{2\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{(\beta_F - \beta_M)(\sum_{i=1}^{n} I_F(i)x_i - \sum_{i=1}^{n} I_M(i)x_i)}{2\sum_{i=1}^{n}(x_i - \bar{x})^2} > 0$$

Therefore, $E(\hat{\beta_1})$ will generally be positive.

(c)

From question (b), we have obtained that: $E(\hat{\beta_1}) > 0$. Since in this particular college the true value is: $\beta_1 = 0$, so: $E(\hat{\beta_1}) > \beta_1$, for which $\hat{\beta_1}$ is a positively biased estimator.

Therefore, using model (2) rather than the correct model (1) could result in a positively biased estimation of β_1, which means it will mislead us to conclude that the high school GPA has a positive effect on the college GPA when in fact there is no true effect.