Homework 3

2023-09-29

Problem 1

```
# Load necessary libraries
library(readxl)

# Read the data
data_1 <- read_excel("Income.xlsx")

# Fit a linear regression model
model_1 <- lm(Income ~ Participation, data = data_1)

# Summarize the model
summary(model_1)
```

```
##
## Call:
## lm(formula = Income ~ Participation, data = data_1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14993  -7644    698   7528  14812
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      24999       1324  18.889   <2e-16 ***
## Participation     1112       1768   0.629    0.531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8779 on 98 degrees of freedom
## Multiple R-squared:  0.00402,    Adjusted R-squared:  -0.006143
## F-statistic: 0.3955 on 1 and 98 DF,  p-value: 0.5309
```

Estimated equation for Income is: Income = 24999 + 1112 * Participation

Here, we do not see any statistically significant effect of Participating in the job training program on Income.

The intercept $\beta 0$ is 24999, which represents the estimated income for individuals who did not participate in the job training program (0). This is the baseline income estimate.

The coefficient for Participation $\beta 1$ tells us how much the income is expected to change for those who did participate compared to those who did not. The coefficient is 1112. Individuals who participated (1) in the job training program earned $1112 more than those who did not participate.

However, we do not see any Significance code near Participation Coefficient, which means, it is not statistically significant. Also, our p-value Pr(>|t|) 0.531 is greater than the typical significance level of 0.05, indicating the same result.

Problem 2

```r
# Read the data
data_2 <- read_excel("Investment.xlsx")

# Fit a linear regression model for Profitability
model_2 <- lm(Profitability ~ `Tax Policy`, data = data_2)

# Summarize the model
summary(model_2)
```

```
##
## Call:
## lm(formula = Profitability ~ `Tax Policy`, data = data_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4711.2 -2013.3  -196.8  2200.7  8931.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5764.3      262.8   21.94   <2e-16 ***
## `Tax Policy`   5072.9      371.6   13.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2628 on 198 degrees of freedom
## Multiple R-squared:  0.4848, Adjusted R-squared:  0.4822
## F-statistic: 186.3 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
# Fit a linear regression model for Investment
model_3 <- lm(Investment ~ `Tax Policy`, data = data_2)

# Summarize the model
summary(model_3)
```

```
##
## Call:
## lm(formula = Investment ~ `Tax Policy`, data = data_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49550 -22379   -930  23857  45673
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     53857       2714  19.845  < 2e-16 ***
## `Tax Policy`    15763       3838   4.107 5.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27140 on 198 degrees of freedom
## Multiple R-squared:  0.07851,    Adjusted R-squared:  0.07385
## F-statistic: 16.87 on 1 and 198 DF,  p-value: 5.859e-05
```

1.

Estimated equation for Profitability is: Profitability = 5764.3 + 5072.9 * Tax Policy

- `Intercept` $\beta 1$: The estimated annual profitability for small businesses that do not do the new tax policy (Tax Policy = 0) is $5,764.3.
- `Tax Policy Coefficient` $\beta 2$: It is 5072.9. This means that, expected effect of small businesses that do the new tax policy (Tax Policy = 1) and they have an estimated annual profitability that is $5,072.9 higher than those who are not affected by the policy.

The p-value Pr(>|t|) is extremely small (2e-16). And the p-value associated with the Tax Policy coefficient has Significance code showing '***' indicating that the effect of the new tax policy on profitability is statistically significant.

Estimated equation for Investment is: Investment = 53857 + 15763 * Tax Policy

- `Intercept` $\beta 1$: The estimated amount of capital investment made by small businesses do not do the new tax policy (Tax Policy = 0) is $53,857.
- `Tax Policy Coefficient` $\beta 2$: It is 15763. This means that, expected effect of small businesses that do the new tax policy (Tax Policy = 1) have an estimated capital investment that is $15,763 higher than those who are not affected by the policy.

The p-value Pr(>|t|) is very small (5.86e-05). And the p-value associated with the Tax Policy coefficient has Significance code showing '***' indicating that the effect of the new tax policy on investment is statistically significant.

Overall, both regressions demonstrate that the new tax policy has a statistically significant and positive impact on both profitability and investment decisions of small businesses. Small businesses subject to the policy appear to perform better in terms of profitability and make greater capital investments compared to those not affected by the policy.

2.

As mentioned above, in both cases, the small p-values (much smaller than the commonly used significance level of 0.05) suggest that the coefficients are highly significant, indicating that the new tax policy has a significant impact on both profitability and investment decisions of small businesses.

3.

The Average Treatment Effect on Profitability (ATP) represents the average difference in profitability between small businesses subject to the new tax policy (Treatment Group) and those not affected by it (Control Group). The new tax policy is associated with an estimated average increase in profitability of approximately 5072.9 units for small businesses that are subject to it, compared to those that are not affected by the policy.

4.

The Average Treatment Effect on Investment (ATI) represents the average difference in capital investment between small businesses subject to the new tax policy (Treatment Group) and those not affected by it (Control Group). The new tax policy is associated with an estimated average increase in investment of approximately 15763 units for small businesses that are subject to it, compared to those that are not affected by the policy.

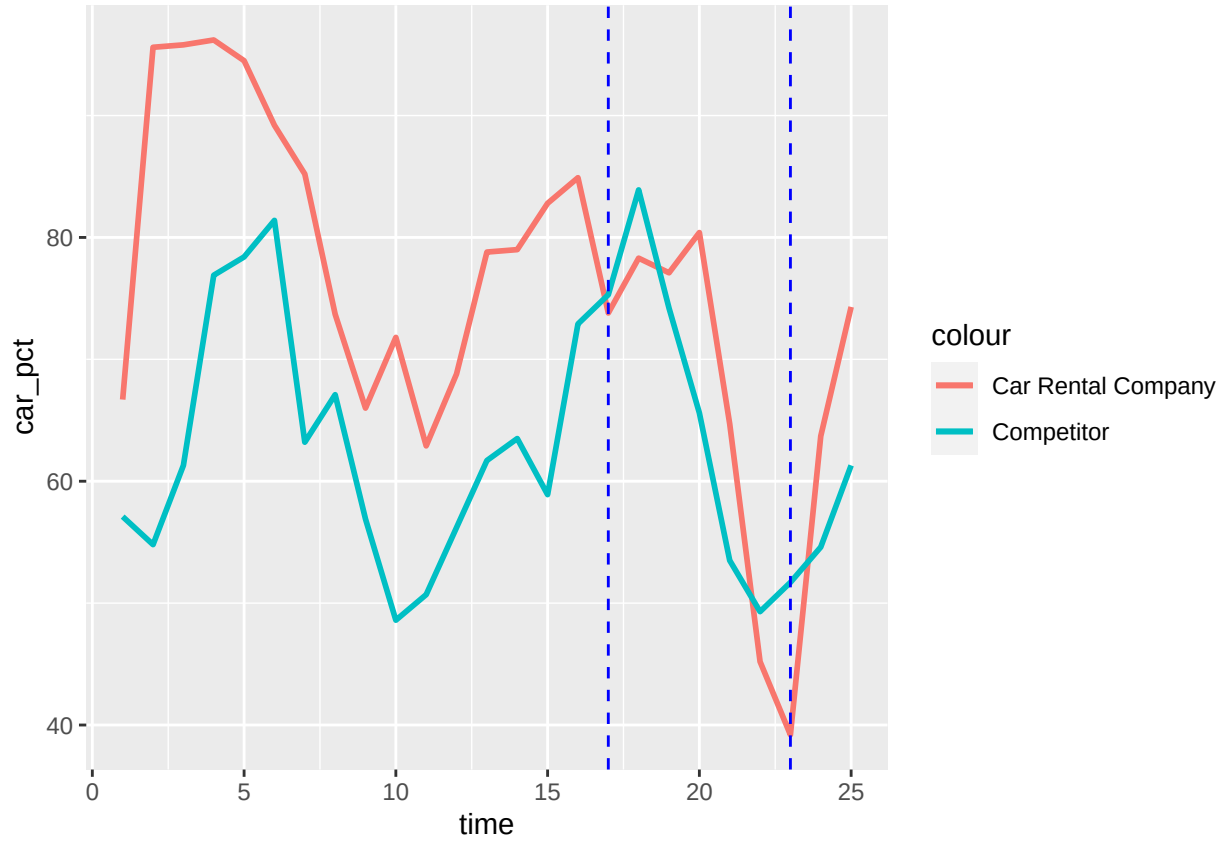
Problem 3

```r
# Load necessary libraries
library(ggplot2)
library(readr)

# Read the dataset
data_3 <- read_csv("cars.csv")
```

```
## Rows: 25 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (6): time, days, car_pct, comp_pct, repair, relprice
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Create a line plot of CAR_PCT and COMP_PCT over TIME
ggplot(data_3, aes(x = time)) +
  geom_line(aes(y = car_pct, color = "Car Rental Company"), linewidth = 1) +
  geom_line(aes(y = comp_pct, color = "Competitor"), linewidth = 1) +
  geom_vline(xintercept = 17, linetype = "dashed", color = "blue") +
  geom_vline(xintercept = 23, linetype = "dashed", color = "blue")
```




```r
  # Vertical line for the start of the repair period
  labs(title = "Occupancy Rates Over Time",
```

```
       x = "Time (Months)",
       y = "Occupancy Rate") +
  scale_color_manual(values = c( "Competitor" = "turquoise")) +
  theme_minimal()
```

```
## NULL
```

a.

Car Rental Company has higher occupancy before the repair period as we see left from the first blue line (17th month). However, during the repair period which is between the two blue lines, we see a dramatic decrease up to 23th month (which was the last one of all 7 months of correcting the defects started from 17th). After that part it is increasing again up to 25th month.

b.

```
# Load necessary libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Separate data for non-repair and repair periods
non_repair_data <- data_3 %>%
  filter(time <= 16 | time> 23)

repair_data <- data_3 %>%
  filter(time > 16 & time <=23)

# Calculate average occupancies during non-repair period
avg_CAR_0 <- mean(non_repair_data$car_pct)
avg_COMP_0 <- mean(non_repair_data$comp_pct)

# Calculate average occupancies during repair period
avg_CAR_1 <- mean(repair_data$car_pct)
avg_COMP_1 <- mean(repair_data$comp_pct)

# Calculate the difference in average occupancies during the non-repair period
diff_occupancy <- avg_CAR_0 - avg_COMP_0

# Estimate the company's occupancy rate that have maintained
# the same relative difference in occupancy if there had been no repairs.
avg_CAR_1_asterix <- avg_COMP_1 + (avg_CAR_0 - avg_COMP_0)
```

```r
# Calculate the "simple" difference estimate of lost occupancy
lost_occupancy <- avg_CAR_1_asterix - avg_CAR_1

# Calculate the amount of revenue lost during the seven-month period
lost_revenue <- lost_occupancy * 215 * 56.61
```

Average occupancy rate during non-repair period for the company:

```r
avg_CAR_0
```

```
## [1] 79.43889
```

Average occupancy rate during non-repair period the competitor:

```r
avg_COMP_0
```

```
## [1] 62.52778
```

Average occupancy rate during repair period the company:

```r
avg_CAR_1
```

```
## [1] 65.52857
```

Average occupancy rate during repair period the competitor:

```r
avg_COMP_1
```

```
## [1] 64.78571
```

Difference in average occupancies during the non-repair period:

```r
diff_occupancy
```

```
## [1] 16.91111
```

Estimated occupancy rate during repair without repairs fot the company:

```r
avg_CAR_1_asterix
```

```
## [1] 81.69683
```

Simple difference estimate of lost occupancy:

```r
lost_occupancy
```

```
## [1] 16.16825
```

Estimated revenue lost during the seven-month period:

```
lost_revenue
```

```
## [1] 196786.2
```

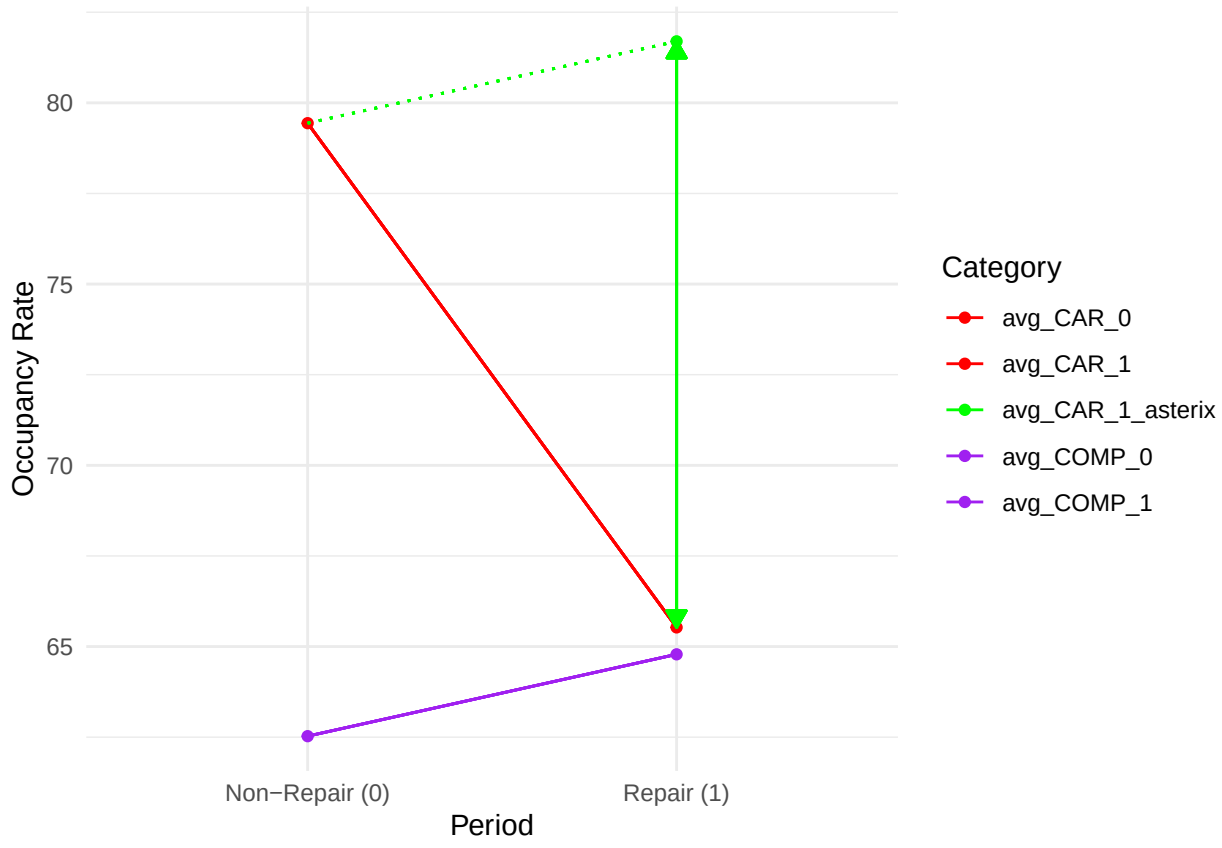
c.

```
library(ggplot2)

# Create a data frame with the specified values
occupancy_data <- data.frame(
  Category = c("avg_CAR_0", "avg_CAR_1", "avg_CAR_1_asterix", "avg_COMP_0", "avg_COMP_1"),
  Value = c(avg_CAR_0, avg_CAR_1, avg_CAR_1_asterix, avg_COMP_0, avg_COMP_1),
  Period = factor(c("Non-Repair (0)", "Repair (1)", "Repair (1)", "Non-Repair (0)",
                    "Repair (1)"))
)

# Create a scatterplot
ggplot(occupancy_data, aes(x = as.factor(Period), y = Value, color = Category)) +
  geom_point() +
  geom_line(aes(group = Category)) +
  geom_segment(aes(x = "Non-Repair (0)", xend = "Repair (1)", y = avg_CAR_0,
                   yend = avg_CAR_1),
               color = "red") +
  geom_segment(aes(x = "Non-Repair (0)", xend = "Repair (1)", y = avg_CAR_0,
                   yend = avg_CAR_1_asterix),
               linetype = "dotted", color = "green") +
  geom_segment(aes(x = "Non-Repair (0)", xend = "Repair (1)", y = avg_COMP_0,
                   yend = avg_COMP_1),
               color = "purple") +
  geom_segment(aes(x = "Repair (1)", xend = "Repair (1)", y = avg_CAR_1_asterix,
                   yend = avg_CAR_1),
               arrow = arrow(type = "closed", length = unit(0.1, "inches")),
               color = "green") +
  geom_segment(aes(x = "Repair (1)", xend = "Repair (1)", y = avg_CAR_1,
                   yend = avg_CAR_1_asterix),
               arrow = arrow(type = "closed", length = unit(0.1, "inches")),
               color = "green") +
  labs(
    x = "Period",
    y = "Occupancy Rate",
    color = "Category"
  ) +
  scale_color_manual(values = c("avg_CAR_0" = "red", "avg_CAR_1" = "red",
                                "avg_CAR_1_asterix" = "green",
                                "avg_COMP_0" = "purple",
                                "avg_COMP_1" = "purple")) +
  theme_minimal()
```

```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```



d.

```
# Define the regression model
model_4 <- lm(car_pct ~ comp_pct + relprice + repair, data = data_3)

# Obtain the least squares estimates of the parameters
summary(model_4)
```

```
##
## Call:
## lm(formula = car_pct ~ comp_pct + relprice + repair, data = data_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.511  -4.999  -2.349   4.142  16.408
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  135.8206    46.5265   2.919   0.0082 **
## comp_pct       0.5783     0.2068   2.797   0.0108 *
## relprice    -122.3011    49.7389  -2.459   0.0227 *
## repair       -20.1898     4.1963  -4.811 9.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 8.525 on 21 degrees of freedom
## Multiple R-squared:  0.693,  Adjusted R-squared:  0.6491
## F-statistic:  15.8 on 3 and 21 DF,  p-value: 1.324e-05
```

- `Intercept`: The intercept represents the estimated company's car occupancy rate when all other variables (comp_pct, relprice, and repair) are zero.
- `comp_pct`: The coefficient for comp_pct is 0.5783. It represents the change in the company's car occupancy rate associated with a one-unit change in competitors' occupancy, holding all other variables constant. The positive sign of this coefficient suggests that as competitors' occupancy increases, the company's occupancy rate also tends to increase. This effect is statistically significant (p-value = 0.0108), indicating that competitors' occupancy has an impact on company's occupancy.
- `Relprice`: The coefficient for relprice is -122.3011. It represents the change in company's car occupancy rate associated with a one-unit change in relative price (relprice), holding all other variables constant. The negative sign of this coefficient suggests that as relative price increases (becomes more expensive), the company's occupancy rate tends to decrease. This effect is also statistically significant (p-value = 0.0227), indicating that relative price has an impact on company's occupancy.
- `Repair`: The coefficient for repair is -20.1898. It represents the change in company's car occupancy rate during the repair period compared to the non-repair period. The negative sign indicates that company's occupancy tends to decrease during the repair period. This effect is highly statistically significant (p-value < 0.0001), suggesting that the repair period has a significant negative impact on company's occupancy.

In summary:

- The intercept represents the baseline car occupancy rate when all other variables are zero.
- Competitors' occupancy has a positive and significant impact on company's occupancy.
- Relative price has a negative and significant impact on company's occupancy.
- The repair period has a negative and highly significant impact on company's occupancy.

e.

```r
# Given coefficient for repair
coef_repair <- coef(model_4)["repair"]

# Estimate the revenue lost during the repair period
lost_revenue_repair<- coef_repair * 215 * 56.61

# The "simple" difference estimate of lost occupancy
lost_occupancy <- avg_CAR_1_asterix - avg_CAR_1

# Compare the two estimates
if (lost_revenue_repair > lost_occupancy) {
  cat("Revenue lost using repair is greater than the simple estimate. \n")
} else if (lost_revenue_repair < lost_occupancy) {
  cat("Simple estimate of lost occupancy is greater than revenue lost using repair. \n")
} else {
  cat("Both estimates are equal. \n")
}
```

```
## Simple estimate of lost occupancy is greater than revenue lost using repair.
```

```
# Calculate the standard error of the coefficient for repair
se_repair <- summary(model_4)$coefficients["repair", "Std. Error"]

# Calculate the t-statistic for a 95% confidence interval (assuming normal distribution)
t_stat <- qt(0.975, df = length(data_3$repair) - 4)

# Calculate the margin of error for the coefficient of repair
margin_of_error <- t_stat * se_repair

# Calculate the lower and upper bounds of the 95% confidence interval
lower_bound <- coef_repair - margin_of_error
upper_bound <- coef_repair + margin_of_error

# Estimate the revenue loss during the repair period using the lower and upper bounds
upper_revenue_estimate <- abs(lower_bound) * 215 * 56.61
lower_revenue_estimate <- abs(upper_bound) * 215 * 56.61
```

95% Confidence Interval for Lost Revenue (Repair Period)

Lower Bound:

```
lower_revenue_estimate
```

```
##   repair
## 139519.2
```

Upper Bound:

```
upper_revenue_estimate
```

```
##   repair
## 351946.5
```

Estimated revenue loss from part (b):

```
lost_revenue
```

```
## [1] 196786.2
```

Yes, as we see, it is within the estimated interval.

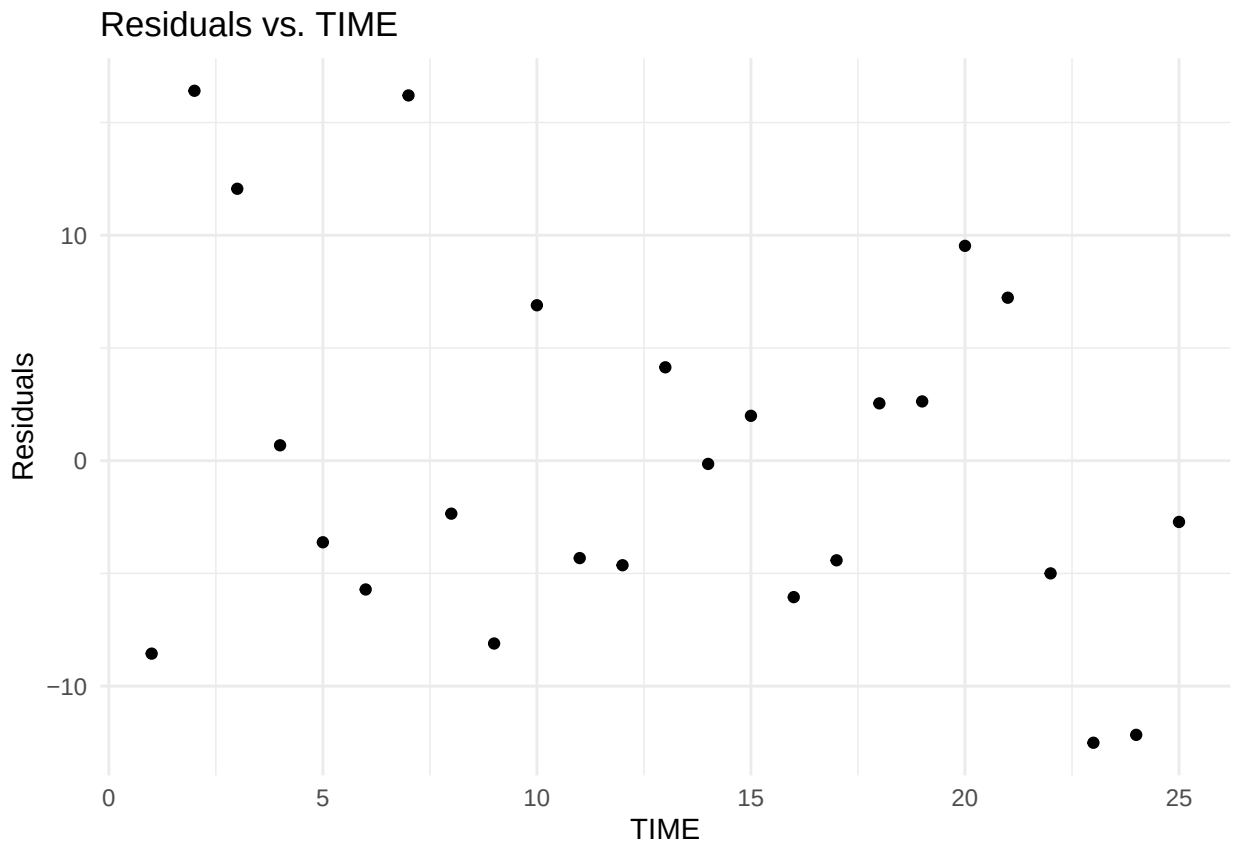
d.

```
# Obtain the residuals from the linear regression model
# (assuming you've already fitted the model)
residuals <- residuals(model_4)

# Create a data frame with TIME and residuals
residual_data <- data.frame(TIME = data_3$time, Residuals = residuals)
```

```r
# Load necessary libraries for plotting
library(ggplot2)

# Create a scatterplot of residuals against TIME
ggplot(residual_data, aes(x = TIME, y = Residuals)) +
  geom_point() +
  labs(title = "Residuals vs. TIME",
       x = "TIME",
       y = "Residuals") +
  theme_minimal()
```



No, I do not see any obvious patterns.