Chapter 3 - Diagnostics & Remedial Measures: 3.4, 3.12, 3.13, 3.17, 3.18, 3.21, 3.23.

Problem 3.4

Refer to **Copier maintenance** Problem 1.20.

a. Prepare a dot plot for the number of copiers serviced X_i. What information is provided by this plot? Are there any outlying cases with respect to this variable?

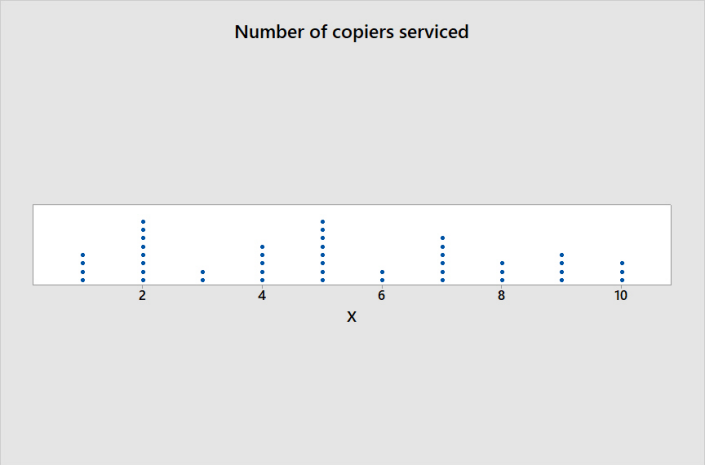


Figure 1: Dot plot for the number of copiers serviced X_i.

The dot plot show that the minimum of 1 and maximum of 10 copiers are serviced. The number of copiers serviced levels are spread throughout this range and there are no outlining cases with respect to this variable. Since there are replications at each X-level, we can estimate Pure Error.

b. The cases are given in time order. Prepare a time plot for the number of copiers serviced. What does your plot show?

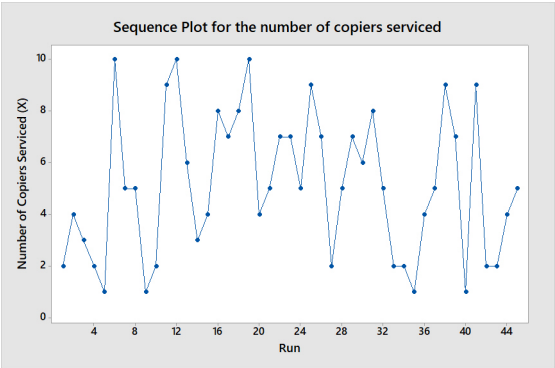


Figure 2: Time plot for the number of copiers serviced.

Based on the plot, there is no pattern with regards to this variable and time series is random.

c. Prepare a stem-and-leaf plot of the residuals. Are there any noteworthy features in this plot? *Based on the plot,*

Stem-and-leaf of Residuals N = 45

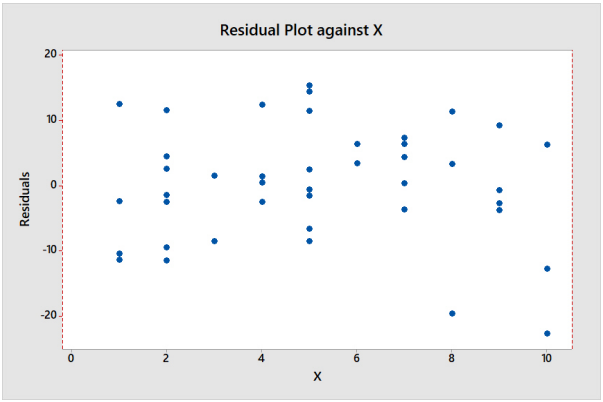
1	-2	2
2	-1	9
6	-1	2110
11	-0	99886
22	-0	33222221100
(10)	0	0011223344
13	0	666779
7	1	111224
1	1	5

Leaf Unit = 1

Figure 3: Stem-and-leaf plot of the residuals.

Based on the stem-and-leaf plot our residuals' distribution close to normal.

d. Prepare residual plots of e_i versus $\hat{Y}_i$ and e_i versus X_i on separate graphs. Do these plots provide the same information? What departures from regression model (2.1) can be studied from these plots? State your findings.



(a) Residual plots of e_i versus X_i

(b) Residual plots of e_i versus $\hat{Y}_i$

Figure 4: Residual plots of e_i versus X_i and versus $\hat{Y}_i$

Both of the graphs above provide the same information about the departures from the regression model (they are identical). The residual plots confirm the constant variance assumption of the regression model and absence of patterns. While most of the residuals are symmetric to $e_i = 0$ there are some interesting departures worth noting. The model tends to "underestimate" when $x = 1$ and $x = 10$ which are the min and max number of copiers serviced. The regression model seems to "overestimate" when $x = 7$ but this departures are not too dramatic. The last interesting point is when $x = 8$ as the model seems to do poorly at this point. There are only two values where the residuals fall below $e_i = -15$ which occur at $x = 8$ and $x = 10$ pointing that we might have outliers.

e. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be tenable here? Use Table B.6 and $\alpha = 0.10$.

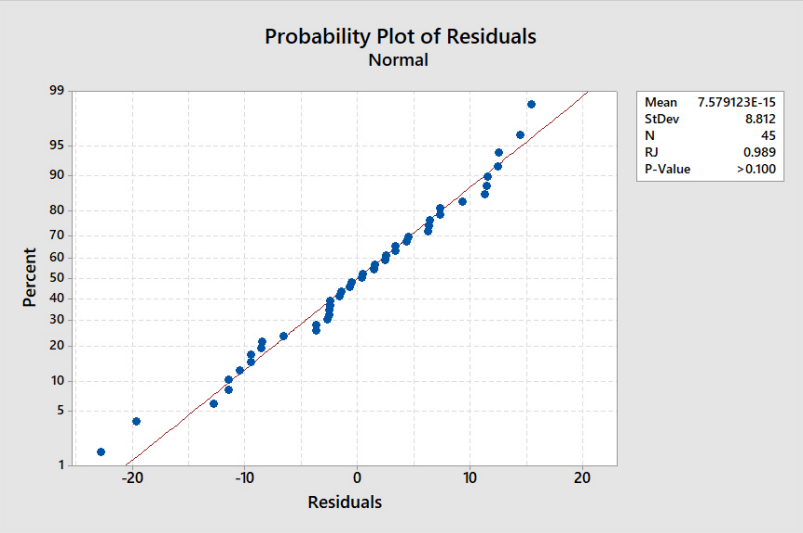


Figure 5: Normal probability plot of the residuals.

The normal probability plot of the residuals shows that the residuals are approximately normally distributed with a few outliers. From the Table B.6 we find that the critical value (percentile) is $r \in (0.977, 0.981)$ *interval for* $\alpha = 0.10$. *This value is less than the Ryan-Joiner coefficient of correlation of 0.989, which indicates that the normality assumption holds.*

f. Prepare a time plot of the residuals to ascertain whether the error terms are correlated over time. What is your conclusion?

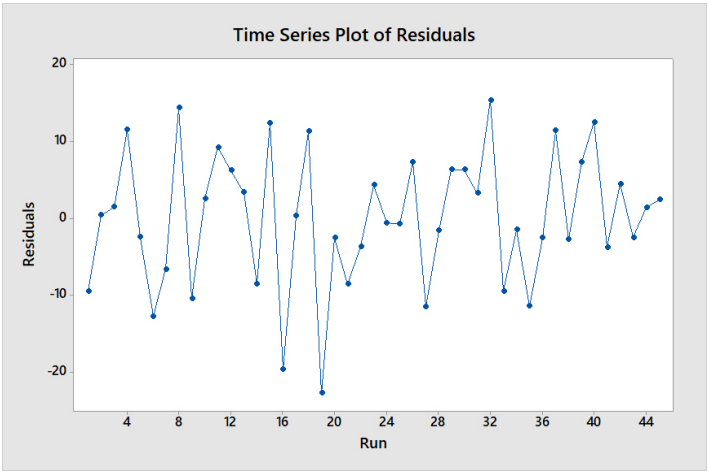


Figure 6: Time plot of the residuals.

Based on the plot, there is no pattern with regards to the residuals over time implying they're independent.

g. Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion.

$H_0 : \gamma_1 = 0$ (error variance is constant)
$H_a : \gamma_1 \neq 0$

Decision rule:

If p-value ≤ 0.05, reject H_0
If p-value > 0.05 fail to reject H_0.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	41.8	32.7	1.28	0.208	
X	6.67	5.64	1.18	0.243	1.00

Regression Equation

(e_i)^2 = 41.8 + 6.67 X

Figure 7: Minitab output for the Breusch-Pagan Test.

Based on the Minitab output for the Breusch-Pagan test, we get the $p-value > 0.05$ *and we fail to reject* H_0*. Hence, error variance appears to be constant.*

Alternatively:
$X^2_{bp} = \frac{15.155/2}{(3416.4/45)^2} = 1.3147$

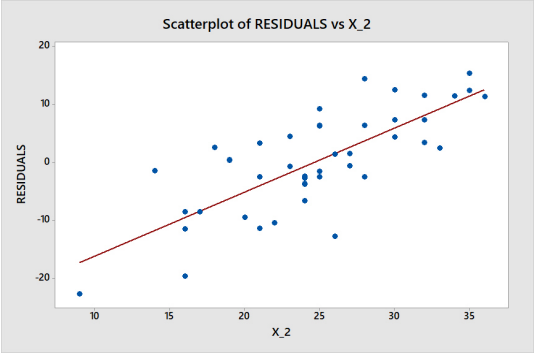
p-value= 0.7485, *we fail to reject* $H_0 : \sigma_i^2 = \sigma^2$*. Hence, error variance is constant.*

h. Information is given below on two variables not included in the regression model, namely, mean operational age of copiers serviced on the call (X_2, in months) and years of experience of the service person making the call (X_3). Plot the residuals against X_2 and X_3 on separate - graphs to ascertain whether the model can be improved by including either or both of these variables. What do you conclude?

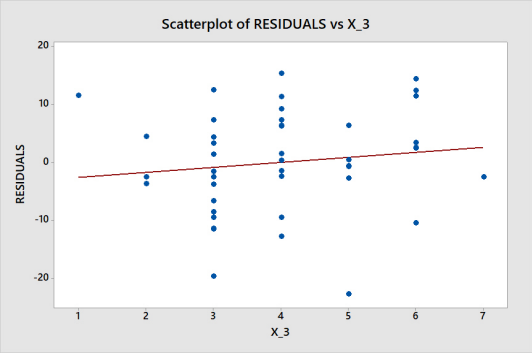
i :	1	2	3	...	43	44	45
X_2 :	20	19	27	...	28	26	33
X_3 :	4	5	4	...	3	3	6

Based on the plots, X_2- *mean operational age of copiers serviced on the call- has strong positive correlation with the residuals while* X_3- *years of experience of the service person making the call-shows no pattern. Hence, including* X_3 *will not add any value towards improving the model. Adding* X_2*, however, would help improve the regression model.*

See graph below:



(a) Plot of the residuals against X_2



(b) Plot of the residuals against X_3

Figure 8: Plot of the residuals against X_2 and X_3

Problem 3.12

A student does not understand why the sum of squares defined in (3.16) is called a pure error sum of squares "since the formula looks like one for an ordinary sum of squares" Explain.

$SSE = \sum(Y_i - \hat{Y})^2$
$SSPE = \sum\sum(Y_{ij} - \bar{Y}_j)^2$

SSE is model-dependent since $\hat{Y}$ is included in its calculation. SSPE on the other hand measures the deviations from "local" means or deviations at each X level (with replicates), making Pure Error model-independent.

Problem 3.13

Refer to Copier maintenance Problem 1.20.

a. What are the alternative conclusions when testing for lack of fit of a linear regression function?

$H_0 : E\{Y_i\} = \beta_0 + \beta_1 X_i$ (μ_j in the full model $Y_{ij} = \mu_j + \epsilon_{ij}$is linearly related to$X_j$)
$H_a : E\{Y_i\} \neq \beta_0 + \beta_1 X_i$

b. Perform the test indicated in part (a). Control the risk of Type I error at 0.05. State the decision rule and conclusion.

Decision rule:

If $F^* \leq F(1-\alpha; c-2, n-c)$, conclude H_0
If $F^* > F(1-\alpha; c-2, n-c)$, conclude H_a.

From Figure 9:
$SSLF = 618.7$
$SSPE = 2797.7$

$n = 45$; $c = 10$ (see Figure 1.)

$$\begin{aligned} SSE &= SSPE + SSLF \\ F^* &= \frac{SSE - SSPE}{(n-2)-(n-c)} \div \frac{SSPE}{n-c} = \frac{SSLF}{c-2} \cdot \frac{n-c}{SSPE} = \frac{618.7}{10-2} \cdot \frac{45-10}{2797.7} = \\ &= 77.3375 \cdot 0.01251 = 0.9675 \\ F(0.95; 8, 35) &= 2.216675033 \simeq 2.217 \end{aligned} \tag{1}$$

Regression Analysis: Y versus X_1

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	76960.4	95.75%	76960.4	76960.4	968.66	0.000
X_1	1	76960.4	95.75%	76960.4	76960.4	968.66	0.000
Error	43	3416.4	4.25%	3416.4	79.5		
Lack-of-Fit	8	618.7	0.77%	618.7	77.3	0.97	0.477
Pure Error	35	2797.7	3.48%	2797.7	79.9		
Total	44	80376.8	100.00%				

Figure 9: Minitab output for the Breusch-Pagan Test.

$F^* = .968 < 2.217$. *Conclude H_0 that the regression function is linear ($p-value = 0.477$).*

c. Does the test in part (b) detect other departures from regression model (2.1), such as lack of constant variance or lack of normality in the error terms? Could the results of the test of lack of fit be affected by such departures? Discuss.

No, the test doesn't detect other departures. We would need to do normality test and regress the residuals over time to check for variance constancy.

Problem 3.17

Sales growth. A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data are as follows, where X is the year (coded) and Y is sales in thousands of units:

i :	1	2	3	4	5	6	7	8	9	10
X_i :	0	1	2	3	4	5	6	7	8	9
Y_i :	98	135	162	178	221	232	283	300	374	395

a. Prepare a scatter plot of the data. Does a linear relation appear adequate here?

A linear relation is adequate for this relationship, with no potential outliers in the data.

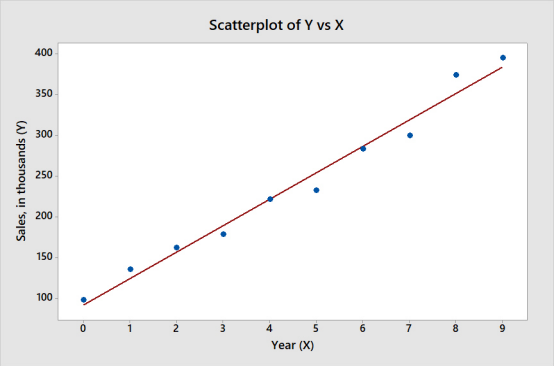


Figure 10: Scatter plot of the sales data against the years coded.

b. Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation of Y. Evaluate SSE for $\lambda = 0.3, 0.4, 0.5, 0.6, 0.7$. What transformation of Y is suggested?

Step 1. *We need to test for normality. From the plot below, Ryan-Joiner coefficient is 0.986 and* $p-value = 0.100 > 0.05$ *indicating that the data is normally distributed:*

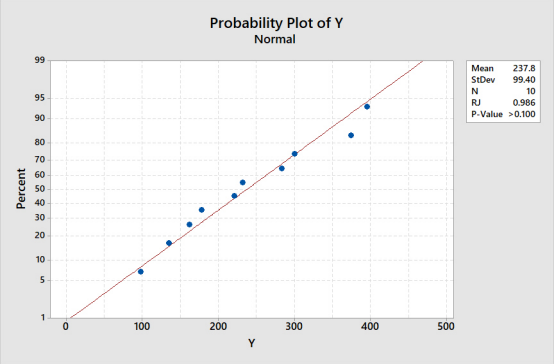


Figure 11: Normal Probability Plot of Y.

Step 2. *We perform Box-Cox Transformation*

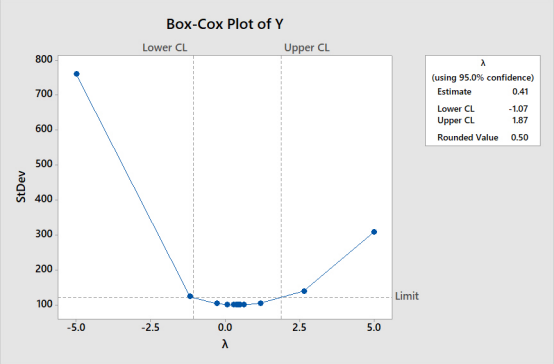


Figure 12: Box-Cox Plot of Y.

Below are the SSE's for various λ values:

λ :	0.3	0.4	0.5	0.6	0.7
SSE :	1099.7093	967.9088	916.4048	942.4498	1044.2384

Based on the Minitab output appropriate $\lambda = 0.41$ which can be rounded to 0.5. From the calculations of the SSE we can also see that the lowest SSE is at $\lambda = 0.5$. Then, the power transformation of Y would be $Y' = \sqrt{Y}$.

c. Use the transformation $Y' = \sqrt{Y}$ and obtain the estimated linear regression function for the transformed data.

Regression Analysis: Y versus X

Method

Box-Cox transformation λ = 0.5

Analysis of Variance for Transformed Response

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	95.568	95.5683	728.37	0.000
X	1	95.568	95.5683	728.37	0.000
Error	8	1.050	0.1312		
Total	9	96.618			

Model Summary for Transformed Response

S	R-sq	R-sq(adj)	R-sq(pred)
0.362227	98.91%	98.78%	98.25%

Coefficients for Transformed Response

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	10.261	0.213	48.20	0.000	
X	1.0763	0.0399	26.99	0.000	1.00

Regression Equation

Y^0.5 = 10.261 + 1.0763 X

Figure 13: The estimated linear regression function output for the transformed data.

$$\boxed{\sqrt{Y} = 10.261 + 1.0763X}$$

d. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

Based on the plot, the regression line does not appear to be a good fit to the transformed data (unlike the original data). There are outliers on the ends that affaect the accuracy of the prediction.

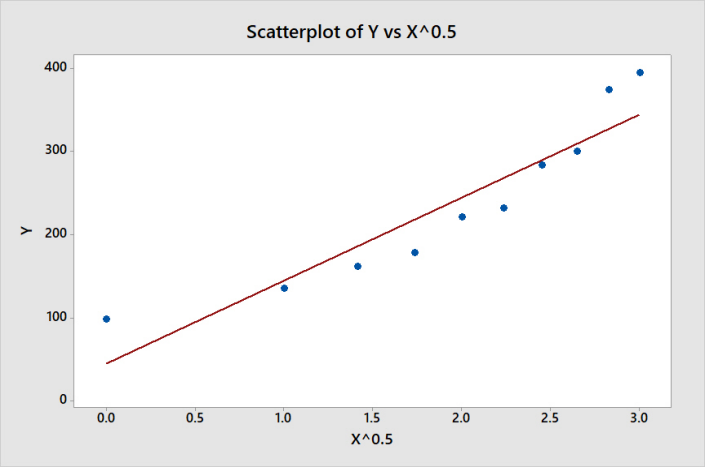
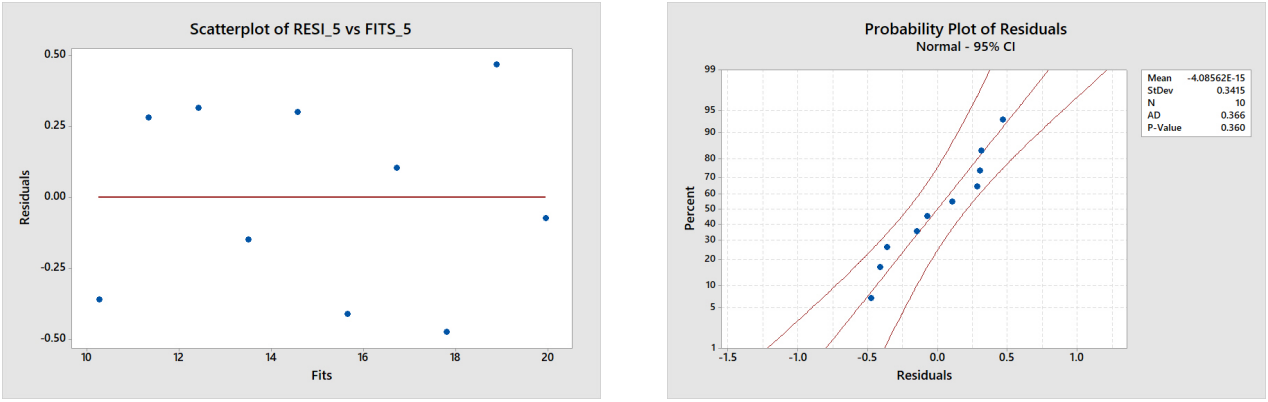


Figure 14: The estimated linear regression function output for the transformed data ($\sqrt{X}$.

e. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?



(a) Plot of the residuals against the fitted values.

(b) Normal probability plot of the residuals.

Figure 15: Residual associated plots.

Plots show high variance associated with the distribution of the residuals. The residuals don't seem to be normally distributed.

f. Express the estimated regression function in the original units.

$$\begin{aligned} Y' = \sqrt{Y} &= 10.261 + 1.0763X \\ \hat{Y} &= (10.261 + 1.0763X)^2 \\ \hat{Y} &= 105.29 + 22.09X + 1.158X^2 \end{aligned} \tag{2}$$

Problem 3.18

Production time. In a manufacturing study, the production times for 111 recent production runs were obtained. The table below lists for each run the production time in hours (Y) and the production lot size (X).

i :	1	2	3	...	109	110	111
X_i :	15	9	7	...	12	9	15
Y_i :	14.28	8.80	12.49	...	16.37	11.45	15.78

a. Prepare a scatter plot of the data Does a linear relation appear adequate here? Would a transformation on X or Y be more appropriate here? Why?

Based on the scatter plot there is a curvilinear association with Lack-of-Fit p-value = 0.055 and a transformation on X seem to be appropriate.

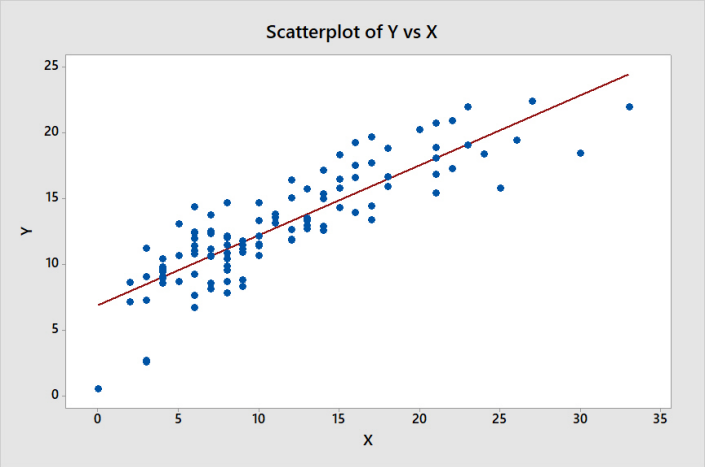


Figure 16: Scatter plot of the data.

b. Use the transformation $X' = \sqrt{X}$ and obtain the estimated linear regression function for the transformed data.

$$\boxed{Y = 1.255 + 3.624\sqrt{X}}$$

c. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

Based on the plot (Figure 17), the regression line appear to be a good fit compared to the previous plot with the Lack-of-Fit p-value = 0.424. We seem to have one potential outlier.

d. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

The plots show that there is somewhat constant variance of the residuals and they seem to be approximately normally distributed meeting the assumptions of a linear model.

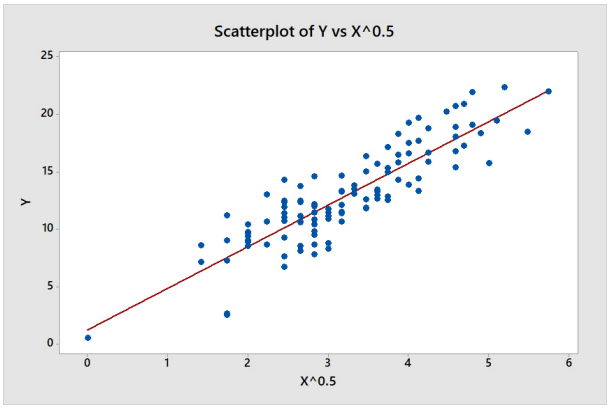
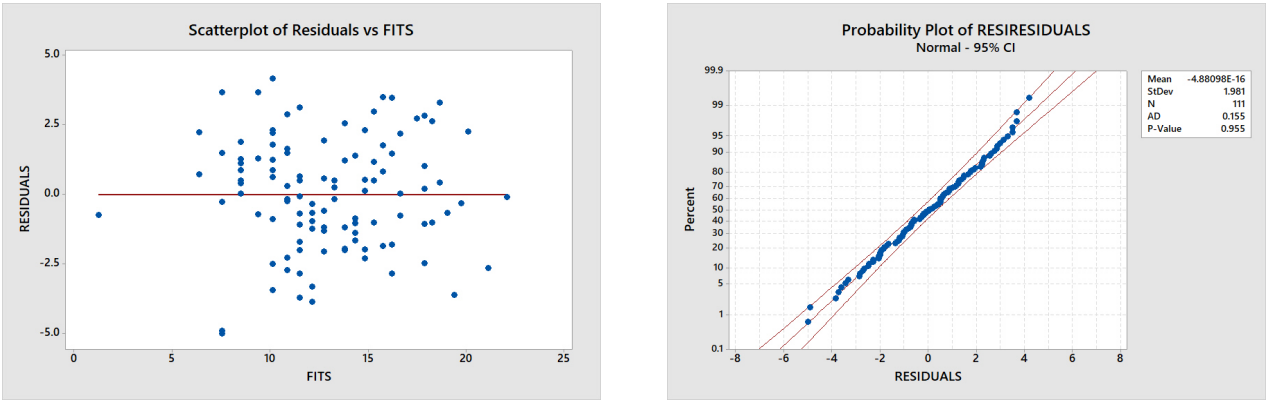


Figure 17: The estimated linear regression function output for the transformed data.



(a) Plot of the residuals against the fitted values. (b) Normal probability plot of the residuals.

Figure 18: Residuals associated plots.

e. Express the estimated regression function in the original units.

$$
\begin{aligned}
Y &= 1.255 + 3.624\sqrt{X} \\
\sqrt{X} &= \frac{Y - 1.255}{3.624} \\
X &= \left(\frac{Y - 1.255}{3.624}\right)^2 = \frac{Y^2 - (2 * 1.255Y) + (1.255)^2}{13.133} \qquad (3) \\
13.133X - 1.575 &= Y^2 - 2.51Y \\
Y(Y - 2.51) &= 13.133X - 1.575
\end{aligned}
$$

$$\boxed{Y(Y - 2.51) = 13.133X - 1.575}$$

Problem 3.21

Derive the result in (3.29).
(3.29) $\sum\sum(Y_{ij} - \hat{Y}_{ij})^2 = \sum\sum(Y_{ij} - \bar{Y}_j)^2 + \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$

$\sum_i \hat{Y}_{ij} = n_j \hat{Y}_{ij}, for \hat{Y}_{ij} = b_0 + b_1 X_j$
$\bar{Y}_{ij} = \frac{\sum_i Y_{ij}}{n_j}$

Prep Work:

$$\begin{aligned}
\sum_j^c \sum_i^{n_j} (Y_{ij} - \bar{Y}_j)(\bar{Y}_j - \hat{Y}_{ij}) &= \sum\sum (Y_{ij}\bar{Y}_j - \bar{Y}_j\bar{Y}_j - Y_{ij}\hat{Y}_{ij} + \bar{Y}_j\hat{Y}_{ij}) = \\
&= \sum\sum Y_{ij}\bar{Y}_j - \sum\sum \bar{Y}_j\bar{Y}_j - \sum\sum Y_{ij}\hat{Y}_{ij} + \sum\sum \bar{Y}_j\hat{Y}_{ij} = \\
&= \sum_j \bar{Y}_j \sum_i Y_{ij} - n_j^2\bar{Y}_j^2 - \sum\sum Y_{ij}\hat{Y}_{ij} + \sum_j \bar{Y}_j \sum_i \hat{Y}_{ij} = \\
&= \bar{Y}_j \sum_i n_j\bar{Y}_j - n_j^2\bar{Y}_j^2 - \sum\sum Y_{ij}\hat{Y}_{ij} + \frac{\sum_i Y_{ij}}{n_j} \sum_i n_j\hat{Y}_{ij} = \\
&= n_j^2\bar{Y}_j^2 - n_j^2\bar{Y}_j^2 - \sum\sum Y_{ij}\hat{Y}_{ij} + \sum\sum Y_{ij}\hat{Y}_{ij} = 0
\end{aligned} \tag{4}$$

Derivation:

$$\begin{aligned}
Y_{ij} - \hat{Y}_{ij} &= Y_{ij} - \bar{Y}_j + \bar{Y}_j - \hat{Y}_{ij} \\
\sum\sum (Y_{ij} - \hat{Y}_{ij})^2 &= \sum\sum \Big[(Y_{ij} - \bar{Y}_{ij}) + (\bar{Y}_j - \hat{Y}_{ij})\Big]^2 = \\
&= \sum\sum (Y_{ij} - \bar{Y}_j)^2 + \sum\sum (\bar{Y}_j - \hat{Y}_{ij})^2 + \sum\sum 2(Y_{ij} - \bar{Y}_j)(\bar{Y}_j - \hat{Y}_{ij}) = \\
&= \sum\sum (Y_{ij} - \bar{Y}_j)^2 + \sum\sum (\bar{Y}_j - \hat{Y}_{ij})^2 + 0 = \qquad \text{(from prep-work)} \\
&= \sum\sum (Y_{ij} - \bar{Y}_j)^2 + \sum\sum (\bar{Y}_j - \hat{Y}_{ij})^2
\end{aligned} \tag{5}$$

$$\therefore \sum\sum (Y_{ij} - \hat{Y}_{ij})^2 = \sum\sum (Y_{ij} - \bar{Y}_j)^2 + \sum\sum (\bar{Y}_j - \hat{Y}_{ij})^2$$

Problem 3.23

A linear regression model with intercept $\beta_0 = 0$ is under consideration. Data have been obtained that contain replications. State the full and reduced models for testing the appropriateness of the regression function under consideration. What are the degrees of freedom associated with the full and reduced models if $n = 20$ and $c = 10$?

Full model: $Y_{ij} = \mu_{ij} + \epsilon_{ij}$ with $df_F = n - c = 20 - 10 = 10$
$\boxed{Y_{ij} = \mu_{ij} + \epsilon_{ij}, df_F = 10}$
Reduced model: $Y_{ij} = \beta_0 + \beta_1 X_1 + \epsilon_{ij} = \beta_1 X_1 + \epsilon_{ij}$ with $df_R = n - 1 = 19$
$\boxed{Y_{ij} = \beta_1 X_1 + \epsilon_{ij}, df_R = 19}$

A R Code for creating the SSE of the Box-Cox procedures with varying values of λ:

```
# Resource for gmean solution:
# http://onlinestatbook.com/2/transformations/box-cox.html

my_data <- read.csv("~/3-17.csv")

model_1 <- lm(Y~X, data = my_data)        # fit the linear model
                                          # calculate geometric mean of response
gmean <- exp(mean(log(my_data$Y)))        # g = e^(log(y)/n)

sse <- c()                                # empty list to store SSE's
lambda<-c(.3, .4, .5, .6, .7)             # list with lambda values
i <- 1
for (l in lambda){                 # loop through the list of lambdas to get SSE's
  if (l != 0){                     # y' = (y^lambda - 1)/ (lambda*g^(lambda-1))
  Y_prime <- (new_data$Y^l - 1) / (l*gmean^(l-1))
  } else {                                # when lambda = 0
  Y_prime <- gmean*log(my_data$Y)         # y' = g*log(y)
  }
  test <- anova(lm(Y_prime ~ my_data$X))
  sse[i] <- test['Residuals','Sum Sq']
  i <- i+1
}

cbind(lambda, sse)
```