# Set_4

Oluwanifemi

2024-10-18

```r
library(ggplot2)
library(stats)
library(readxl)
```

```r
 # Read the .xlsx file
data <- read_excel("birth_dat.xlsx")

# Check the first few rows of the data
head(data)
```

```
## # A tibble: 6 x 5
##   Nation      Birthrate PerCapIncome PopFarms InfantMort
##   <chr>           <dbl>        <dbl>    <dbl>      <dbl>
## 1 Venezuela        46.4          392     0.4        68.5
## 2 Mexico           45.7          118     0.61       87.8
## 3 Ecuador          45.3           44     0.53      116.
## 4 Colombia         38.6          158     0.53      107.
## 5 Ceylon           37.2           81     0.53       71.6
## 6 Puerto Rico      35            374     0.37       60.2
```

```r
model <- lm(Birthrate ~ PerCapIncome+PopFarms+InfantMort, data = data)

summary(model)
```

```
##
## Call:
## lm(formula = Birthrate ~ PerCapIncome + PopFarms + InfantMort,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.232  -4.208  -1.710   4.699  18.006
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.553868   7.898060   0.703  0.48818
## PerCapIncome 0.006566   0.006239   1.052  0.30225
## PopFarms     9.104755  12.828869   0.710  0.48420
## InfantMort   0.242690   0.072845   3.332  0.00259 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.4 on 26 degrees of freedom
## Multiple R-squared:  0.4647, Adjusted R-squared:  0.403
## F-statistic: 7.525 on 3 and 26 DF,  p-value: 0.0008779
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

**n is the number of observation**

**p is number of predictor variables including intercept**

```
n <- nrow(data)
p <- 3
cat("n is", n, "and p is", p)
```

```
## n is 30 and p is 3
```

The underlying mean function for predicting the birth rate can be expressed as:

$$\text{BirthRate} = \beta_0 + \beta_1 \times \text{PerCapIncome} + \beta_2 \times \text{PopFarms} + \beta_3 \times \text{InfantMort} + \epsilon$$

```
team <- head(data, 3)  # Get the first 3 rows

# Convert the Birthrate column into a 3x1 data frame and print
birthrate_df <- data.frame(Birthrate = team$Birthrate)
print(birthrate_df)
```

```
##   Birthrate
## 1      46.4
## 2      45.7
## 3      45.3
```

```
# Print the rest of the columns (excluding 'Birthrate')
team_without_birthrate <- team[, !(names(team) %in% "Birthrate")]
print(team_without_birthrate)
```

```
## # A tibble: 3 x 4
##   Nation    PerCapIncome PopFarms InfantMort
##   <chr>            <dbl>    <dbl>      <dbl>
## 1 Venezuela          392     0.4        68.5
## 2 Mexico             118     0.61       87.8
## 3 Ecuador             44     0.53      116.
```

```
x_9 <- data[9, ]
x_9_without_birthrate <- x_9[, !(names(x_9) %in% "Birthrate")]
print(x_9_without_birthrate)
```

```
## # A tibble: 1 x 4
##   Nation        PerCapIncome PopFarms InfantMort
##   <chr>                <dbl>    <dbl>      <dbl>
## 1 United States         1723     0.12       27.2
```

```r
x2 <- data$PerCapIncome[1:3]
print(x2)
```

```
## [1] 392 118  44
```

```r
subset_data <- data.frame(
  Nation = c("Venezuela", "Mexico", "Ecuador"),
  PerCapIncome = c(392, 118, 44)
)


print(subset_data)
```

```
##      Nation PerCapIncome
## 1 Venezuela          392
## 2    Mexico          118
## 3   Ecuador           44
```

### Sum of Squared Errors (SSE)

The equation we are minimizing can be expressed as:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

### Linear Model

Given the linear model:

$$\hat{y}_i = \beta_0 + \beta_1 \times \text{PerCapIncome}_i + \beta_2 \times \text{PopFarms}_i + \beta_3 \times \text{InfantMort}_i$$

### Expanded Sum of Squared Errors

The sum of squared errors can be expanded as follows:

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 \times \text{PerCapIncome}_i + \beta_2 \times \text{PopFarms}_i + \beta_3 \times \text{InfantMort}_i)\right)^2$$

## Problem 2

## True or False Questions

### Question a

**Statement:** The equation $E[y_i|x_i] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ is no longer a linear model.

**Answer: False**

**Explanation/Justification:** The equation represents the regression model of $y_i$ given $x_i$ as a linear function of the predictor variables $x_{i1}, x_{i2}, \ldots, x_{ip}$. Despite the presence of $E$ (the expectation operator), this relationship is still linear in the parameters $\beta_0, \beta_1, \ldots, \beta_p$. A model is considered linear if it can be expressed as a linear combination of parameters, regardless of the distribution of $y_i$ or whether we take expectations. Thus, the equation describes a linear regression model.

---

### Question b

**Statement:** If

$$A = \begin{pmatrix} 4 & 22 & 3 \\ 5 & 11 & 2 \\ 6 & 0 & 1 \end{pmatrix}$$

then the transpose of $A$ is

$$A' = \begin{pmatrix} 4 & 5 & 6 \\ 22 & 11 & 0 \\ 3 & 2 & 1 \end{pmatrix}$$

**Answer: True**

**Explanation/Justification:** The first row of $A$ becomes the first column of $A'$, the second row becomes the second column, and the third row becomes the third column. Therefore, the transpose of $A$ is correctly given by:

$$A' = \begin{pmatrix} 4 & 5 & 6 \\ 22 & 11 & 0 \\ 3 & 2 & 1 \end{pmatrix}$$

---

### Question c

**Statement:** If

$$a' = \begin{pmatrix} 1.5 & 2 & 3 \end{pmatrix}$$

and

$$b' = \begin{pmatrix} 0 & 10 & 11 \end{pmatrix}$$

then the inner product of $a$ and $b$ is $\langle a, b \rangle = 53$.

**Answer: True**

**Explanation/Justification:** The inner product (dot product) of two vectors $a$ and $b$ is calculated as follows:

$$\langle a, b \rangle = a_1 b_1 + a_2 b_2 + a_3 b_3$$

Substituting the values from the vectors:

$$\langle a, b \rangle = (1.5 \cdot 0) + (2 \cdot 10) + (3 \cdot 11) = 0 + 20 + 33 = 53$$

---

## Summary

- **a)** False
- **b)** True
- **c)** True