# STAT 5385: Homework 3

Willliam Ofosu Agyapong

22/02/2022

## Problem 3.4: Copier maintenance (Problem 1.20)

To begin, we first import the `copier mainternance` dataset and fit a simple regression model relating service time in minutes to number of copiers serviced. The estimated regression coefficients with associated measures for the model is presented in the table that follows.

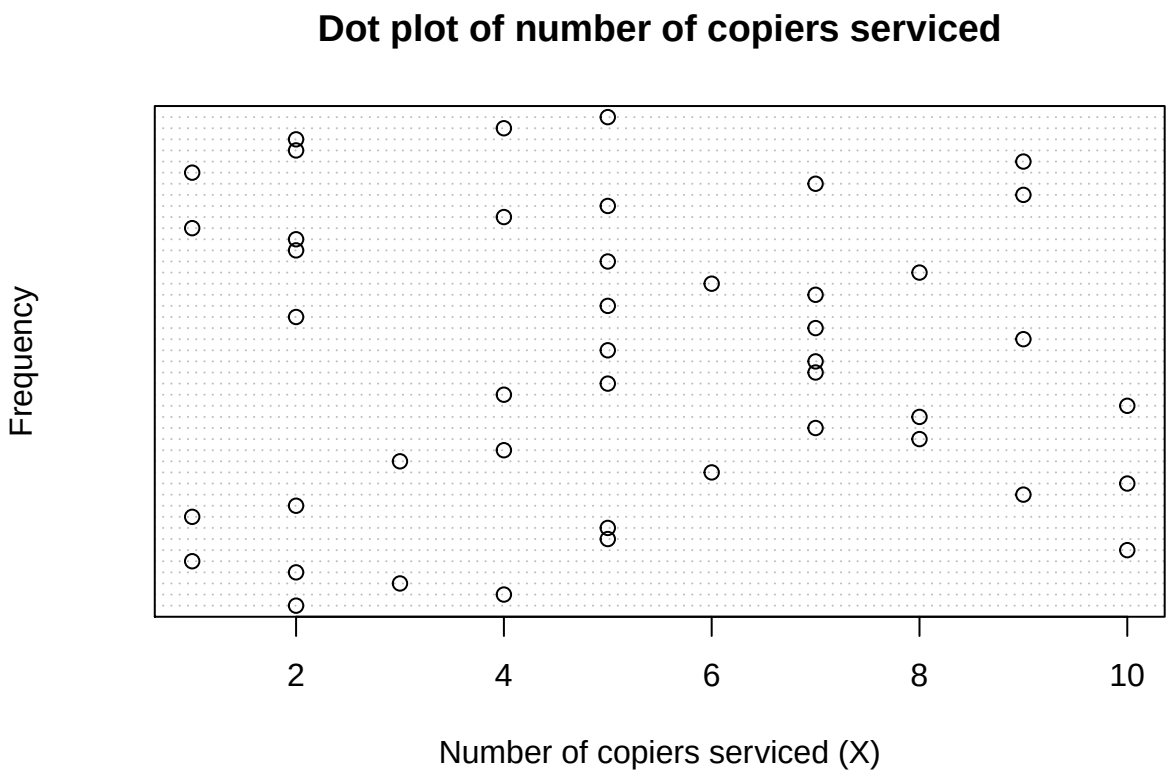
```
# Importing the updated copier maintenance dataset
copier <- read.table("../Data Sets/Chapter  3 Data Sets/CH03PR04.txt",
                     header = F,
                     col.names = c("service_time", "copiers_serviced", "X2", "X3")
                   )

# Fit a linear regression model for later use.
mdl_copier <- lm(service_time~copiers_serviced, data =  copier)

mdl_copier %>%
  tidy(conf.int = TRUE, conf.level = 0.95) %>%
  mutate(term = c("$\\beta_0$","$\\beta_1$")) %>%
  kable(caption = "Parameter Estimates with 95% Confidence Interval")
```

Table 1: Parameter Estimates with 95% Confidence Interval

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| $\beta_0$ | -0.5802 | 2.8039 | -0.2069 | 0.8371 | -6.235 | 5.074 |
| $\beta_1$ | 15.0352 | 0.4831 | 31.1233 | 0.0000 | 14.061 | 16.009 |

### Part (a)

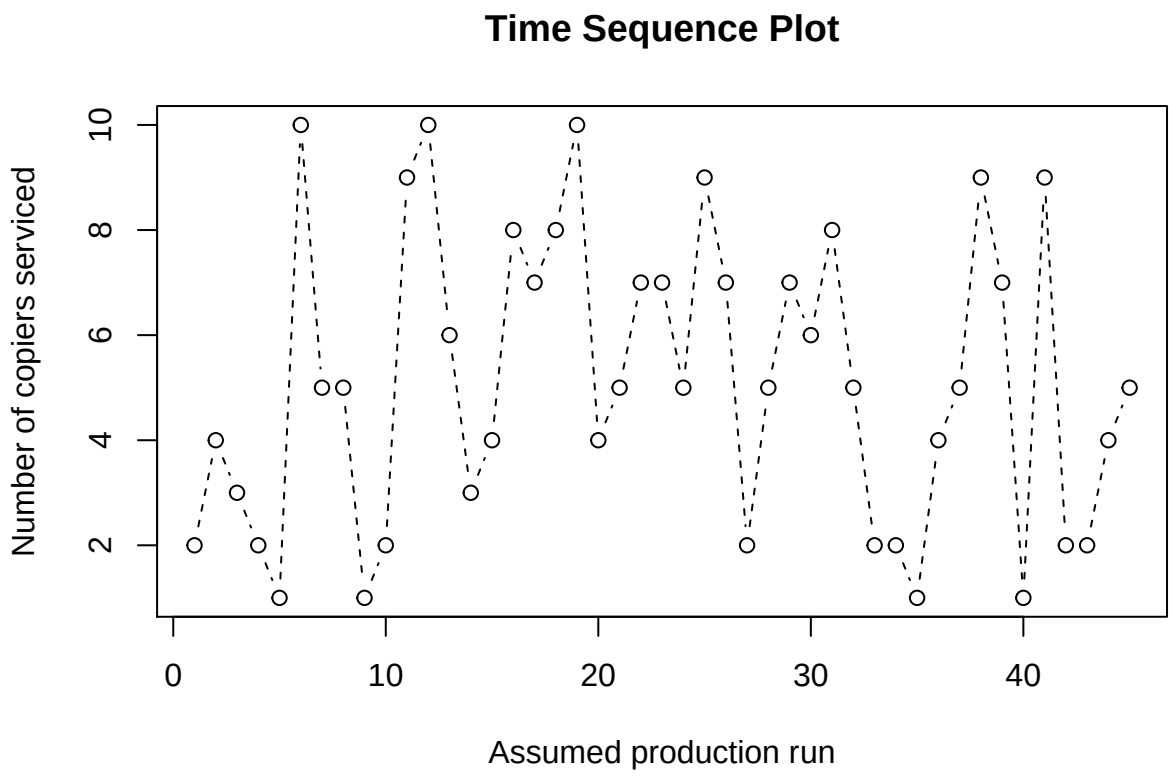
```
# create a dot plot of X, number of copiers serviced
dotchart(copier$copiers_serviced, xlab="Number of copiers serviced (X)",
         ylab = "Frequency", main = "Dot plot of number of copiers serviced")
```




The above dot plot tells us that the minimum and maximum number of copiers serviced are 1 and 10, respectively. It can also be seen that in a number of cases several copiers took the same number of minutes to be serviced. Between the minimum and maximum values, the number of copiers serviced appear to spread evenly throughout, with no apparent outlying cases.

## Part (b)

```
# Prepare a time plot
plot(copier$copiers_serviced, type = "b", lty = 2, xlab = "Assumed production run", ylab = "Number of copi
title("Time Sequence Plot")
```




The points in the plot are connected to show more effectively the time sequence. Here, the sequence plot shows no special pattern. We cannot make any general conclusion about the number of copiers serviced and the production runs. This suggests that the number of copiers serviced are, perhaps, not correlated with time.

### Part (c): Stem-and-leaf plot of residuals

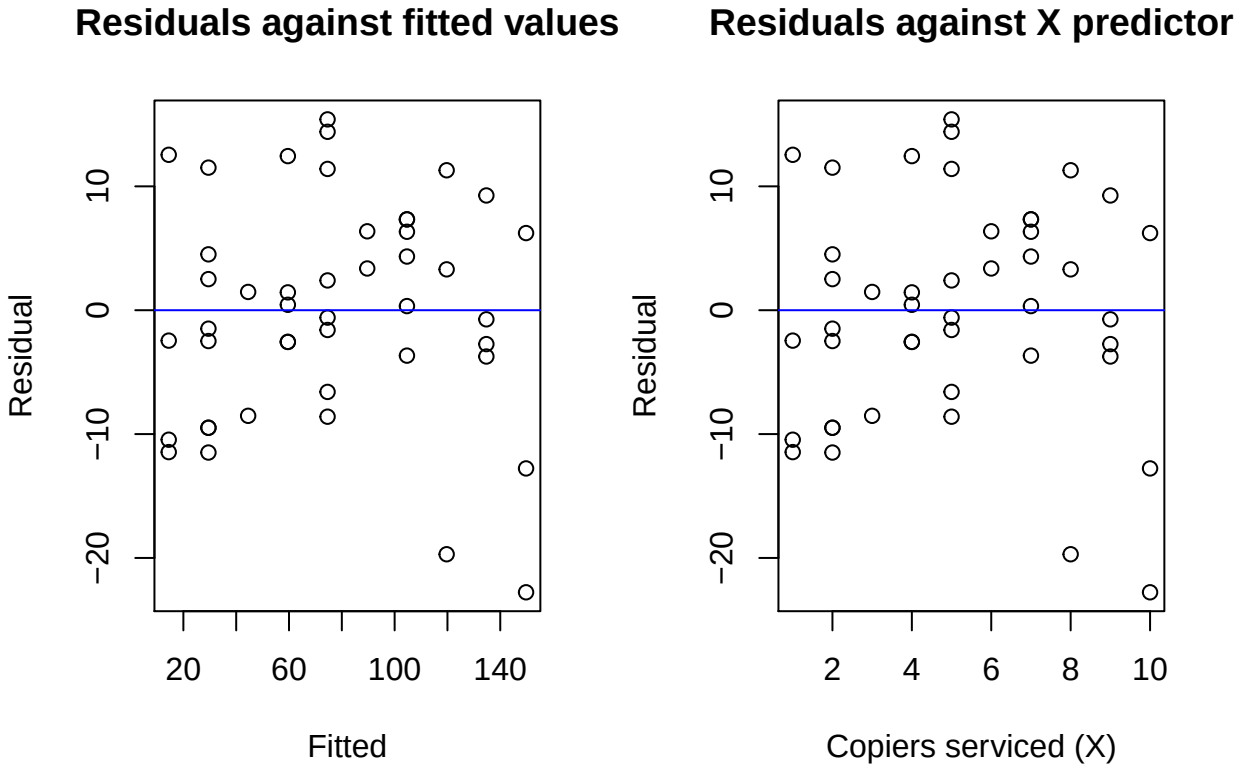
```
# stem-and-leaf plot
stem(resid(mdl_copier), scale=3)
```

```
##
##   The decimal point is at the |
##
##   -22 | 8
##   -20 |
##   -18 | 7
##   -16 |
##   -14 |
##   -12 | 8
##   -10 | 555
##    -8 | 5565
##    -6 | 6
##    -4 |
##    -2 | 7776655
##    -0 | 6576
##     0 | 3445
##     2 | 4534
##     4 | 35
##     6 | 23433
##     8 | 3
##    10 | 345
##    12 | 45
##    14 | 44
```

This plot suggests that there are about the same number of residuals above and below 0. Most of the residuals lie between -14 and 14 which makes the distribution of residuals appear roughly symmetric with that range. However, for those two extremely low residuals (-22.8 and -18.7), the distribution can be said to be negatively skewed.

## Part (d)

```
par(mfrow = c(1,2))
# residual vs. fitted plot
plot(fitted(mdl_copier), resid(mdl_copier), main = "Residuals against fitted values",
     xlab = "Fitted", ylab = "Residual"
     )
abline(0,0, col="blue")

plot(copier$copiers_serviced, resid(mdl_copier), main = "Residuals against X predictor",
     xlab = "Copiers serviced (X)", ylab = "Residual")
abline(0,0, col="blue")
```




Interestingly, the two plots provide exactly the same information. Meaning, they can both be used to study departures of the regression function from a **linear relationship** as well as **nonconstancy of error variance**. Standardizing the residuals in these plots can also help us detect the **presence of outlirs**.

From the last plot, one can see that the residuals fluctuate randomly from 0 across the number of copiers serviced. This suggests that the error variance is reasonably the same for all levels of copiers serviced. In other words, the error terms have constant variance.

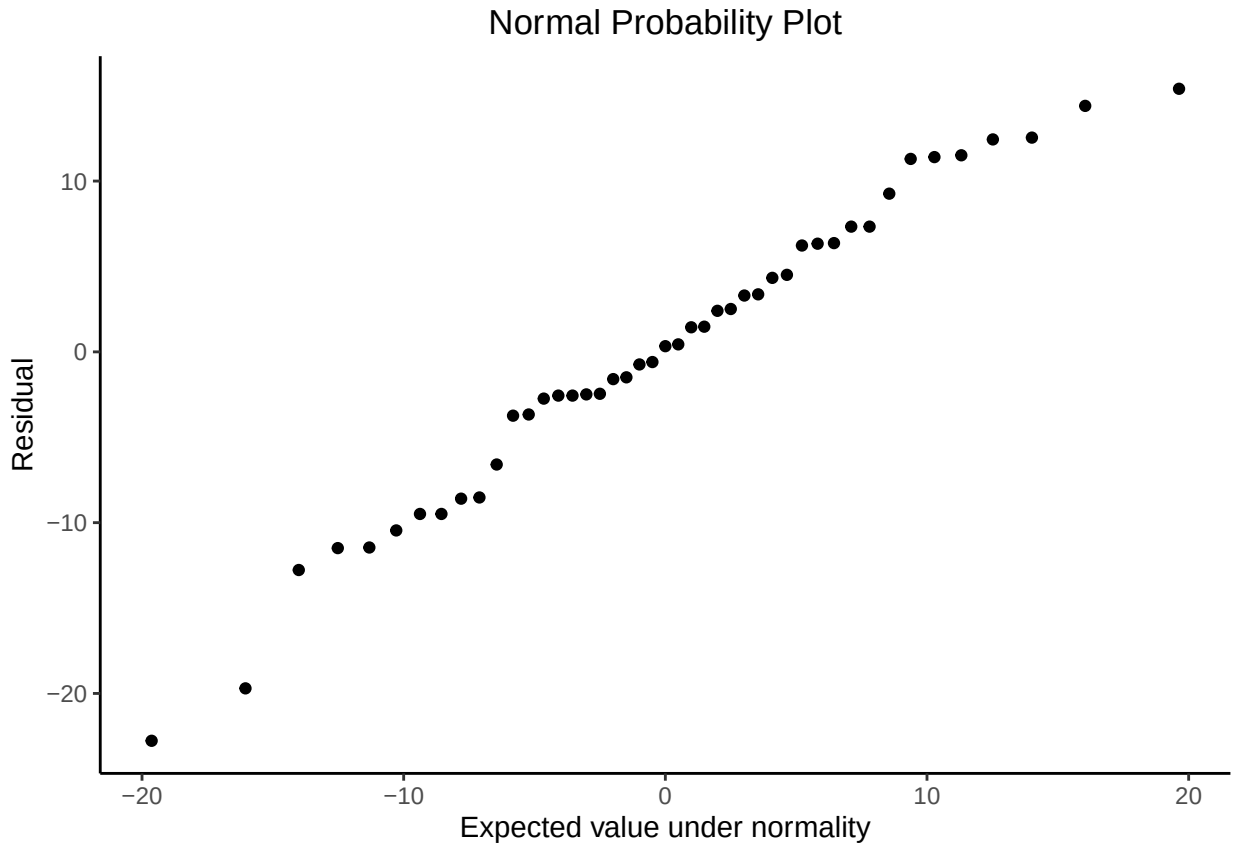
## Part (e)

```
# Obtain the ordered residuals in ascending order
resid_df <- data.frame(resid_ord = sort(resid(mdl_copier)),
                       k = 1:nrow(copier))
# computing the expected values
MSE_sqrt <- mdl_copier %>% glance() %>% pull(sigma)
```

```
resid_df <- resid_df %>%
  mutate(expected_value = MSE_sqrt *(qnorm((k-0.375)/(nrow(copier)+0.25))))

# view few rows
kable(head(resid_df), col.names = c("Ordered residual", "Rank(k)", "Expected value"))
```

| | Ordered residual | Rank(k) | Expected value |
|---|---|---|---|
| 19 | -22.77 | 1 | -19.63 |
| 16 | -19.70 | 2 | -16.05 |
| 6 | -12.77 | 3 | -14.01 |
| 27 | -11.49 | 4 | -12.52 |
| 35 | -11.46 | 5 | -11.31 |
| 9 | -10.46 | 6 | -10.28 |

```
# Obtain the expected values under normality using suggested procedure from the textbook
ggplot(resid_df, aes(expected_value, resid_ord)) +
  geom_point() +
  labs(title = "Normal Probability Plot",
       x = "Expected value under normality",
       y = "Residual") +
  theme(plot.title = element_text(hjust = 0.5))
```




The above normal probability plot suggests that the distribution of the error terms does not depart substantially from normality, since majority of the points appear to fall along a straight line.

```
# compute the correlation between ordered residuals and the expected values
(obs_cor_coef <- cor(resid_df$resid_ord, resid_df$expected_value))
```

```
## [1] 0.9891
```

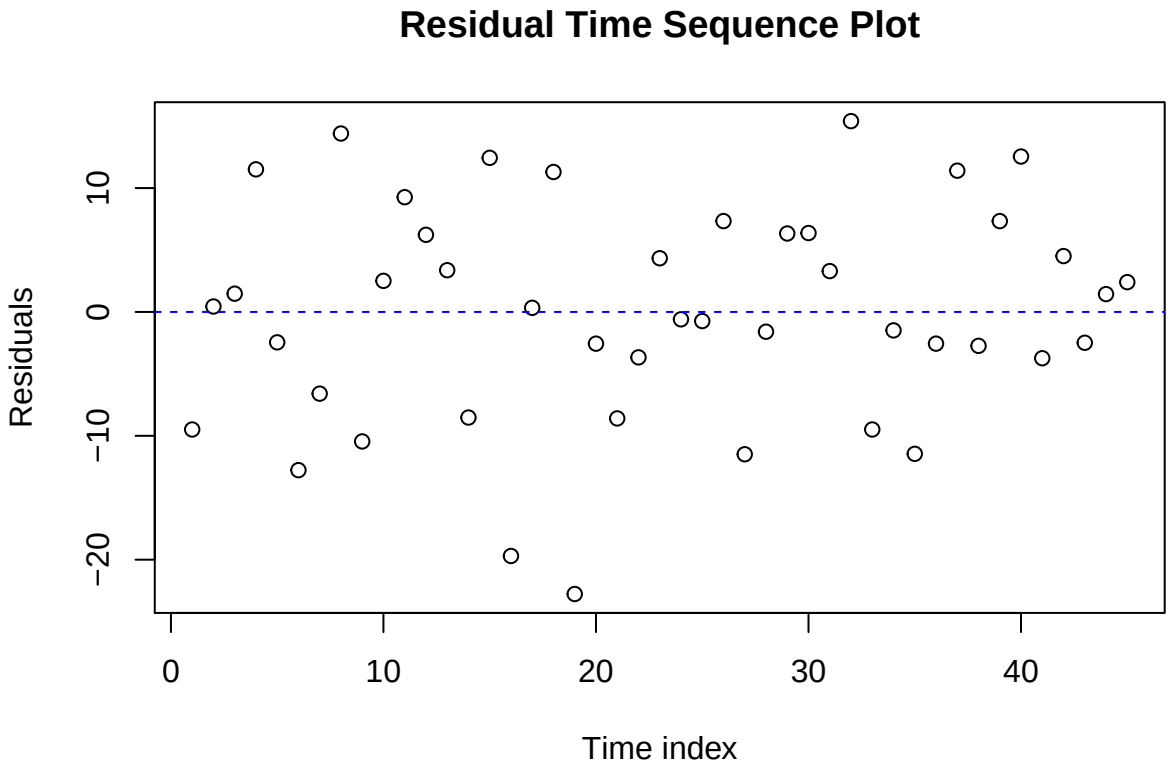
```r
# Use linear interpolation to obtain critical value for correlation coefficient
# Two points at alpha = 0.10: (n=40, cor=.977), (n=50, cor=.981)
(crit_cor_coef <- 0.977 + (nrow(copier)-40) * (0.981-0.977)/(50-40))
```

```
## [1] 0.979
```

Controlling the $\alpha$ risk at 0.10, we found from **Table B6**, by linear interpolation, that the critical value for $n = 45$ is `0.979`. Since the observed correlation coefficient of `0.9891` exceeds this level, we have support for our earlier conclusion from the normal probability plot that the distribution of the error terms does not depart substantially from a normal distribution.

## Part (f)

```r
# a time plot of the residuals
plot(resid(mdl_copier), main = "Residual Time Sequence Plot", xlab = "Time index", ylab = "Residuals")
abline(h=0, lty=2, col="blue")
```




The residuals fluctuate in a random pattern around the 0 base line, indicating that the error terms are not correlated over time. This means that there is no effect connected with time when studying the relation between number of copiers serviced ($X$) and the service time ($Y$), which leads us to conclude that the error terms are independent.

## Part (g): Breusch-Pagan Test

```r
# conducting a Bruesch-Pagan test
library(lmtest)
# Get the BP test results
bptest(mdl_copier, studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  mdl_copier
```

```
## BP = 1.3, df = 1, p-value = 0.3
```

```
# Compute the chi-square acceptance region
qchisq((1-0.05), df=1)
```

```
## [1] 3.841
```

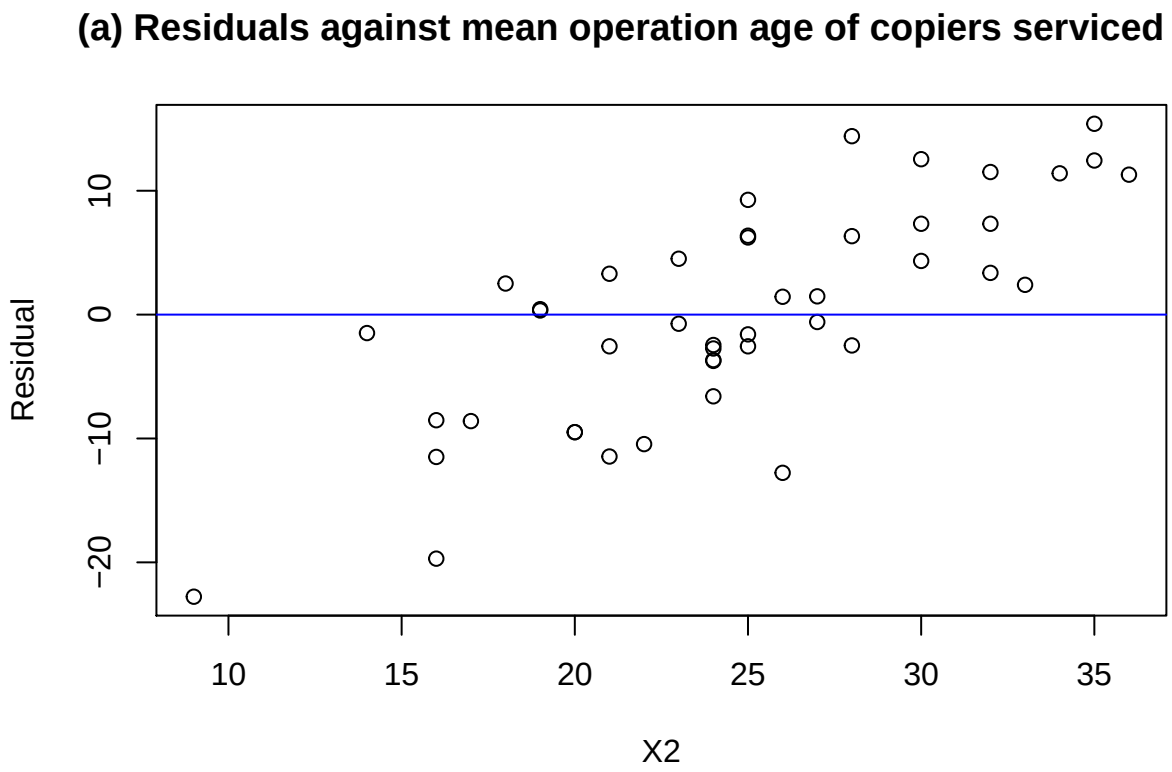
- Significance level, $\alpha = 0.05$.
- The two alternative hypotheses are as follows:

$$H_0 : \text{The error variance is constant. Versus } H_a : \text{The error variance is not constant.}$$

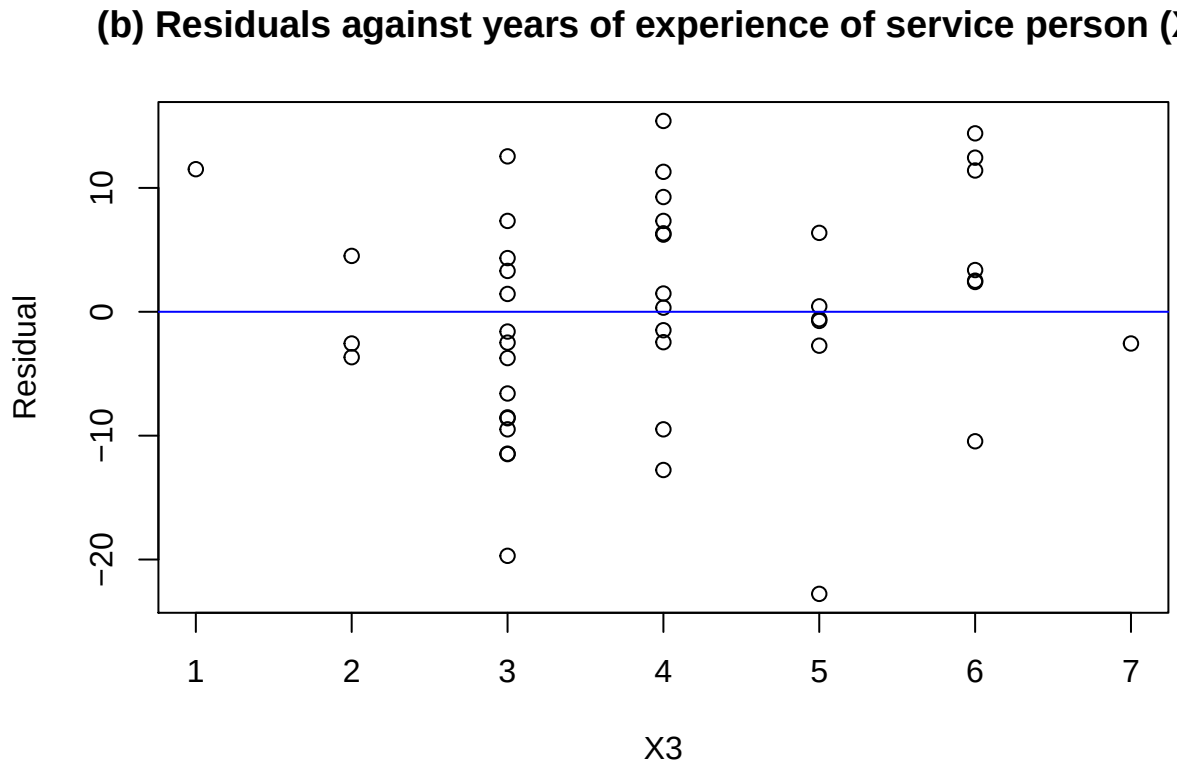
- If $\chi^2_{BP} \leq \chi^2_{(0.95;1)} = 3.841$, conclude $H_0$, otherwise, conclude $H_a$.
- Since $\chi^2_{BP} = 1.3 < \chi^2_{(0.95;1)} = 3.841$, we conclude $H_0$, that the error variance is constant at 5% level of significance.

## Part (h)

```
plot(copier$X2, resid(mdl_copier), main = "(a) Residuals against mean operation age of copiers serviced (X
     xlab = "X2", ylab = "Residual")
abline(0,0, col="blue")
```




```
plot(copier$X3, resid(mdl_copier), main = "(b) Residuals against years of experience of service person (X3
     xlab = "X3", ylab = "Residual")
abline(0,0, col="blue")
```




From plot (a), the residuals are seen to increase with increasing levels of the second variable (operational age of copiers serviced), showing that this variable appears to have an explicit effect on the service time, while in plot (b) there is no clear effect of the third variable (years of experience of the service person) on the residuals. Hence, I will conclude that the mean operational age of copiers serviced on the call ($X_2$) can lead to improved performance when included in the model.

## Problem 3.5: Airfreight breakage (Problem 1.21)

We begin this part of the assignment by importing the `Airfreight breakage` dataset and fitting a simple regression model between ampules broken and the number of transfers made. Summary information about the model's estimated coefficients is presented below.

```
# Importing the dataset
airfreight <- read.table("../Data Sets/Chapter  1 Data Sets/CH01PR21.txt",
                         header = F,
                         col.names = c("broken_ampules", "transfer_made")
                       )

# Fit a linear regression model and display estimates
mdl_airfreight <- lm(broken_ampules~transfer_made, data =  airfreight)

mdl_airfreight %>%
  tidy(conf.int = TRUE, conf.level = 0.95) %>%
  mutate(term = c("$\\beta_0$", "$\\beta_1$")) %>%
  kable(caption = "Parameter Estimates with 95% Confidence Interval")
```

Table 3: Parameter Estimates with 95% Confidence Interval

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| $\beta_0$ | 10.2 | 0.6633 | 15.377 | 0 | 8.670 | 11.730 |
| $\beta_1$ | 4.0 | 0.4690 | 8.528 | 0 | 2.918 | 5.082 |

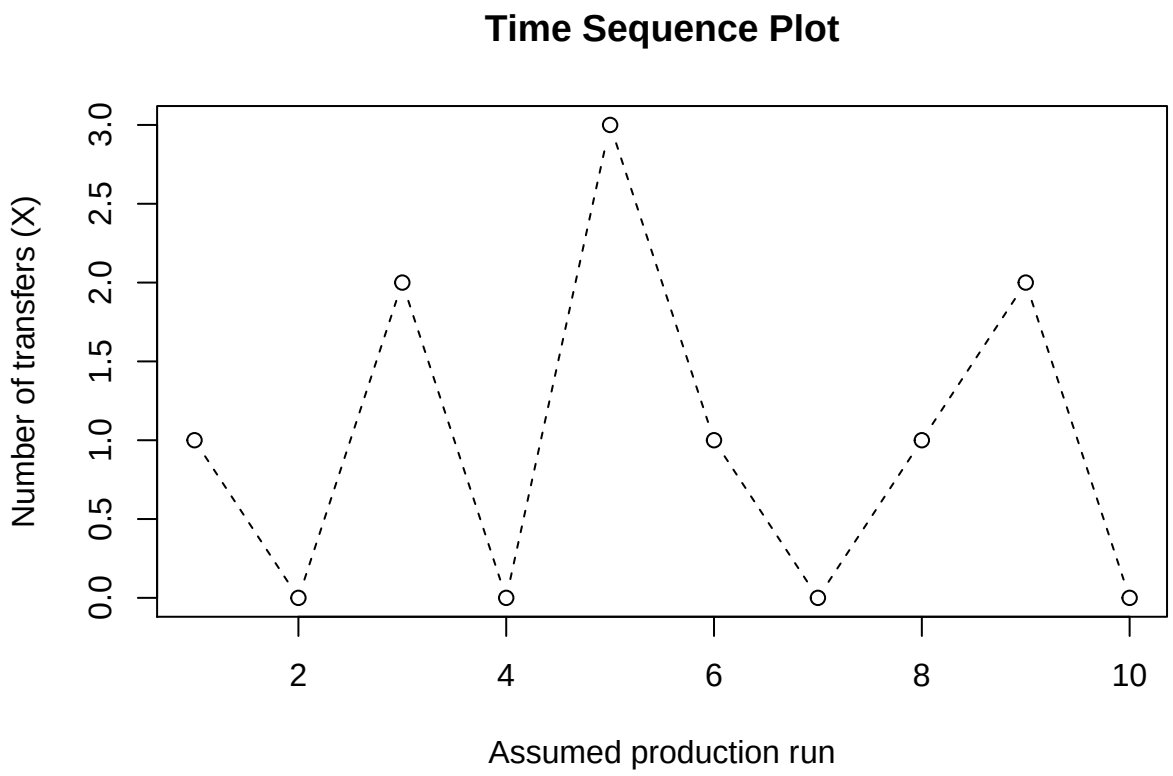
## Part (a)

```
# create a dot plot of X, number of copiers serviced
dotchart(airfreight$transfer_made, xlab="Number of transfers (X)",
         ylab = "Frequency", main = "Dot plot of number of transfers made")
```


The dot plot shows that the number of transfers made ranges between 0 and 3, with most observations occurring at 0. Here, the distribution of the number of transfers made does not appear to be symmetrical (that is, it is asymmetrical). The point at 3 appears to be an outlier.

## Part (b)

```
# Prepare a time plot for number of transfers
plot(airfreight$transfer_made, type = "b", lty = 2, xlab = "Assumed production run", ylab = "Number of tra
title("Time Sequence Plot")
```




Here, the sequence plot shows no evidence of a systematic pattern, suggesting that the number of transfers made are, perhaps, not correlated with time.

### Part (c): Stem-and-leaf plot of the residuals

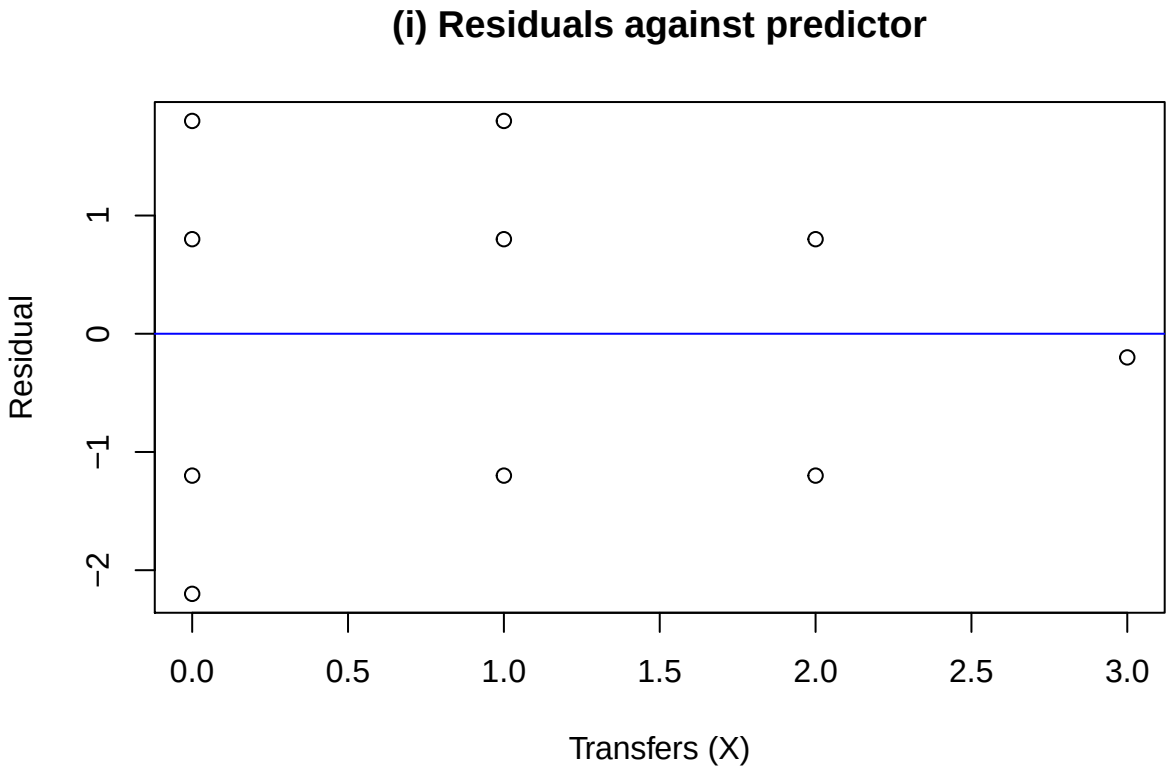
```
# stem-and-leaf plot
stem(resid(mdl_airfreight), scale=3)
```

```
##
##   The decimal point is at the |
##
##   -2 | 2
##   -1 |
##   -1 | 222
##   -0 |
##   -0 | 2
##    0 |
##    0 | 888
##    1 |
##    1 | 88
```

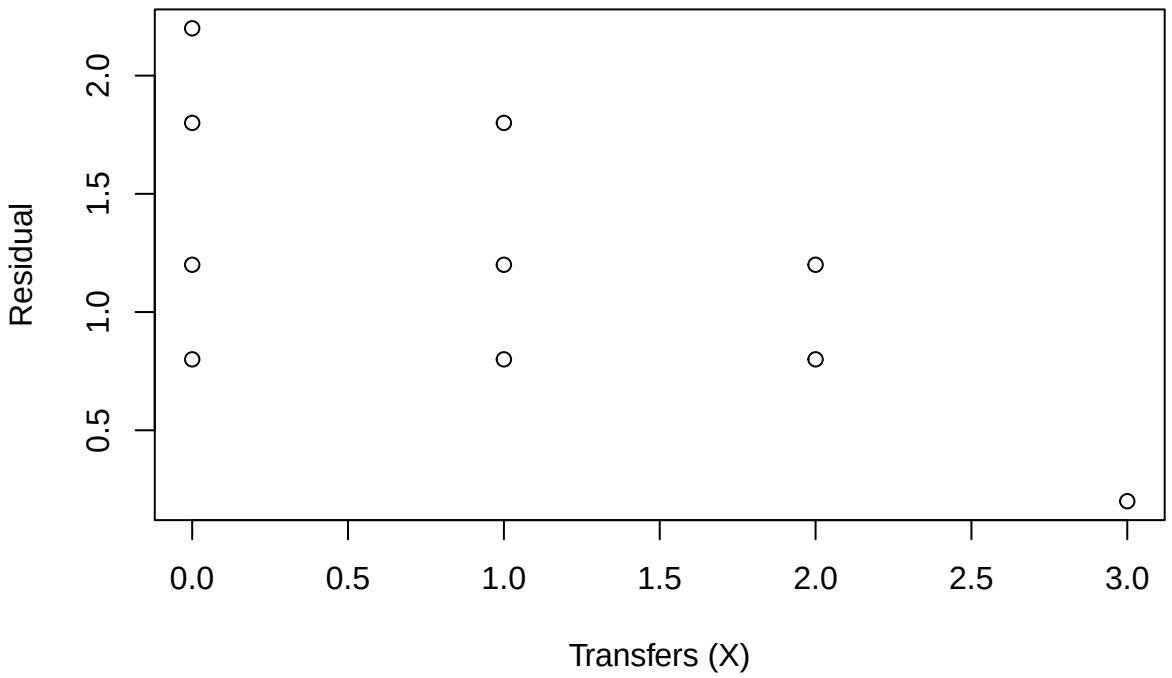
The plot provides information about the shape, modality and spread of the residuals. This plot shows that the distribution of the residuals is bimodal, with one peak occurring at -1.2 and the other at 0.8. The residuals ranges between -2.2 and 1.8 which are quite close. We can also see that close to half of the observations have residuals are around zero (0).

### Part (d)

```
# residual vs. predictor plot
plot(airfreight$transfer_made, resid(mdl_airfreight), main = "(i) Residuals against predictor",
     xlab = "Transfers (X)", ylab = "Residual")
abline(0,0, col="blue")
```




```
plot(airfreight$transfer_made, abs(resid(mdl_airfreight)), main = "(ii) Absolute residuals against predict
     xlab = "Transfers (X)", ylab = "Residual")
abline(0,0, col="blue")
```




- Plot (i) provides no evidence of departure of the regression function from linearity. Also, there is no clear evidence of outlying cases from looking at this plot.
- However, there appears to be an issue of the error variance not being constant as reflected by the decreasing

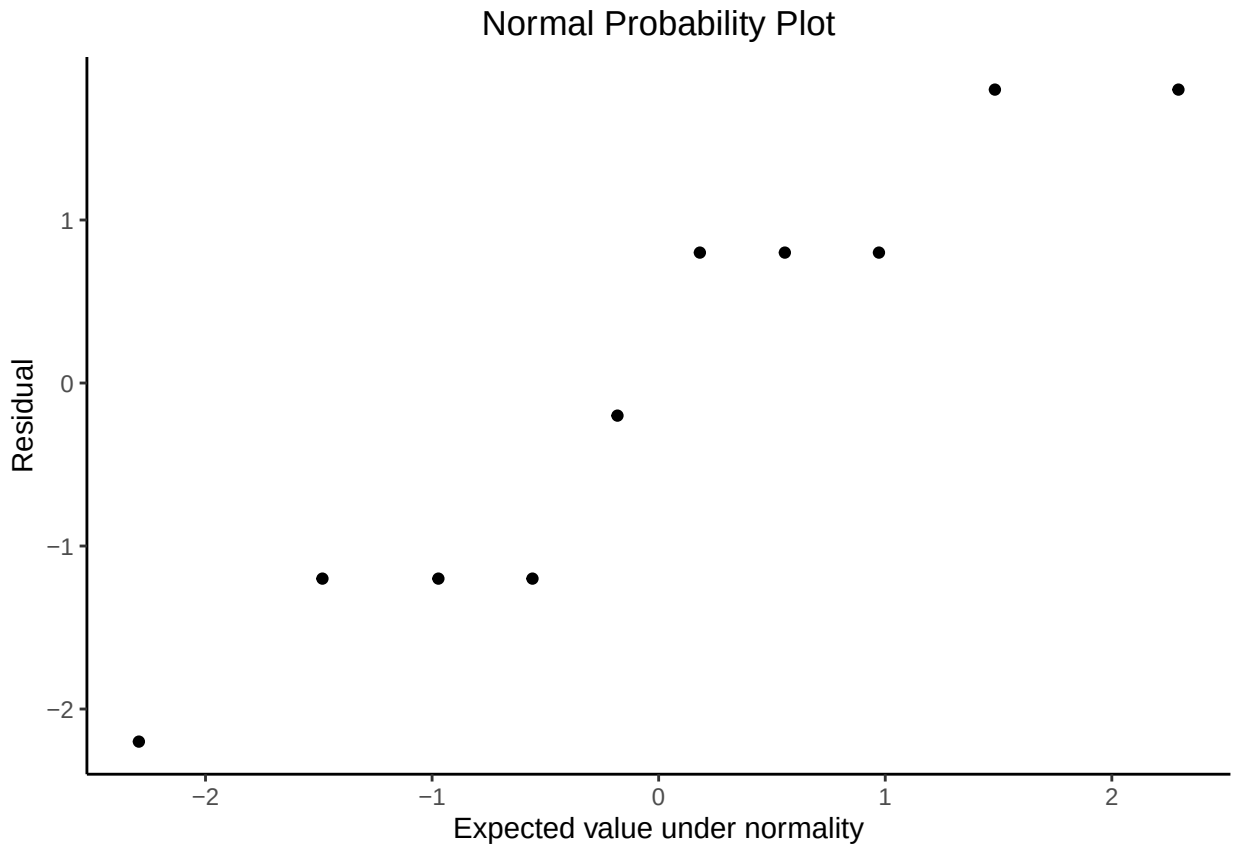
nature of the magnitude of the residuals as the number of transfers increases. The decreasing behavior of the residuals is somehow concealed probably due to the small number of cases in the dataset, so I included a plot of the absolute values of the residuals against the predictor variable $X$, plot (ii), to help me see clearly the true nature of the error variance. As a confirmation, plot (ii) shows more clearly that the residuals tend to be smaller in absolute magnitude for higher $X$ values. Thus, by using these plots I can say that the error variance is not constant.

## Part (e)

```
# Obtain the ordered residuals in ascending order
resid_df <- data.frame(resid_ord = sort(resid(mdl_airfreight)),
                       k = 1:nrow(airfreight))
# computing the expected values
MSE_sqrt <- mdl_airfreight %>% glance() %>% pull(sigma)
resid_df <- resid_df %>%
  mutate(expected_value = MSE_sqrt *(qnorm((k-0.375)/(nrow(airfreight)+0.25))))

# view few rows
kable(head(resid_df), col.names = c("Ordered residual", "Rank(k)", "Expected value"))
```

| | Ordered residual | Rank(k) | Expected value |
|---|---|---|---|
| 7 | -2.2 | 1 | -2.2940 |
| 6 | -1.2 | 2 | -1.4840 |
| 3 | -1.2 | 3 | -0.9722 |
| 2 | -1.2 | 4 | -0.5569 |
| 5 | -0.2 | 5 | -0.1818 |
| 10 | 0.8 | 6 | 0.1818 |

```
# Obtain the expected values under normality using suggested procedure from the textbook
ggplot(resid_df, aes(expected_value, resid_ord)) +
  geom_point() +
  labs(title = "Normal Probability Plot",
       x = "Expected value under normality",
       y = "Residual") +
  theme(plot.title = element_text(hjust = 0.5))
```




The plot appears to depart somehow from linearity. However the departure of the plot from linearity is not that substantial. Therefore, I will conclude that the error distribution does not depart substantially from a normal distribution.

```
# compute the correlation between ordered residuals and the expected values
(obs_cor_coef <- cor(resid_df$resid_ord, resid_df$expected_value))
```

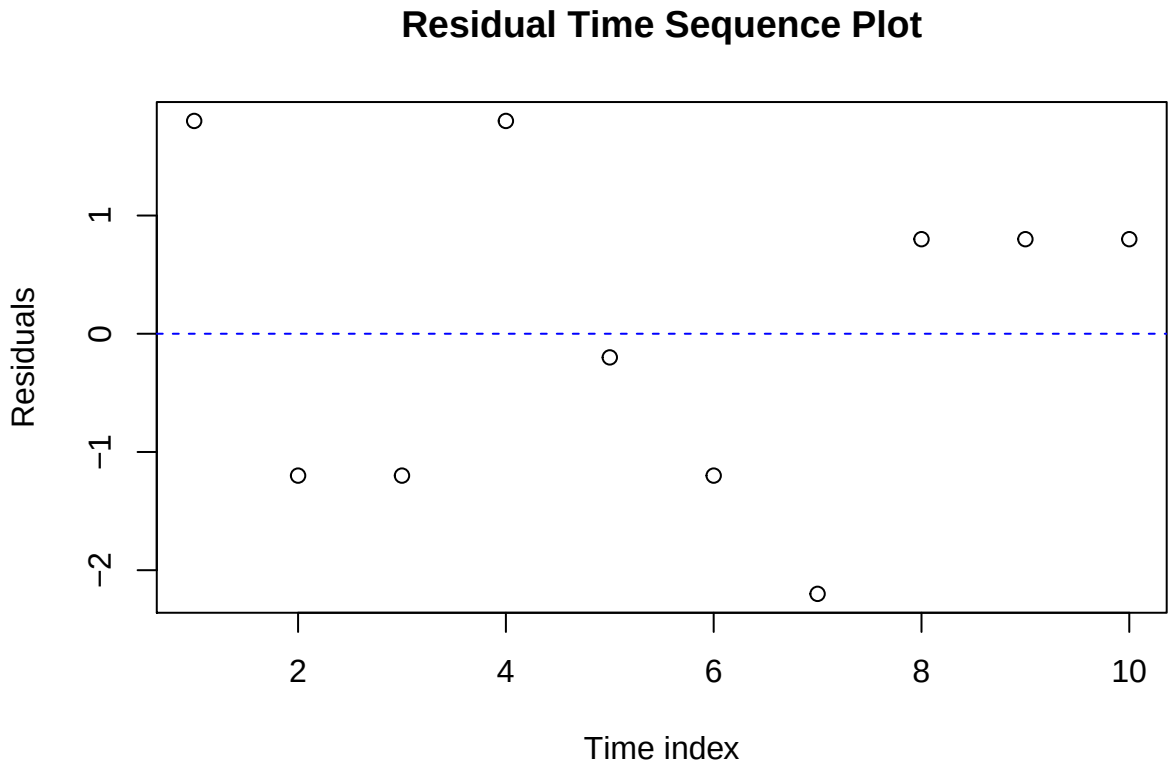
```
## [1] 0.961
```

```
crit_cor_coef <- 0.879 # read directly from Table B6 at alpha = 0.01 and n =10.
```

Controlling the $\alpha$ risk at 0.01, we found from **Table B6** that the critical value for $n = 10$ is `0.879`. Since the observed correlation coefficient of `0.961` exceeds this level, we conclude that the distribution of the error terms does not depart substantially from a normal distribution. This result happens to be consistent with my initial assessment using the normal probability plot above, though I would rather proceed with caution when basing my assessment solely on the normal probability plot.

## Part (f)

```
# a time plot of the residuals
plot(resid(mdl_airfreight), main = "Residual Time Sequence Plot", xlab = "Time index", ylab = "Residuals")
abline(h=0, lty=2, col="blue")
```




The residuals fluctuate in a random manner around the 0 base line, indicating that the error terms are not correlated over time. This means that there is no effect associated with time when studying the relation between number of transfers ($X$) and the number of broken ampules ($Y$), which leads us to conclude that the error terms are independent.

### Part (g): Breusch-Pagan Test

```
#----- conducting a Bruesch-Pagan test -----
# Get the BP test results
bptest(mdl_airfreight, studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  mdl_airfreight
## BP = 1, df = 1, p-value = 0.3
```

```
# Compute the chi-square acceptance region
qchisq((1-0.10), df=1)
```

```
## [1] 2.706
```

- Significance level, $\alpha = 0.10$.
- The two alternative hypotheses are as follows:

$$H_0 : \text{The error variance does not varry with the level of X (prdictor).}$$

$$H_a : \text{The error variance varies with the level of X (predictor).}$$

- If $\chi^2_{BP} \leq \chi^2_{(0.90;1)} = 2.706$, conclude $H_0$, otherwise, conclude $H_a$.
- Therefore, since $\chi^2_{BP} = 1 < \chi^2_{(0.90;1)} = 2.706$, we conclude $H_0$, at 10% level of significance, that the error variance is constant at all levels of $X$.

My conclusion here does not support my preliminary findings in part (d), since in part (d) I concluded that the error variance was not constant.

## Remarks

I found some contrasting results between some of the diagnostic plots for the `airfreight maintenance` data and their corresponding numerical tests. For instance, the residual versus predictor plot appears to suggest that the error variance is not constant, while the **Bruesch-Pagan** test provided sufficient evidence of the error variance being constant with the level of $X$. I therefore suspects that the relatively small sample size of 10 could be a contributory factor for such differences.

## Reference


- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Wasserman, W. (2004). Applied linear regression models (Vol. 4). New York: McGraw-Hill/Irwin..
- RPubs site for ALSR (Chapter 3 – Diagnostics and Remedial Measures): https://rpubs.com/bryangoodrich/5217.