Simple Linear Regression - Homework 1 - STAT 571A

Daniel Flick

Due: September 8, 2023

```r
install.packages('tinytex')
tinytex::install_tinytex()
```

I. Mathematical Derivations

(1) Show that Cov(e_i,e_j) = -$\frac{\sigma^2}{n}$

Cov(e_j) = E[(Y_i - $\bar{Y}$)(Y_j-$\bar{Y}$)] - E[(Y_i - $\bar{Y}$)]E[(Y_j - $\bar{Y}$)]

= E[Y_iY_j - $Y_i\bar{Y}$ - $Y_j\bar{Y}$ + $\bar{Y}^2$] - 0 (as E[Y_i] = E[Y_j] = E[$\bar{Y}$] = μ)

= E[Y_i]E[Y_j] - E[Y_i[$\frac{\sum_{k=1}^{n} Y_k}{n}$]] - E[$Y_j$[$\frac{\sum_{k=1}^{n} Y_k}{n}$]] + E[$\bar{Y}^2$]

We note: Var($\bar{Y}$) = $\frac{\sigma^2}{n}$ = E[$\bar{Y}^2$] - E[$\bar{Y}^2$], ==> $\frac{\sigma^2}{n}$ = E[$\bar{Y}^2$] - μ^2 ==> E[$\bar{Y}^2$] = $\frac{\sigma^2}{n}$ + μ^2

So we have:

μ^2 - $\frac{1}{n}$E[Y_i^2 + $\sum_{k\neq i} Y_iY_k$] - $\frac{1}{n}$E[Y_j^2 + $\sum_{k\neq j} Y_jY_k$] + ($\frac{\sigma^2}{n}$ + μ^2)

We note: Var(Y_i) = Var(Y_j) = E[Y_i^2] - μ^2 ==> E[Y_i^2] = E[Y_j^2] = σ^2 + μ^2

So, finally, we have:

μ^2 - $\frac{1}{n}$[σ^2 + μ^2 + (n-1)μ^2] - $\frac{1}{n}$[σ^2 + μ^2 + (n-1)μ^2] + $\frac{\sigma^2}{n}$ + μ^2

= μ^2 - $\frac{\sigma^2}{n}$ - $\frac{\mu^2}{n}$ - $\frac{(n-1)\mu^2}{n}$ - $\frac{\sigma^2}{n}$ - $\frac{\mu^2}{n}$ - $\frac{(n-1)\mu^2}{n}$ + $\frac{\sigma^2}{n}$ + μ^2

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

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

ALRM 1.5:

No. E[Y_i] = $\beta_0 + \beta_1 X_i$ + E[ϵ_i] since $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where E[$\epsilon_i$] = 0

Therefore, E[Y_i] = β_0 + $\beta_1 X_i$ + 0 = β_0 + $\beta_1 X_i$

(We note that β_0, β_1 and X_i are all constants, so their expected values and constant values are identical)

ALRM: 1.7

(a) No. While we know that σ^2=25, we do not know the underlying distribution with which this variance is associated. We could asssume the underlying distribution is approximately normal and then calculate the *approximate* probability of falling between 195 and 205 - but without knowing the underlying distribution governing the error terms, we cannot calculate an *exact* probability.

(b) Yes. Since we know the underlying distribution is N (0,σ^2=25) - and because we know β_0, β_1 and X=5, we can calculate that:

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i =$ 100 + 20(5) + $\epsilon_i =$ 200 + ϵ_i

The P(195 < Y_i < 205) = P (-5 < ϵ_i < 5) = P(-1 < z < 1) = 0.68268

ALRM: 1.12

(a) Observational. The subjects were not randomly assigned to a specific exercise time. Instead, the study merely made us of data wherein exercise time and frequency of colds were captured/analyzed.

(b) Increased exercise is associated with a lower freqeuncy of colds, but this does not mean exercise *causes* cold frequency to decrease. Association is not causation.

(c)

(1) Those who exercise more may also drink more water, and higher water intake may cause lower cold frequency.

(2) Those who exercise more may eat more fruits/vegetables, which may lower cold frequency.

(3) Those who exercise more may have unique genetic compositions that make them less prone to catching a cold.