# STA108 Discussion2

```
#data are from Problem 1.21 in textbook
X = c(1,0,2,0,3,1,0,1,2,0)
Y = c(16, 9, 17, 12, 22, 13, 8, 15, 19, 11)
n = length(X)
```

## 1. Parameter estimation

### Method 1.direct computation using R code

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$mse = s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$se(\hat{\beta}_1) = \sqrt{\frac{mse}{\sum_{i=1}^{n}(x_i - \bar{X})^2}}$$

$$se(\hat{\beta}_0) = \sqrt{mse(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2})}$$

```
b1hat = t(X-mean(X))%*%(Y-mean(Y))/sum((X-mean(X))^2)
#Note: sum((X-mean(X))^2) is the same as typing t(X-mean(X))%*%(X-mean(X))

b0hat = mean(Y) - b1hat*mean(X)

fit.y = b0hat[1] + b1hat[1]*X

mse = 1/(n-2)*sum((Y - fit.y)^2)

se.b1hat = sqrt(mse/sum((X-mean(X))^2))
se.b0hat = sqrt(mse*(1/n+ mean(X)^2/sum((X - mean(X))^2)))
```

### Method 2. Using R function lm()

```
fit = lm(Y~X)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.3    0.8    1.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
## X             4.0000     0.4690   8.528 2.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```

```
#coefficients estimated by model:
coef = fit$coefficients
b0hat = coef[1]
b1hat = coef[2]
MSE = summary(fit)$sigma^2
b0hat
```

```
## (Intercept)
##        10.2
```

```
b1hat
```

```
## X
## 4
```

se($\beta_0$) and se($\beta_1$): The column named "Std. Error" in the output table.

### Test statistic for testing hypothesis $H_0 : \beta_1 = \beta_{10}, H_a : \beta_1 \neq \beta_{10}$

Ex. $H_0 : \beta_1 = 1, H_a : \beta_1 \neq 1$

Test statistic:

$$t = \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)}$$

where $\hat{\beta}_1$ is the least square estimate of $\beta_1$, and se($\hat{\beta}_1$) is the standard error of the least square estimate.

- Under the null hypothesis $H_0$, $t \sim t_{n-2}$. Given $\alpha$, we compare $t_{1-\alpha/2,n-2}$ with $|t|$ and reject $H_0$ if $|t| > t_{\alpha/2,n-2}$.
- In the output table produced by R function lm() function, the column "t value" contains the test statistic for $H_0 : \beta_i = 0, H_a : \beta_i \neq 0$ (i = 0(intercept), 1(slope)).

### (1-alpha) ci for beta

$$\hat{\beta}_k \pm t_{n-2,1-\alpha/2} \times se(\hat{\beta}_k) \ (k = 0, 1)$$

```
alpha = 0.01
p = 1-alpha/2
lb.b1hat = b0hat - qt(p, df = n - 2)*se.b1hat
ub.b1hat = b0hat + qt(p, df = n - 2)*se.b1hat
```

# Proof of unbiasedness of least square estimates of regression coefficients.

The regression model is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where E$(\epsilon_i) = 0$, Var$(\epsilon_i) = \sigma^2, i = 1, \cdots, n$. The least square estimates can be obtained as

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}\end{aligned}$$

Show that E$(\hat{\beta}_1) = \beta_1$,E$(\hat{\beta}_0) = \beta_0$.

*Proof*:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i - \bar{Y}\sum_{i=1}^n X_i + n\bar{X}\bar{Y}}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\begin{aligned}\mathrm{E}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (X_i - \bar{X})\mathrm{E}(Y_i)}{\sum_{i=1}^n (X_i - \bar{X})^2} &= \frac{\sum_{i=1}^n (X_i - \bar{X})(\beta_0 + \beta_1 X_i)}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\beta_0 \sum_{i=1}^n X_i - n\bar{X}\beta_0 + \beta_1 \sum_{i=1}^n X_i^2 - \beta_1 \bar{X}\sum_{i=1}^n X_i}{\sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2} \\ (using \sum_{i=1}^n X_i = n\bar{X}) &= \frac{\beta_1(\sum_{i=1}^n X_i^2 - n\bar{X}^2)}{\sum_{i=1}^n X_i^2 - n\bar{X}^2} \\ &= \beta_1\end{aligned}$$

$$\mathrm{E}(\hat{\beta}_0) = \mathrm{E}(\bar{Y} - \hat{\beta}_1\bar{X}) = \mathrm{E}(\bar{Y}) - \beta_1\bar{X} = \beta_0$$

Here, we use E$(\hat{\beta}_1) = \beta_1$ when proving the unbiasedness of $\hat{\beta}_0$. And the last equality can be obtained from the model: E$(\bar{Y}) =$ E$(\frac{1}{n}\sum_{i=1}^n Y_i) =$ E$(\beta_0 + \beta_1\bar{X} + \frac{1}{n}\sum_{i=1}^n \epsilon_i) = \beta_0 + \beta_1\bar{X} +$ E$(\frac{1}{n}\sum_{i=1}^n \epsilon_i) = \beta_0 + \beta_1\bar{X}$.