Homework Assignment 1

Rafael Romero

January 24, 2025

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

1.) I do not agree because it should not be expectation of (Y_i), since this implies the expected value of Y_i which also does not involve the random error term and you instead get the regression function line

$$E(y_i) = \beta_0 + \beta_1 x_i$$

The proper form would be

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

for the simple linear regression model.

2.)

a.) The implication for the regression function if $\beta_0 = 0$ means that regression function simplifies to $y_i = \beta_1 X_i + \varepsilon_i$. The line will be completely linear based on β_1 and X. The regression function will have a slope that goes through y=0 (the intercept) when X = 0,therefore, it will pass through the origin (0,0). If $\beta_1 > 0$ the line will go upward as X increases and if $\beta_1 < 0$ the line will slope downward as X increases.

b.) The least squares estimate of B_1 with the assumption of $\beta_0 = 0$ is

$$\beta_1 = \frac{\sum_{i=1}^{n} x_1 y_i}{\sum_{i=1}^{n} x_i^2}$$

c.) using lm(y~x-1) will fit a model without the intercept.

3.)

a.) The implication for the regression function if $\beta_1 = 0$ means that regression function simplifies to $y_i = \beta_0 + \varepsilon_i$. The line will be horizontal at Y = β_0. This implies the regression function only depends on β_0 and not β_1. This indicates that there is no relationship between X and Y.

b.) The least squares estimate of B_0 with the assumption of $\beta_1 = 0$ is

$$\beta_0 = \bar{Y} \ where \ \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

c.) using lm(y~1) will fit a model with only the intercept.

4.) That does not imply that the summation of random errors in the model equals zero because residuals pertains the summation of the difference between the observed data and the predicted values. Random errors on the other hand are random and pertain to the parameters β_0 *and* β_1 which are unknown to us. The expectation of the summation of random errors does equal to zero because the expected value of the errors are assumed to be equal to zero and due to independence the summation also equals to zero.

5.) To minimize the sum of squared residuals, we start with the estimates for β_0 and β_1.

$$\hat{Y}_i = b_0 + b_1 X_i$$

The overall squared discrepancy between observed response Y_i and the fitted response $\hat{Y}_i$ is:

$$Q = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

We want to minimize Q and find b_0 and b_1.

We take the derivative of Q with respect to b_0 and b_1, and set them equal to zero. We get the normal equations:

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i) = 0$$

We solve the above equations, first for b_0:

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = nb_0 + nb_1 \bar{X} - n\bar{Y} = 0$$

This gives:

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Substituting b_0 into the second equation, we have:

$$\sum_{i=1}^{n} X_i (b_0 + b_1 X_i - Y_i)$$

$$= \sum_{i=1}^{n} X_i (\bar{Y} - b_1 \bar{X} + b_1 X_i - Y_i)$$

$$= \sum_{i=1}^{n} X_i (-b_1 \bar{X} + b_1 X_i) + \sum_{i=1}^{n} X_i (\bar{Y} - Y_i)$$

$$= b_1 \sum_{i=1}^{n} X_i (X_i - \bar{X}) + \sum_{i=1}^{n} X_i (\bar{Y} - Y_i)$$

$$= b_1 \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(X_i - \bar{X})(\bar{Y} - Y_i) = 0$$

Finally, solving for b_1:

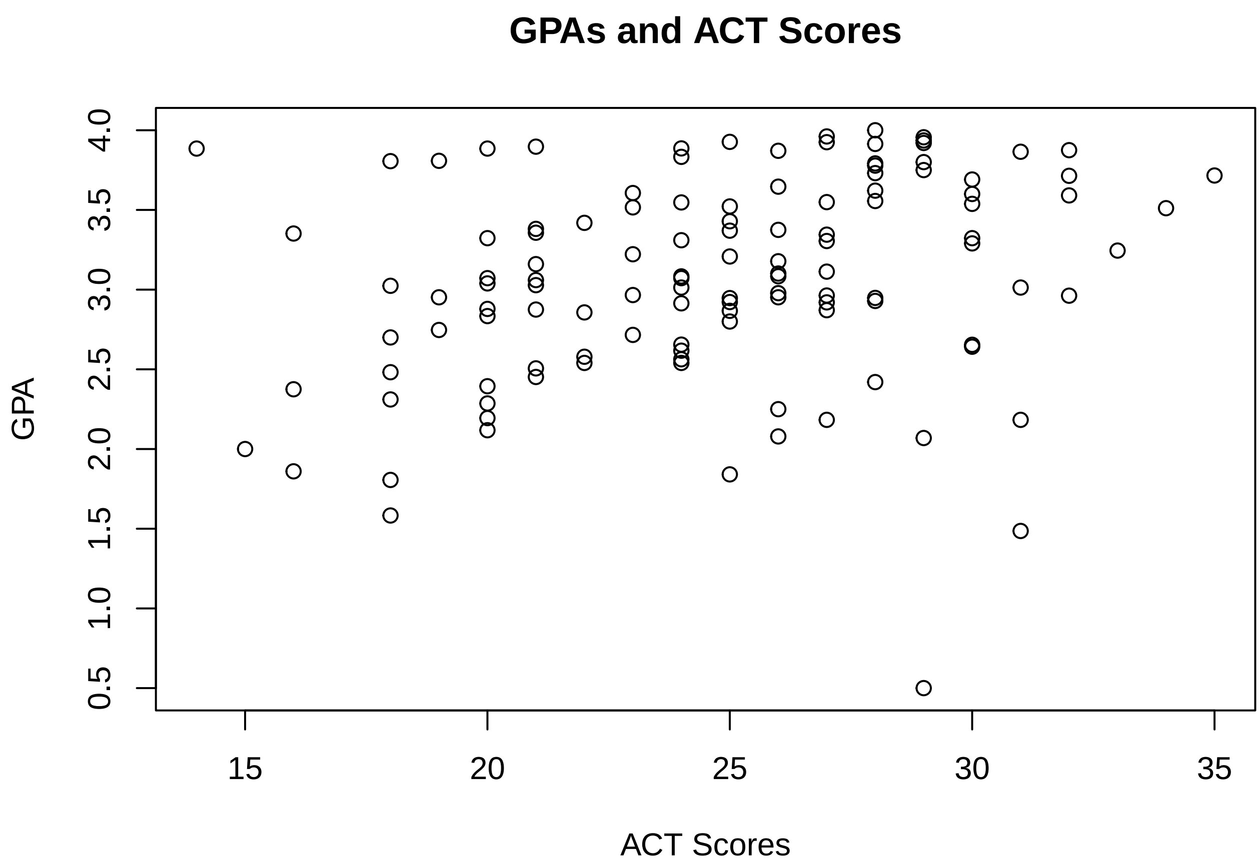
$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

6.)

a.) The predictor variable is ACT test scores. The response variable is a student's grade point average (GPA).

b.)

```
data <- read.table("data/GPA.txt", header = TRUE)
GPA <- data$GPA #response variable, dependent
ACT <- data$ACT #predictor variable, independent
plot(ACT, GPA, xlab = "ACT Scores", main = "GPAs and ACT Scores")
```



Based on the above plot, there does not seem to be a linear relationship between GPA scores and ACT Scores.

c.) The estmiated resgression function is

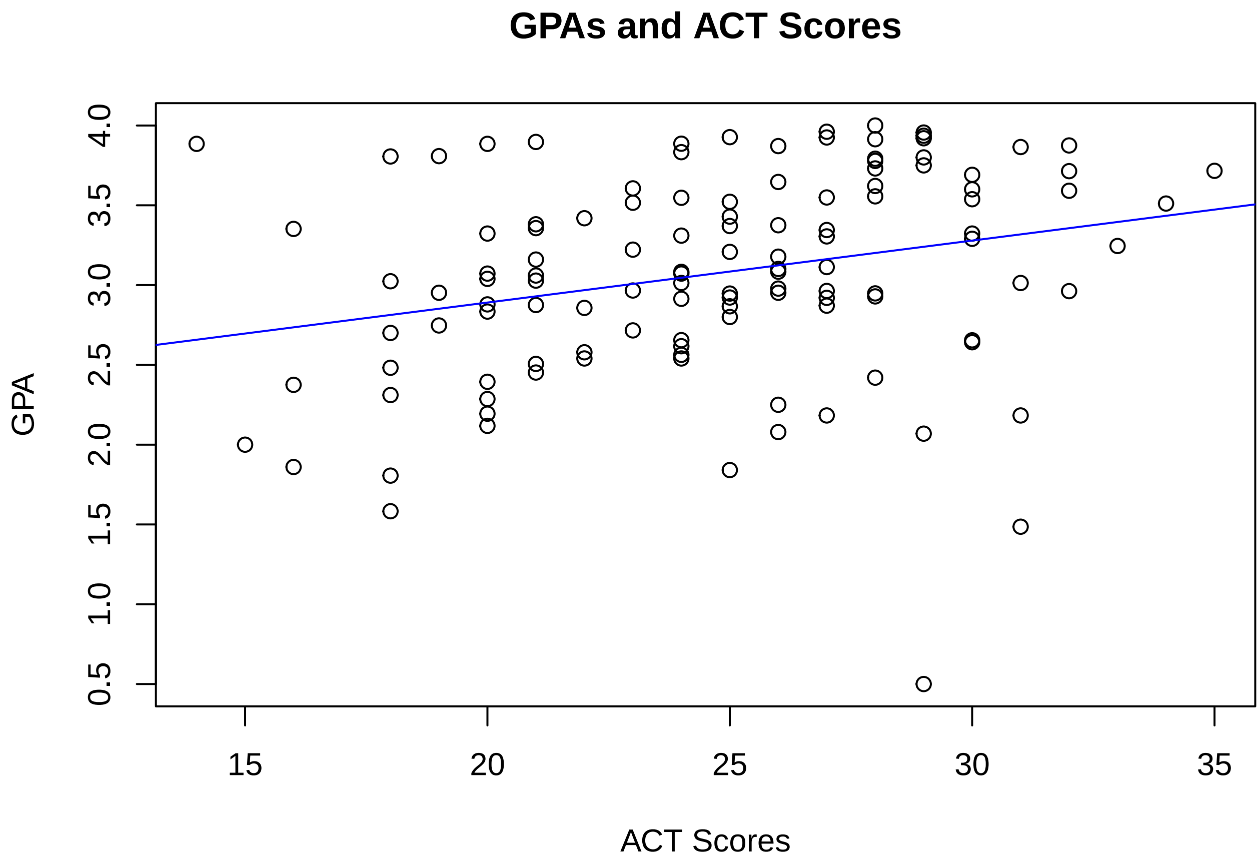
$$\hat{Y} = 2.1140 + 0.0388 * X$$

```r
model <- lm(GPA~ACT)
summary(model)
```

```
##
## Call:
## lm(formula = GPA ~ ACT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7400 -0.3383  0.0406  0.4406  1.2274
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1140     0.3209    6.59  1.3e-09 ***
## ACT           0.0388     0.0128    3.04   0.0029 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.623 on 118 degrees of freedom
## Multiple R-squared:  0.0726, Adjusted R-squared:  0.0648
## F-statistic: 9.24 on 1 and 118 DF,  p-value: 0.00292
```

d.)

```r
plot(ACT, GPA, xlab = "ACT Scores", main = "GPAs and ACT Scores")
abline(model, col= "blue")
```



The estimated regression function does not seem to fit the data well since it a lot of the data deviates either above or below the fitted line, so therefore the regression function does not accurately capture the relationship between GPA and ACT Scores.

e.) The point estimate of the change in the mean response when the entrance test score increases by one point is **3.279** GPA.

```
predict(model, newdata = data.frame(ACT=30))
```

```
##     1
## 3.279
```

f.) The point estimate of the change in the mean response when the entrance test score increases by one point can be found in the estimated regression function as b_1 and looking for **coefficients** then **estimate of ACT** we get 0.0388.

g.) The residuals are approximately zero.

```
residuals <- model$residuals

sum(residuals)
```

```
## [1] -1.138e-15
```

h.)

```
n <- nrow(data)
rss <- sum(residuals(model)^2)
sigma_squared <- rss/(n-2)
sigma <- sqrt(sigma_squared)
sigma_squared
```

```
## [1] 0.3883
```

```
sigma
```

```
## [1] 0.6231
```

$\sigma^2 = 0.3883\, GPA^2$ $\sigma = 0.6231\, GPA$

7.)

a.) The estimated regression line is

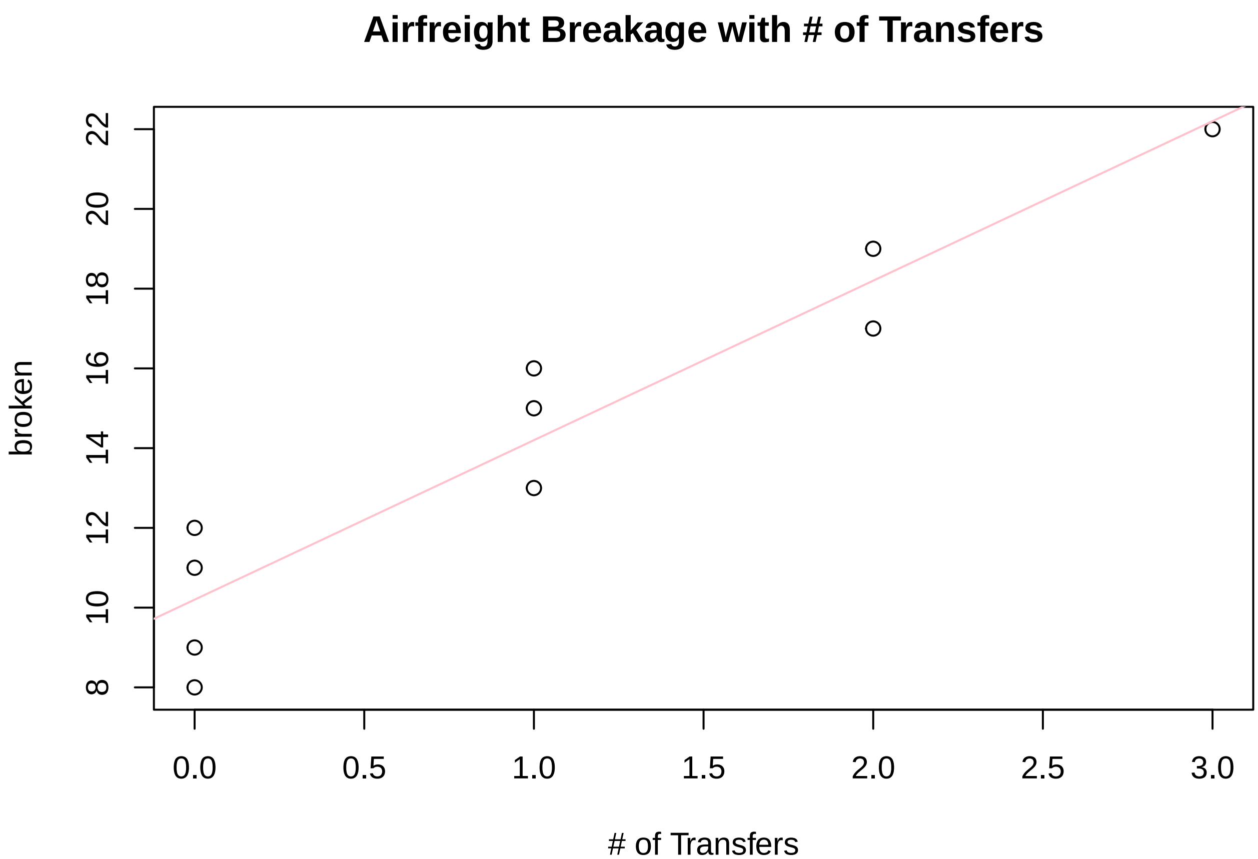
$$\hat{Y} = 10.200 + 4.000 * X$$

```
airfreight <- read.table("data/airfreight_breakage.txt", header = FALSE)

broken <- airfreight$V1 #response variable, dependent
transfers <- airfreight$V2 #predictor variable, independent
plot(transfers, broken, xlab = "# of Transfers", main = "Airfreight Breakage with # of Transfers")

linmodel <- lm(broken~transfers)
summary(linmodel)
```

```
##
## Call:
## lm(formula = broken ~ transfers)
##
## 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)   10.200      0.663   15.38  3.2e-07 ***
## transfers      4.000      0.469    8.53  2.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.48 on 8 degrees of freedom
## Multiple R-squared:  0.901,  Adjusted R-squared:  0.889
## F-statistic: 72.7 on 1 and 8 DF,  p-value: 2.75e-05
```

```
abline(linmodel, col = "Pink")
```



The linear regression function appears to give a good fit as it follows the trend of the real data.

b.) The point estimate of the expected number of broken ampules when 1 transfer is made is **14.2**

```
estimate <- 10.200 + 4*1
estimate
```

```
## [1] 14.2
```

```
predict(linmodel, newdata = data.frame(transfers=1))
```

```
##    1
## 14.2
```

c.) There is an **increase** in expected number of ampules broken when there are 2 transfers as compared to 1 transfer. 18.2 > 14.2

```r
predict(linmodel, newdata = data.frame(transfers=2))
```

```
##    1
## 18.2
```

d.) To verify that the fitted regression line goes through the point $(\bar{X}, \bar{Y})$ we must first calculate the mean of the transfers and broken ampules

```r
mean_transfers <- mean(transfers)
mean_broken <- mean(broken)
```

Next we must plug in our mean for transfers as our X and the predicted Y must equal to our mean of broken ampules

```r
predictedY <- predict(linmodel, newdata = data.frame(transfers=mean_transfers))
predictedY #with sample mean as the X
```

```
##    1
## 14.2
```

```r
mean_broken #mean of the broken ampulees
```

```
## [1] 14.2
```

Because our predicted Y and mean of broken ampulees are the same, we can concur that our fitted regression line goes through the point $(\bar{X}, \bar{Y})$