HW2

Robert Chandler

2024-06-23

Setup

Begin by reading the dataset for this set of problems:

```
le <- read.csv("../../datasets/life_expectancy.csv")
```

```
x <- le$Nurses
y <- le$X2015Life.expectancy
```

Problem 1

Using the R-generated summary and ANOVA table for the model $Y \sim X$, answer the following questions.

First, set up X, Y:

```
n <- length(x)
model <- lm(y ~ x)
model_summary <- summary(model)
model_anova <- anova(model)
alpha <- 0.05
t_star <- model_summary$coefficients["x", "t value"]
t_crit <- qt(1 - alpha / 2, n - 2)
p <- 2 * (1 - pt(t_star, n - 2))

model_summary
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0640  -3.7798   0.2097   4.6965  13.4882
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 66.06723    0.66064  100.00   <2e-16 ***
## x            0.11834    0.01012   11.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.103 on 176 degrees of freedom
## Multiple R-squared:  0.4373, Adjusted R-squared:  0.4341
```

```
## F-statistic: 136.8 on 1 and 176 DF,  p-value: < 2.2e-16
```

```
model_anova
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## x           1 5093.0  5093.0  136.75 < 2.2e-16 ***
## Residuals 176 6554.6    37.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.a

For a two-sided hypothesis test on the linear impact, $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$ if a T-test is used, the test statistic is computed with the formula: $t^* = \frac{b_1 - \beta_1}{s\{b_1\}} = \frac{b_1}{s\{b_1\}}$, which is computed as 11.69423; The critical value (*assuming* $\alpha = 0.05$) has the notation of $t(1 - \alpha/2, n - 2) = t(1 - 0.025, 178 - 2)$, and a value of 1.9735344.

The p-value of the test can be computed with the formula $P[t(n-2) < -t^*] + P[t(n-2) > t^*] = 2P[t(n-2) > t^*]$, which is double the one-sided p-value. As shown above, to calculate this in R, we can use the `pt()` function to calculate the CDF of the t-distribution at our t^* value, which yields the complement of the one-sided p-value, so we can then subtract the value from 1 to get the one-sided p-value and double it to get the two-sided value. The value is 0.

1.b

The answer can be verified by observing the `Coefficients` section of the summary table, at the intersection of row `x` and column `Pr(>|t|)`:

```
model_summary$coefficients
```

```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 66.0672276  0.6606387 100.00509 5.209812e-157
## x            0.1183351  0.0101191  11.69423  9.599518e-24
```

The value is $9.5995182 \times 10^{-24}$, which is effectively 0.

1.c

Adjust the HT components from a two-sided test to a one-sided test. Consider the one-sided HT $H_0 : \beta_1 = 0, H_a : \beta_1 > 0$, the test statistic is the same as the two-sided test, but the p-value needs to be adjusted with the formula $P[t(n-2) > t^*]$ and computed as 0 in this question.

Problem 2

For a two-sided hypothesis test on the significance of the linear correlation coefficient between X and Y, $H_0 : \rho = 0, H_a : \rho \neq 0 \ldots$

2.a

If a T-test is used, the test statistic is computed with the sample correlation, r, with the formula

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which is computed as

```
r <- cor(x, y)
n <- length(x)
t_star <- r * sqrt(n - 2) / sqrt(1 - r^2)
```

$$t^* = \frac{0.6612558\sqrt{178-2}}{\sqrt{1-0.6612558^2}} = 11.69423$$

The critical t-value is

```
t_crit <- qt(1 - alpha / 2, n - 2)
```

$$t(1-\alpha/2, n-2) = t(0.975, 176) = 1.9735344 < 11.69423 = t^*$$

so we can reject H_0 and conclude that X, Y are linearly associated.

2.b

Is this test statistic the same as the t-test in part (a)? **YES**.

2.c

Discuss when the results of the hypothesis test on the linear impact and the linear association are equivalent.

Tests on linear impact and linear association are equivalent when the population is bivariate normal, but the regression model still holds even if they are not bivariate normal, so long as the conditional distributions $Y_i|X_i$ are normal and independent, with conditional means $\beta_0 + \beta_1 X_i$ and conditional variance σ^2 and the X_i are independent random variables whose distribution does not depend on $\beta_0, \beta_1, \sigma^2$. In this case, the results of the hypothesis tests are still equivalent.

2.d

Use R to compute a 95% confidence interval for the linear correlation coefficient between Y and X. Use the confidence interval to verify the hypothesis test in (c). (hint: if the confidence interval contains the hypothesized value, then the two-sided hypotheses should be rejected or not?)

We need to transform our data using a Fisher z transformation:

$$z' = \frac{1}{2} log_e(\frac{1+r_{12}}{1-r_{12}})$$
$$\sigma^2\{z'\} = \frac{1}{n-3}$$

```
alpha <- 0.05
z_prime <- 0.5 * log((1 + r) / (1 - r))
sigma_z_prime <- 1 / sqrt(n - 3)
z <- qnorm(1 - alpha / 2)
z_prime_upper <- z_prime + z * sigma_z_prime
z_prime_lower <- z_prime - z * sigma_z_prime
r_upper <- (exp(2 * z_prime_upper) - 1) / (exp(2 * z_prime_upper) + 1)
r_lower <- (exp(2 * z_prime_lower) - 1) / (exp(2 * z_prime_lower) + 1)
```

The transformed confidence interval for ρ is then

$$z' \pm z(1-\alpha/2)\sigma\{z'\} = 0.7950419 \pm z(1-0.05/2)0.0755929 = (0.9432013, 0.6468826)$$

and transforming back using $r = (e^{2z'} - 1)/(e^{2z'} + 1)$, the final interval is

$$(0.7366896, 0.5695676)$$

and since $r = 0$ does not fall within this interval, we reject H_0 and conclude that the data are linearly associated.

Problem 3

Consider a simple linear regression model with $Y \sim X$ on the following table. Practice doing the lack-of-fit by hand.

```
library(knitr)
data_p3 <- data.frame(
  x = c(0.24, 0.22, 0.23, 0.24, 0.24, 0.24, 0.22, 0.21, 0.24, 0.21),
  y = c(16, 40, 32, 13, 1, 1, 2, 3, 8, 14)
)
x <- data_p3$x
y <- data_p3$y
kable(data_p3, caption = "Problem 3 Dataset")
```

Table 1: Problem 3 Dataset

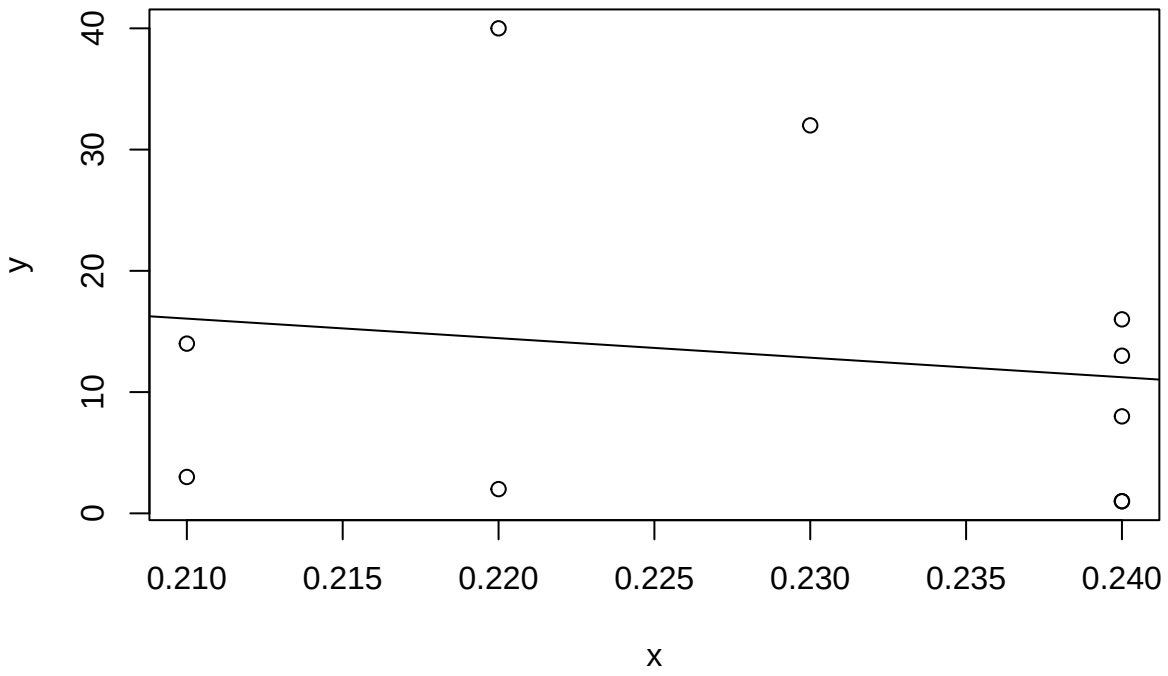
x	y
0.24	16
0.22	40
0.23	32
0.24	13
0.24	1
0.24	1
0.22	2
0.21	3
0.24	8
0.21	14

3.a

Based on a (R-generated) scatter plot of X and Y with the regression line, comment on whether the Simple Linear Regression (SLR) exhibits a lack-of-fit issue.

Plot the SLR curve of Y regressed on X:

```
model <- lm(y ~ x)
plot(y ~ x)
abline(model)
```



This is a difficult case to judge at a first glance. First of all, we note that X does not seem to have much of an impact on Y. Regarding the goodness of fit, the points do seem to have a decent amount of scatter about the line, but this could just imply a high variance in the error. It does not necessarily seem that a linear model is incorrect, but it is still tough to judge by eye, so this is a great use case for the lack-of-fit test.

3.b

Compute the components for the lack of fit test: c, $\hat{Y}_i$, $\bar{Y}_i$, $\bar{Y}$, $SSPE$, $SSLF$, SSE, $DFPE$, $DFLF$, DFE.

Extract coefficients from the model first:

```
b0 <- model$coefficients[["(Intercept)"]]
b1 <- model$coefficients[["x"]]
```

Compute requested values:

```
n <- length(x)

df_unq <- aggregate(y ~ x, data_p3, mean)
colnames(df_unq)[2] <- "ybar"

c <- length(df_unq$x)
df_unq$yhat <- sapply(df_unq$x, function(x) b0 + b1 * x)

ybar_overall <- mean(y)

# Build up SS values iteratively
sspe <- 0
```

```r
sslf <- 0
sse <- 0
for (i in seq_along(y)) {
  # Find the ybar and yhat values associated with the current y... this takes
  # care of the summation across the i index since these are already
  # pre-calculated
  ybar_i <- df_unq$ybar[df_unq$x == x[i]]
  yhat_i <- df_unq$yhat[df_unq$x == x[i]]

  # This is good, but FIXME on yhat
  spe <- (y[i] - ybar_i)^2
  sspe <- sspe + spe

  slf <- (ybar_i - yhat_i)^2
  sslf <- sslf + slf

  se <- (y[i] - yhat_i)^2
  sse <- sse + se
}

# Ensure SSPE + SSLF = SSE
stopifnot(sspe + sslf - sse < 1e-10)

dfpe <- n - c
dflf <- c - 2
dfe <- n - 2
```

The resulting values are:

$$
\begin{aligned}
c &= 4 \\
\hat{Y}_i &= 16.0604027, 14.4496644, 12.8389262, 11.2281879 \\
\bar{Y}_i &= 8.5, 21, 32, 7.8 \\
\bar{Y} &= 13 \\
SSPE &= \sum\sum(Y_{ij} - \bar{Y}_i)^2 = 969.3 \\
SSLF &= \sum\sum(\bar{Y}_i - \hat{Y}_{ij})^2 = 626.0422819 \\
SSE &= \sum\sum(Y_{ij} - \hat{Y}_i j)^2 = 1595.3422819 \\
DFPE &= 6 \\
DFLF &= 2 \\
DFE &= 8
\end{aligned}
$$

3.c

Next, consider the lack-of-fit test for the SLR. Define H_0, H_a, calculate test statistic, define reject region, compute the p-value, and state the conclusion.

The hypotheses are as follows:

$$
\begin{aligned}
H_0 &: E\{Y\} = \beta_0 + \beta_1 X \\
H_a &: E\{Y\} \neq \beta_0 + \beta_1 X
\end{aligned}
$$

The test statistic is

```
mslf <- sslf / (c - 2)
mspe <- sspe / (n - c)
f_star <- mslf / mspe
```

$$F^* = \frac{SSE - SSPE}{(n-2)-(n-c)} \div \frac{SSPE}{n-c} = \frac{SSLF}{c-2} \div \frac{SSPE}{n-c} = \frac{626.0422819}{4-2} \div \frac{969.3}{10-4} = 1.9376115$$

and the critical F value is

```
f_crit <- qf(1 - alpha, c - 2, n - c)
```

$$F(1-\alpha; c-2, n-c) = 5.1432528$$

The rejection region is:

$$\text{If } F^* \leq F(1-\alpha; c-2, n-c), \text{ fail to reject } H_0$$
$$\text{If } F^* > F(1-\alpha; c-2, n-c), \text{ reject } H_0$$

The p-value is

```
p_f_star <- 1 - pf(f_star, c - 2, n - c)
```

$$p = P\{F(c-2, n-c) > F^*\} = P\{F(2,6) > 1.9376115\} = 0.2242916$$

so, because F^* and p both fall outside of the rejection region, we fail to reject H_0, and we conclude that **the model does not have a lack-of-fit problem** at $\alpha = 0.05$.

3.d

Utilize R to conduct the lack-of-fit test and identify as many components as possible from the ones computed in part (b) in the R output.

Our hand-calculated values above are validated in the ANOVA table below:

```
anova(model, lm(y ~ as.factor(x), data_p3))
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ as.factor(x)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      8 1595.3
## 2      6  969.3  2    626.04 1.9376 0.2243
```

The values for SSE and $SSPE$ are readily found in the first and second rows of the `RSS` column, respectively, and $SSLF$ is in the `Sum of Sq` column. The different DOF values are all in the table: DFE and $DFPE$ are in the first and second rows of the `Res.Df` column, respectively, and $DFLF$ is in the `Df` column. c can be found indirectly from $c = 2 + DFLF$. The rest of the values in 3.b are derived from Y and are not shown in the table.

Problem 4

Consider a lack of fit test on the following data on $Y \sim X$:

```
data_p4 <- data.frame(
  x = c(0.19, 0.44, 0.35, 0.32, 0.29),
  y = c(65, 30, 22, 31, 9)
)
kable(data_p4)
```

x	y
0.19	65
0.44	30
0.35	22
0.32	31
0.29	9

4.a

Can you perform a lack of fit test on this data?

As the data currently stands, a lack of fit test cannot be performed on it because there are no replicate values in X. However, we can create a kind of pseudo-replicate dataset by grouping the data if we needed to perform the test and could not gather any new samples.

4.b

Suppose a new row is added: $X = 0.19, Y = 59$ and the sample size is now 6. Before, the $SSPE$ was simply 0 because we had no replicates, and now it is just a function of the $X = 0.19$ terms:

```
data_p4 <- rbind(data_p4, c(0.19, 59))
y_i <- data_p4[data_p4$x == 0.19, "y"]
sspe <- sum(sapply(y_i, function(y) (y - mean(y_i))^2))
```

$$SSPE = (Y_{ij} - \bar{Y}_i)^2 = 18$$

so $SSPE$ has **increased** by 18.

```
n <- nrow(data_p4)
c <- length(unique(data_p4$x))
dfpe <- n - c
dflf <- c - 2
```

$DFPE = n - c = 1$ and $DFLF = c - 2 = 3$.

4.c

Suppose the data is grouped by the tenth digits as follows,

```
data_p4 <- data.frame(
  x = c(0.1, 0.4, 0.3, 0.3, 0.2),
  y = c(65, 30, 22, 31, 9)
)
kable(data_p4)
```

x	y
0.1	65
0.4	30
0.3	22
0.3	31
0.2	9

Then (when compared against original table given at the start of the problem), the $SSPE$ is **increased** by...

```
anova_results <- anova(lm(y ~ x, data_p4), lm(y ~ as.factor(x), data_p4))
sspe <- anova_results[2, "RSS"]
dfpe <- anova_results[2, "Res.Df"]
dflf <- anova_results[2, "Df"]
```

$$SSPE = (Y_{ij} - \bar{Y}_i)^2 = 40.5$$

$SSPE$ has **increased** by 40.5 from its original value of 0; the $DFPE = n - c = 1$, and the $DFLF = c - 2 = 2$.

Problem 5

Utilize the life expectancy data and examine a simple linear regression $Y \sim X$, where $X = X_3$ (pharmacists), and Y represents life expectancy. Employ R to assess assumptions for the SLR.

Conduct screening through the scatter plot, residual plot on X, residual plot on $\hat{Y}$, Shapiro test, Brown-Forsythe test, Breusch-Pagan test, and provide comments on:

- Linear relationship
- Constant variance
- Normal errors
- Outliers

First, update X, Y and generate the SLR model:

```
x <- le$Pharmacists
y <- le$X2015Life.expectancy
model <- lm(y ~ x)
```

Scatter Plot and Regression Line

```
plot(y ~ x)
abline(model)
```

From the scatter plot, it appears that X does seem to have a moderate impact on Y as we observe a pattern of slight increase in Y with an increase in X. The amoult of data is much more dense at X values near 0 than elsewhere. There appears to be a slight amount of heteroskedasticity occurring with a higher variance at lower X values, but this could just be a misconception due to the more sparsely distributed data at higher X values. It also appears possible that the underlying population could be curvilinear in nature, but once again there are so few points at the higher X values that this could be a misinterpretation.

Residual Plot on X

```
plot(model$residuals ~ x, ylab = "Residuals")
abline(h = 0)
```

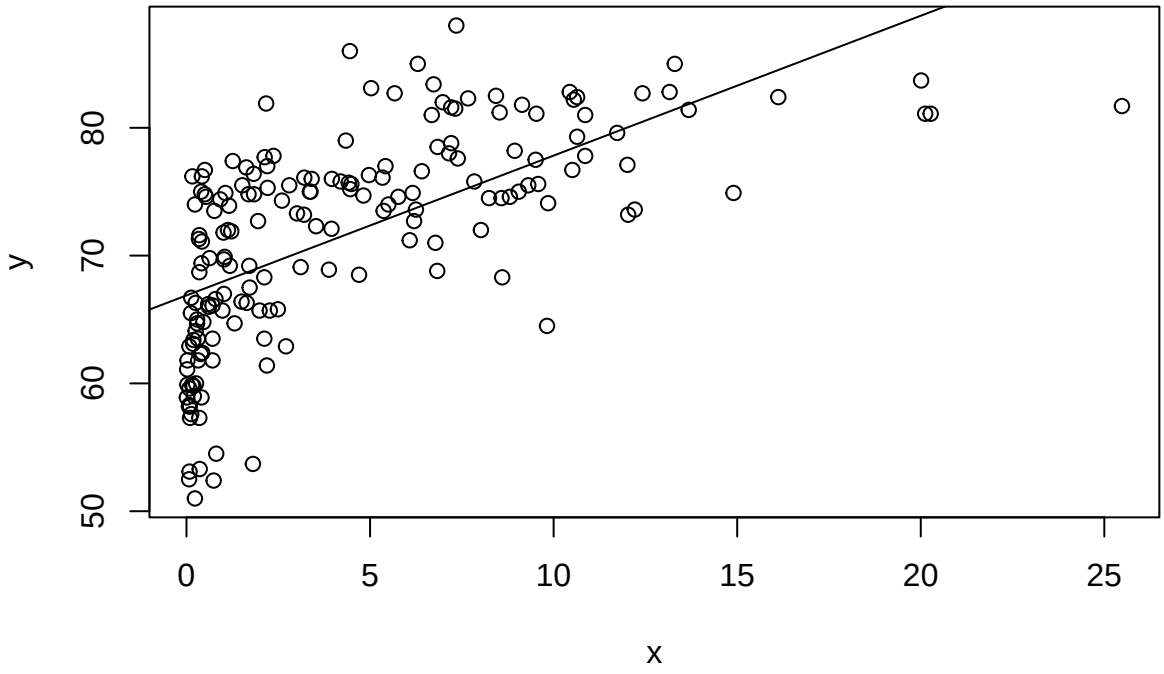


Figure 1: Regression Line and Scatter Plot

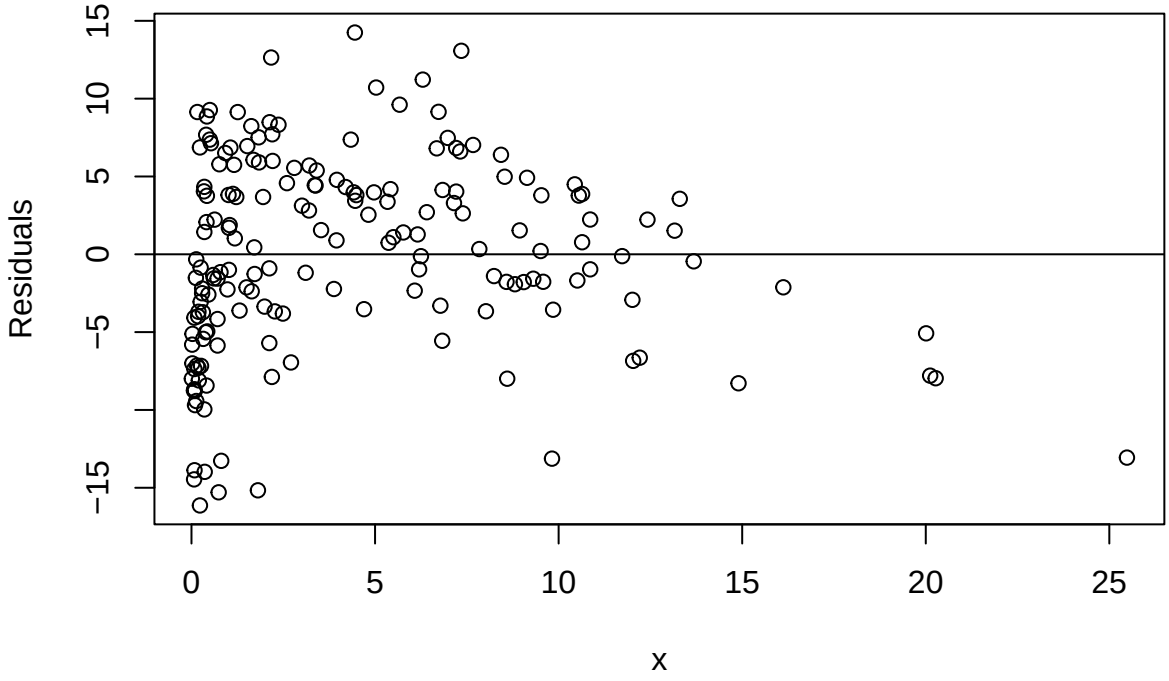


Figure 2: Residuals Plot

The residuals plot reinforces the conclusions we drew from the previous plot, however it does show that there may be a slight positive skew in the residuals around the $X = 2.5$ to $X = 7.5$ range. And then they start to skew in the negative direction. This may be indicative of a curvilinear population distribution. Once again, the variance does seem to decrease as X increases. There do not seem to be any extremely obvious outliers.

Residual Plot on the Fitted Values $\hat{Y}$

```
plot(
  model$residuals ~ model$fitted.values,
  xlab = "Fitted y-values", ylab = "Residuals"
)
abline(h = 0)
```

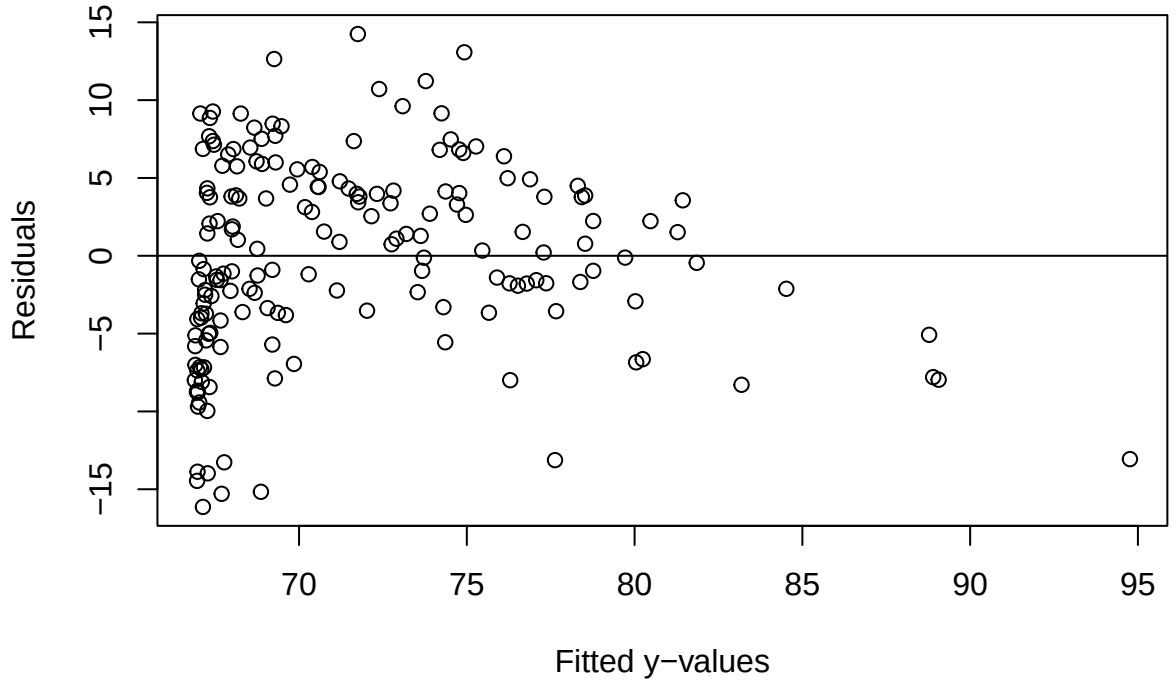


Figure 3: Residuals Plot Against Fitted Values

This is the same plot as the previous one, just with a different scale on the x-axis since the $\hat{Y}_i$ are a linear function of the X_i values, so though the scale is different, the pattern remains the same, which our R plots show quite clearly thanks to their automatic re-scaling.

Shapiro-Wilk Test for Normality

Our hypotheses are:

$$H_0: \text{ residuals follow a normal distribution}$$
$$H_a: \text{ residuals deviate significantly from a normal distribution}$$

```
shapiro_results <- shapiro.test(model$residuals)
shapiro_results
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.98335, p-value = 0.03218
```

$p = 0.0321766 < 0.05 = \alpha$, so we reject H_0 and conclude that the data does not follow a normal distribution. This makes sense given that we do not see the residuals more tightly distributed about the $e = 0$ line, tapering out as they move away from it, but rather they seem to be scattered somewhat uniformly.

Brown-Forsythe Test

Our hypotheses are:

$$H_0: \text{ residuals have constant variance}$$
$$H_a: \text{ residuals have non-constant variance}$$

The tricky part about this test is determining where to split the data into two groups. If we choose a central measure of X as the value at which to split, such as the mean (4.3483989) or median (2.32), we will have a two groups with a relatively even number of samples in them, but the range of the group to the left of the split will be much smaller than that of the group on the right since we have so many more measurements at small X values. If we try to split the groups to have equal ranges at a value like 12, then the group on the right will have significantly less samples than the one on the left. To compromise, we choose to split at a value of 7, striking a balance between the two issues.

```
library(ALSM)
x_split <- 7
groups <- x < x_split
bf_results <- bftest(model, groups)
bf_results
```

```
##       t.value    P.Value alpha  df
## [1,] 1.834897 0.06820907  0.05 176
```

$p = 0.0682091 > 0.05 = \alpha$, so we fail to reject H_0 and we cannot say that the residuals exhibit heteroskedasticity. However, the p value is quite close to α and if we were to choose a value closer to the center of the data, like 6, we observe different results:

```
x_split <- 6
groups <- x < x_split
bf_results <- bftest(model, groups)
bf_results
```

```
##       t.value   P.Value alpha  df
## [1,] 2.009946 0.0459642  0.05 176
```

This time, $p = 0.0459642 < 0.05 = \alpha$, so we *reject* H_0 and conclude that the data *do* exhibit heteroskedasticity.

Since our 2-group BF test is somewhat ambiguous, we can perform another with 4 groups:

```
library(onewaytests)
df_bf <- data.frame(
  group = cut(x, 4),
  residual = model$residuals
```

```
)
bf_4_results <- bf.test(residual ~ group, df_bf)
```

```
##
##   Brown-Forsythe Test (alpha = 0.05)
## ------------------------------------------------------------
##   data : residual and group
##
##   statistic  : 5.475029
##   num df     : 3
##   denom df   : 26.77801
##   p.value    : 0.004562011
##
##   Result     : Difference is statistically significant.
## ------------------------------------------------------------
```

This time around, our p value is much smaller, we reject H_0, and conclude that the residuals are heteroskedastic. Even still, two of our groups in this test had less than 5 samples in them:

```
table(df_bf$group)
```

```
##
## (-0.0155,6.38]    (6.38,12.7]    (12.7,19.1]    (19.1,25.5]
##            126             43              5              4
```

so we would prefer to collect more data at the higher X values before making any strong conclusions about the data, but we will tentatively conclude that the residuals are heteroskedastic based on this test.

Breusch-Pagan Test

Once again, our hypotheses are:

$$H_0: \text{ residuals follow a normal distribution}$$
$$H_a: \text{ residuals deviate significantly from a normal distribution}$$

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bp_results <- bptest(model)
bp_results
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 1.4969, df = 1, p-value = 0.2211
```

$p = 0.2211469 > 0.05 = \alpha$, so we fail to reject H_0 and we cannot say that the residuals exhibit heteroskedasticity based on this test.

Comments

Linear Relationship

Considering the linear relationship between Y and X, it seemed from the Scatter Plot and Regression Line that X had an impact on Y, which we can confirm with an ANOVA analysis:

```
anova_results <- anova(model)
anova_results
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## x           1 4727.6  4727.6  120.24 < 2.2e-16 ***
## Residuals 176 6920.0    39.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is very low here, so we can conclude that X does contribute to a reduction in the variance of Y.

Regarding the goodness of the linear fit, we noted previously that it seems possible the population may be curvilinear. To test this, we can perform a lack-of-fit test with hypotheses:

$$H_0 : E\{Y\} = \beta_0 + \beta_1 X$$
$$H_a : E\{Y\} \neq \beta_0 + \beta_1 X$$

```
anova_results <- anova(model, lm(y ~ as.factor(x)))
anova_results
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ as.factor(x)
##   Res.Df    RSS  Df Sum of Sq      F Pr(>F)
## 1    176 6920.0
## 2     18  542.9 158    6377.1 1.3382 0.2422
```

$p = 0.242194 > 0.05 = \alpha$, so we fail to reject H_0 and we cannot conclude that there is a lack of fit problem with the SLR.

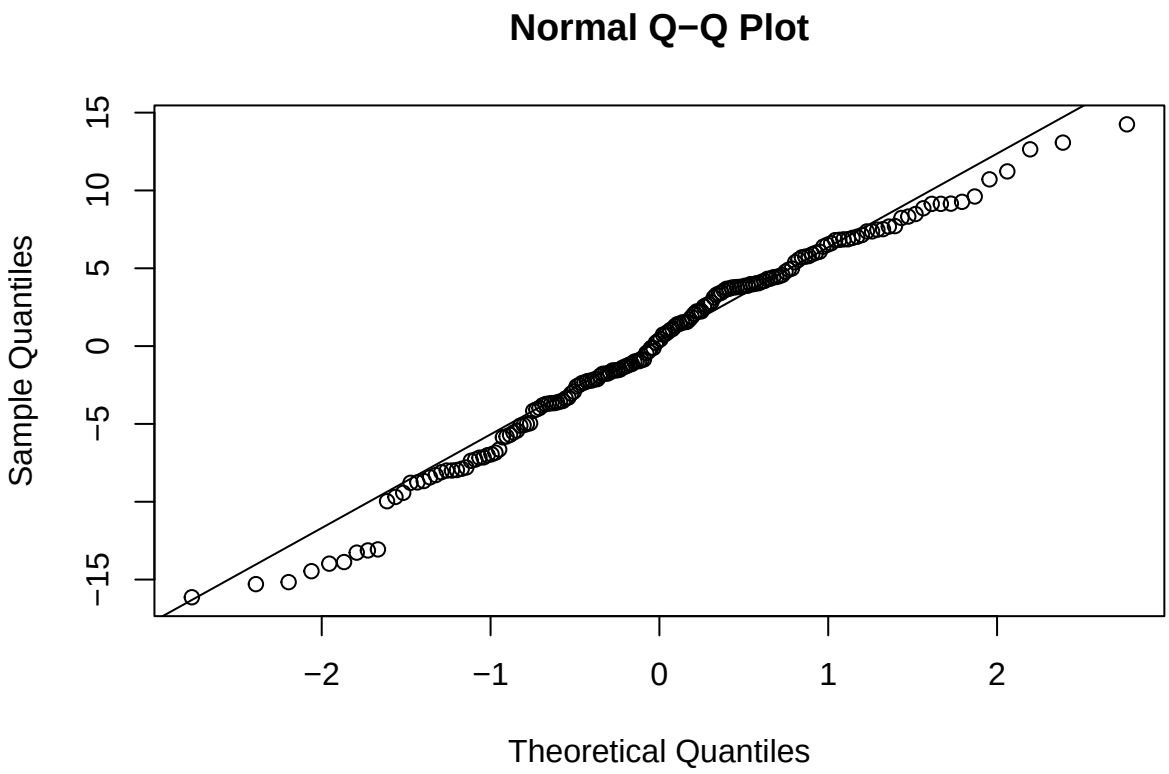
Constant Variance

As we covered thoroughly in the Brown-Forsythe Test and Breusch-Pagan Test sections above, the results of the various tests were ambiguous as to whether or not the residuals exhibit non-constant variance. The ideal solution would be to gather more samples in the larger X regions and re-compute these tests.

Normal errors

In the Shapiro-Wilk Test for Normality, we concluded that the residuals are not normally distributed. One other diagnostic tool we can use for this metric is the normal quantile-quantile plot:

```
qqnorm(model$residuals)
qqline(model$residuals)
```

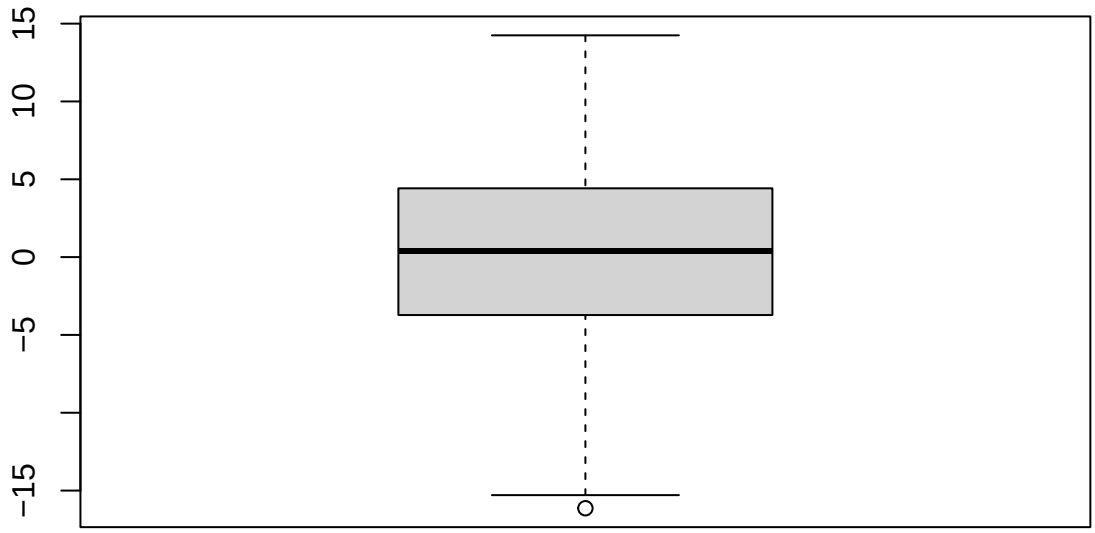


In this plot, we can see a fair amount of deviation from the ideal normal QQ line, especially at the more extreme theoretical quantile levels.

Outliers

Looking at the various residual plots above, there do not seem to be any extremely obvious outliers, but we can utilize a box plot to see whether it determines there are any:

```
boxplot(model$residuals)
```



The plot does show one outlier at the far-negative end of the residuals. This is typically determined by some metric like being below $Q_1 - 1.5IQR$, referring to the first quartile and the interquartile range $Q3 - Q1$.

Problem 6

Do you think it is necessary to transform X or Y?

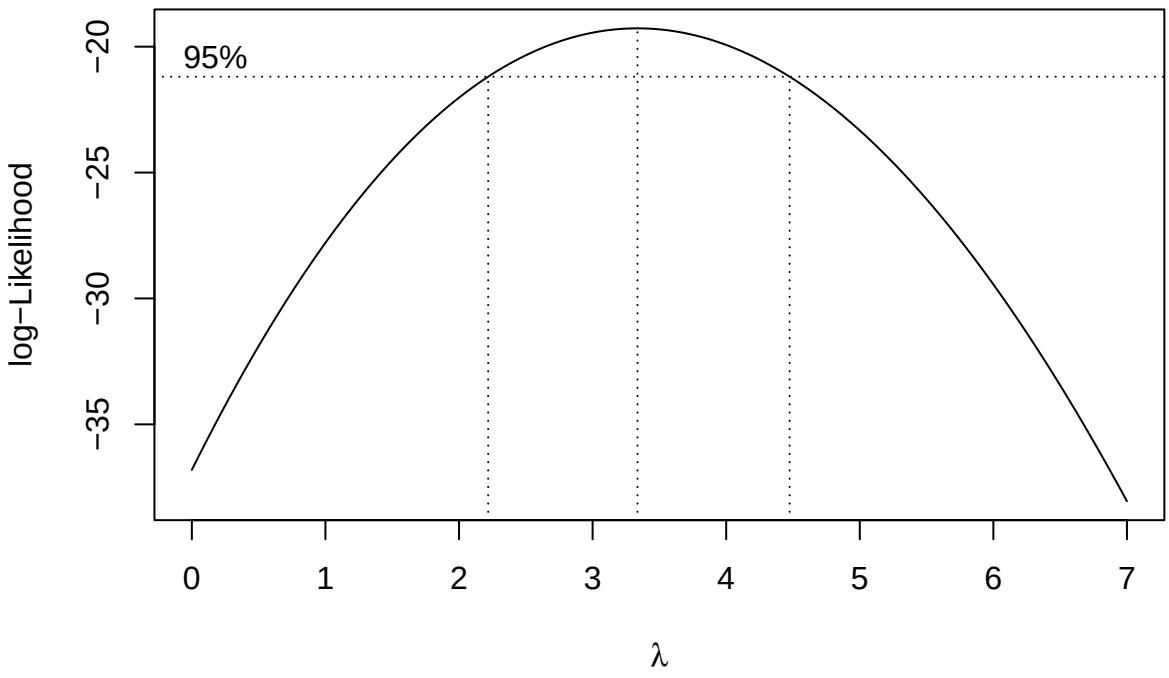
I do not think it is necessary to transform X or Y because the linear model seems to do an adequate job at representing the relationship between these two variables. The lack-of-fit test that we performed in the Linear Relationship section above showed that we could not conclude with statistical significance that there was a lack of fit in the SLR line, so we probably do not stand to gain much from transforming in this case. The one area where it might improve the model would be in making the variance more constant with respect to X.

Problem 7

Utilize R to conduct a Box-Cox transformation on Y, then proceed with the same diagnostic process as in Problem 5. Compare the transformed model with the original model.

```
library(MASS)

# 0.001 is overkill but still fast to compute
bc_mle <- boxcox(model, lambda = seq(0, 7, by = 0.001))
```




```
lambda_mle <- bc_mle$x[which.max(bc_mle$y)]
yt <- y^lambda_mle
```

$$\lambda = 3.336$$

Proceeding with the same diagnostic procedure as before, we will make comments on the plots and compare them to the pre-transformed ones, comparing the overall shape and relative distances in the plots between the two since the values are obviously not directly comparable in the different scales.

First, generate the new transformed SLR model:

```
model <- lm(yt ~ x)
```

Transformed Scatter Plot and Regression Line

```
plot(yt ~ x)
abline(model)
```

This plot looks similar to the pre-transformed one, but when comparing side by side it is possible to see where the values have been stretched and squeezed by the transformation. In particular, the values near $X = 0$ that were quite far from the regression line have been brought in closer to it, showing less (relative) variance overall and a better linear fit.

Transformed Residual Plot on X

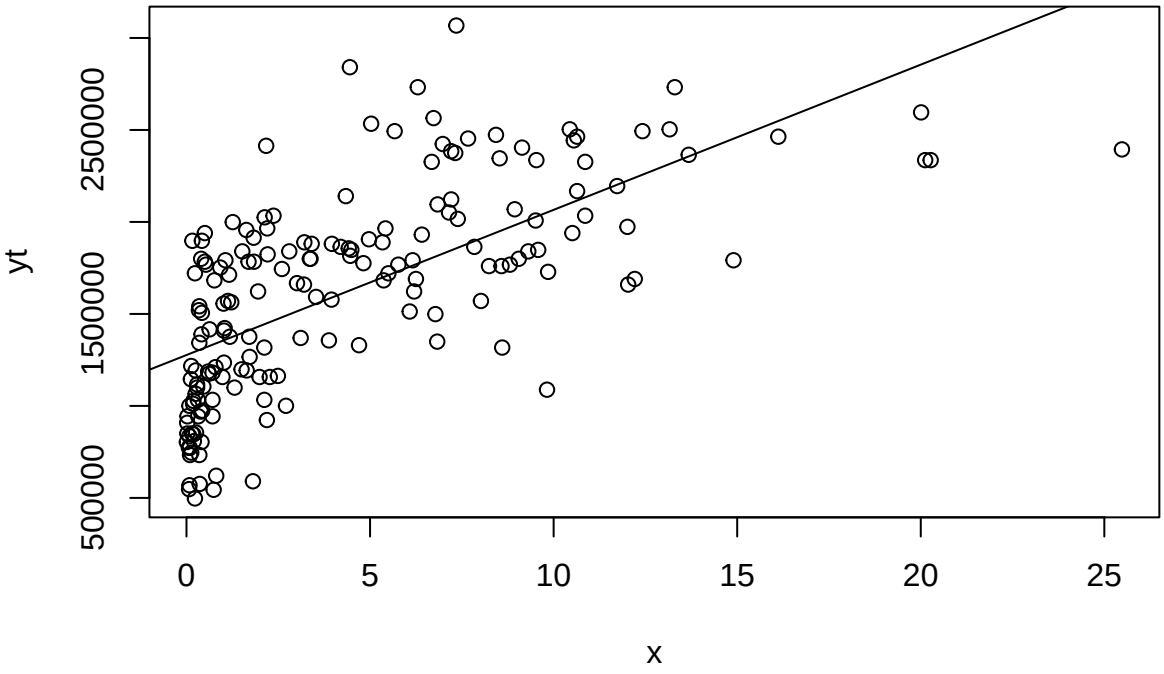


Figure 4: Transformed Regression Line and Scatter Plot

```r
plot(model$residuals ~ x, ylab = "Residuals")
abline(h = 0)
```

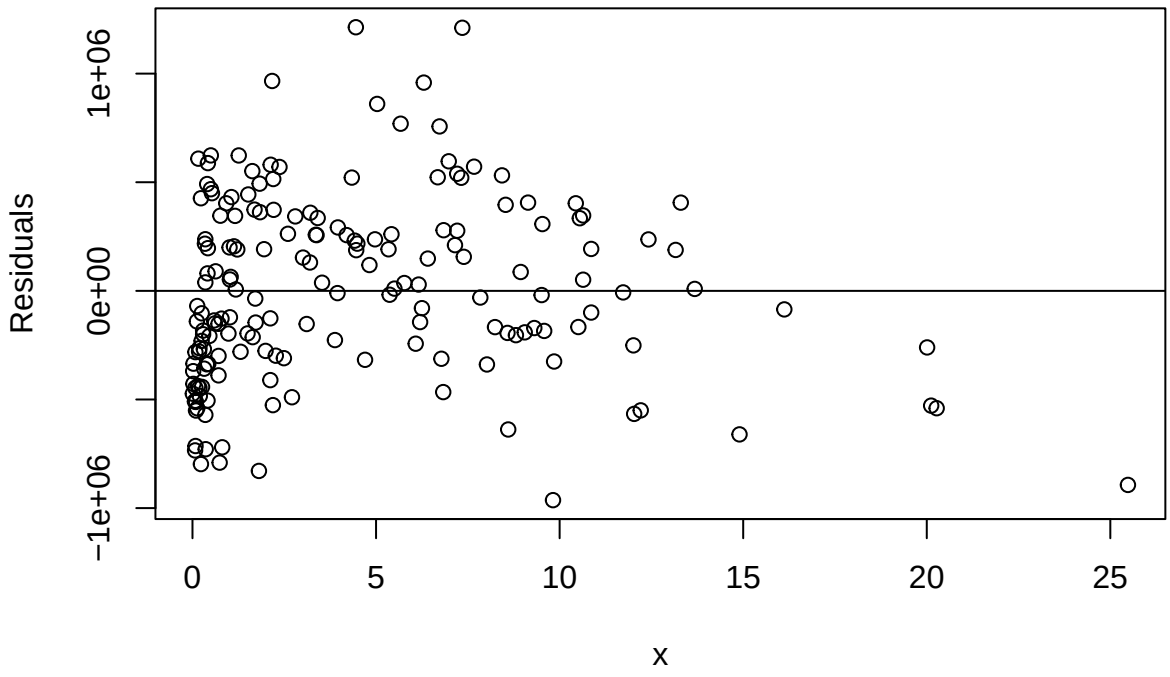


Figure 5: Transformed Residuals Plot

The residuals plot exhibits the same behavior noted in the scatter plot, particularly regarding the values near $X = 0$. The values seem to have more uniform variance throughout, but there does still seem to be a slight skew toward the positive residuals overall.

Transformed Residual Plot on the Fitted Values $\hat{Y}$

```r
plot(
  model$residuals ~ model$fitted.values,
  xlab = "Fitted transformed y-values", ylab = "Residuals"
)
abline(h = 0)
```

Like last time, this is a duplication of the previous plot.

Transformed Shapiro-Wilk Test for Normality

Our hypotheses are:

$$H_0: \text{ residuals follow a normal distribution}$$
$$H_a: \text{ residuals deviate significantly from a normal distribution}$$

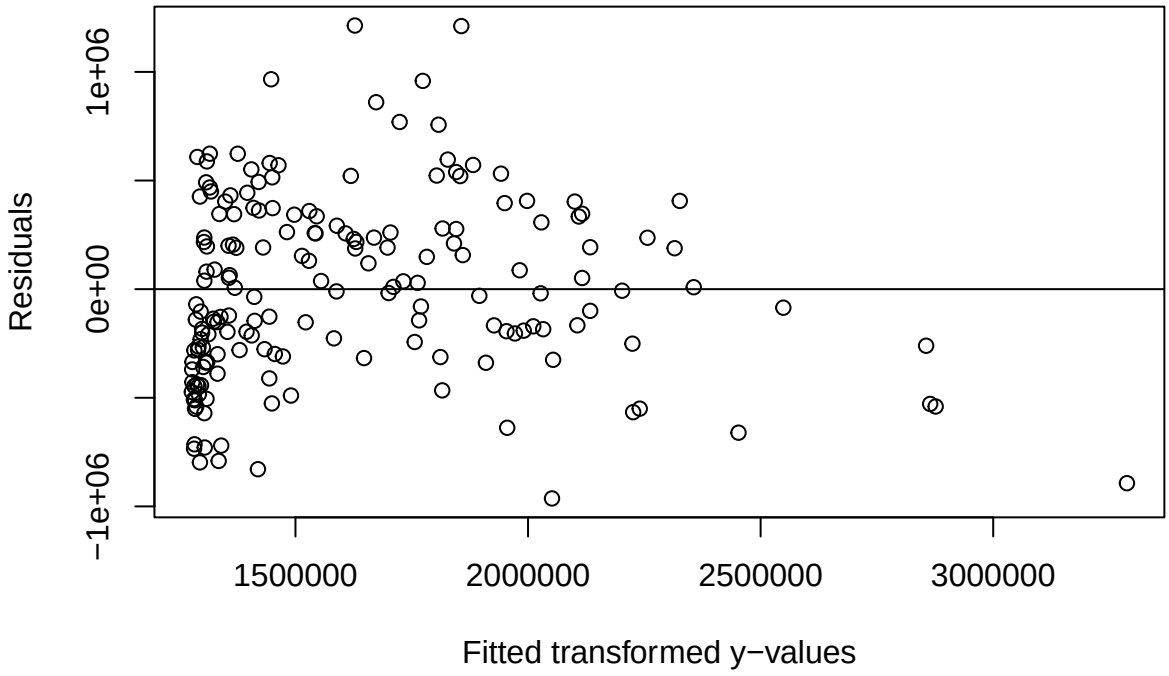


Figure 6: Residuals Plot Against Fitted Values

```
shapiro_results <- shapiro.test(model$residuals)
shapiro_results
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.98964, p-value = 0.2224
```

$p = 0.2224015 > 0.05 = \alpha$, so we fail to reject H_0 and we cannot conclude the data does not follow a normal distribution. This makes sense given that we saw the residuals tighten in toward the 0 line, mirroring a normal distribution more closely this time after the transformation.

Transformed Brown-Forsythe Test

Our hypotheses are:

$$H_0: \text{ residuals have constant variance}$$
$$H_a: \text{ residuals have non-constant variance}$$

We will start off with the same first split as previously at $X = 7$:

```
library(ALSM)
x_split <- 7
groups <- x < x_split
bf_results <- bftest(model, groups)
bf_results
```

```
##       t.value   P.Value alpha  df
## [1,] 0.208244 0.8352793  0.05 176
```

$p = 0.8352793 > 0.05 = \alpha$, so we fail to reject H_0 and we cannot say that the residuals exhibit heteroskedasticity. This time, the p value is nowhere near α and even if we drop down to a split at the median, we still fail to reject H_0. This indicates that the transform has done a good job at making the variance constant:

```
x_split <- median(x)
groups <- x < x_split
bf_results <- bftest(model, groups)
bf_results
```

```
##        t.value   P.Value alpha  df
## [1,] 0.3798522 0.7045131  0.05 176
```

We are satisfied with the results of the transformation as concerns heteroskedasticity, so we proceed on.

Transformed Breusch-Pagan Test

Once again, our hypotheses are:

$$H_0: \text{ residuals follow a normal distribution}$$
$$H_a: \text{ residuals deviate significantly from a normal distribution}$$

```
library(lmtest)
bp_results <- bptest(model)
bp_results
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 0.62765, df = 1, p-value = 0.4282
```

$p = 0.428217 > 0.05 = \alpha$, so we fail to reject H_0 and we cannot say that the residuals exhibit heteroskedasticity based on this test. Once again, the transform has done its job.

Comments

Linear Relationship

Considering the linear relationship between Y and X, it still seemed from the Transformed Scatter Plot and Regression Line that X had an impact on Y, which we can confirm with an ANOVA analysis:

```
anova_results <- anova(model)
anova_results
```

```
## Analysis of Variance Table
##
## Response: yt
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## x           1 2.4601e+13 2.4601e+13  137.33 < 2.2e-16 ***
## Residuals 176 3.1528e+13 1.7913e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is very low here, so we can conclude that X does contribute to a reduction in the variance of Y.

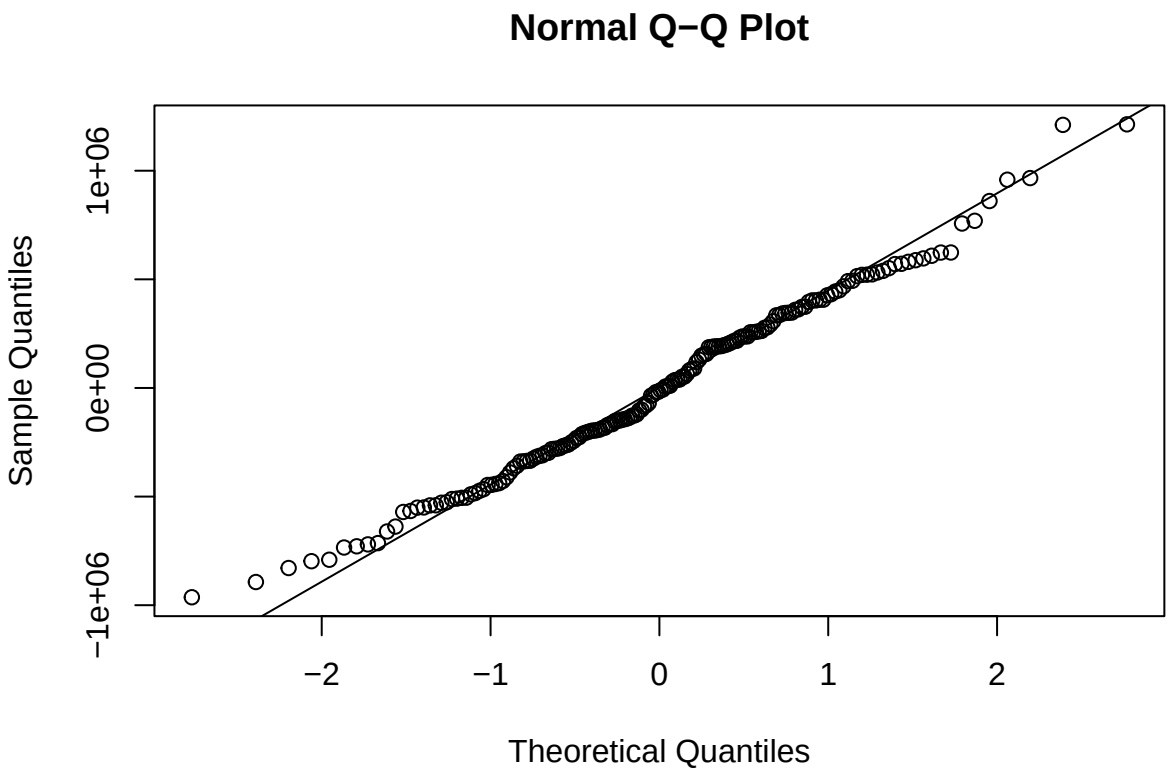
Constant Variance

As we covered in the Transformed Brown-Forsythe Test and [Tranformed Breusch-Pagan Test] sections above, the results of the tests support the conclusion that the transformation has kept us from concluding that the data is heteroskedastic anymore. In other words, the transformation has made the variance more constant as a function of X.

Normal errors

In the Transformed Shapiro-Wilk Test for Normality, we concluded that the residuals were normally distributed. One other diagnostic tool we can use for this metric is the normal quantile-quantile plot:

```
qqnorm(model$residuals)
qqline(model$residuals)
```

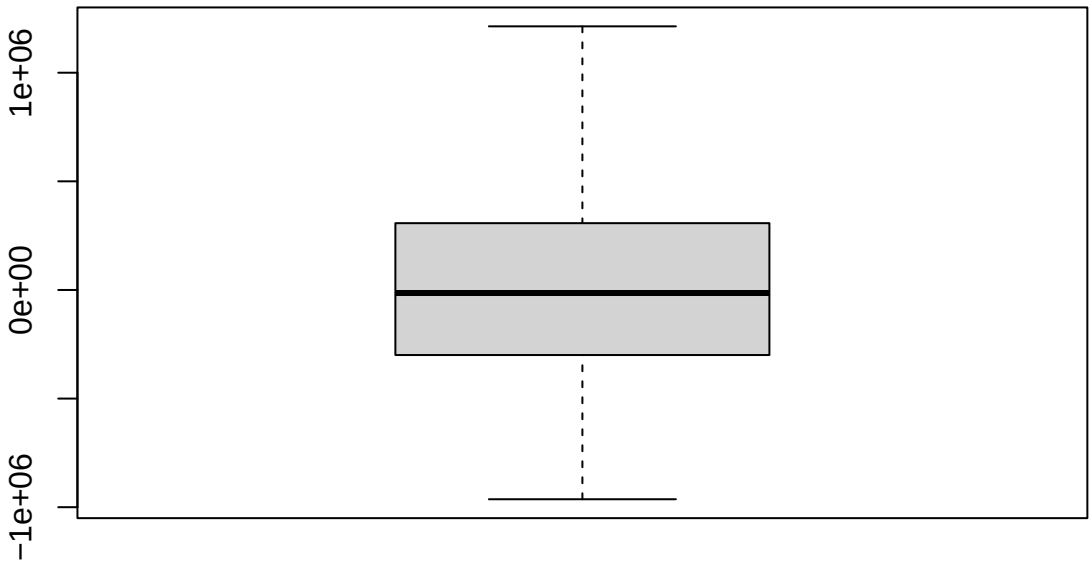


We can see how the plot has tightened up to the ideal normal line, especially when looking at the values in the more extreme theoretical quantiles.

Outliers

Looking at the various residual plots above, there do not seem to be any extremely obvious outliers, but we can utilize a box plot to see whether it determines there are any:

```
boxplot(model$residuals)
```



The one outlier observed previously has been eliminated due to the transform.

Conslusion

In conclusion, the transformation did not make any drastic changes to the data as the linear fit did a fair job in the first place. One area that was significantly improved was the heteroskedasticity of the data, with the transformation bringing it much closer to homoskedasticity. Given the extra work needed to inpterpret the data, especially if applying back-transformations, might not be worth the extra effort though.