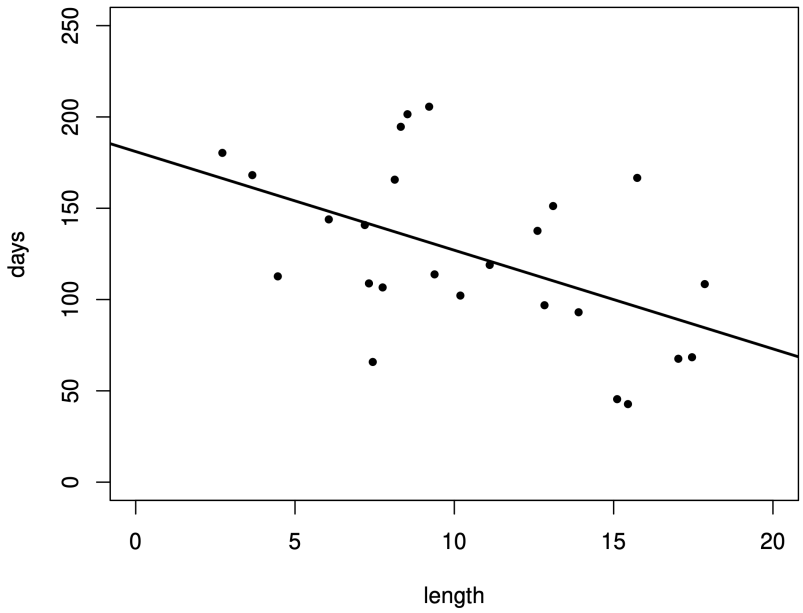
Carson Slater STA 5381 Homework #2

Beginning after the multiple choice questions, please start each problem on a new page.

A researcher claims that the length of a fetus (as measured on a ultrasound) and the number of gestational days remaining is negatively linearly related. He selects a SRS of 25 expectant mothers and measures the length of each one's fetus as well as the number of days until each gives birth. Below is some of the output from the analysis. Use this information to answer the following 5 multiple choice questions.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  181.043     21.929   8.256 2.49e-08 ***
length        -5.403      1.931  -2.798   0.0102 *

Analysis of Variance Table
Response: days
          Df Sum Sq Mean Sq F value  Pr(>F)
length     1  13599   13599   7.828 0.01022 *
Residuals 23  39957    1737
```



1. **What hypotheses should he use to test his claim?**

   (a) $H_0 : \beta_0 \geq 0 \quad H_a : \beta_0 < 0$

   (b) $H_0 : \beta_0 = 0 \quad H_a : \beta_0 \neq 0$

   (c) $H_0 : \beta_1 \geq 0 \quad H_a : \beta_1 < 0$

   (d) $H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0$

   (e) $H_0 : \hat{\beta}_1 \geq 0 \quad H_a : \hat{\beta}_1 < 0$

2. **What is the p-value he should use to test his claim?**

   (a) 2.49×10^{-8}
   (b) 0.0102
   (c) 0.9949
   (d) 0.0051
   (e) None of the above.

3. **What proportion of the variability in "days to birth" is explained by the fetal length?**

```
SSReg <- 13599
SST <- 39957

(rsq <-  (SSReg/(SSReg + SST)) |> round(digits = 4))

## [1] 0.2539
```

   (a) 7.83%
   (b) 25.39%
   (c) 34.03%
   (d) 74.61%
   (e) 88.67%

4. **The fetal length for another expectant mother (one who is not part of the original 25) is recorded to be 15.3 inches. Which of the following values might the expectant mother be least interested in? Explain your choice.**

   (a) The 95% confidence interval for average "days to birth" of a 15.3 inch fetus.
   (b) The 95% prediction interval for a particular "days to birth" of a 15.3 inch fetus.
   (c) The point estimate of "days to birth."
   (d) The proportion of variability in "days to birth" that is explained by "fetal length."
   (e) The sample size of the study used to generate the relationship between "days to birth" and "fetal length."

5. **The fetal length for another expectant mother (one who is not part of the original 25) is recorded to be 15.3 inches. Which of the following values might a doctor who is using the results of the study be least interested in? Explain your choice.**

   (a) The 95% confidence interval for average "days to birth" of a 15.3 inch fetus.
   (b) The 95% prediction interval for a particular "days to birth" of a 15.3 inch fetus.
   (c) The point estimate of "days to birth."
   (d) The proportion of variability in "days to birth" that is explained by "fetal length."
   (e) The sample size of the study used to generate the relationship between "days to birth" and "fetal length."

Multiple Choice Answers

1. C
2. D
3. B
4. A - The mother is probably least interested in the aggregate. She cares if the model used to generate what would predict her "days to birth" fulfills modeling assumptions, and she is a particular case, and would be able to have more variable but firm expectations. Hence she is probably more interested in the prediction interval than the confidence interval, which is is interested in the least.
5. B - The doctor is a person who treats many expectant mothers, and wants a reliable model for 'days to birth' based on the length of the fetus. In this case, they are concerned about how good the model estimates the aggregate. Therefore this doctor is most likely concerned about what makes this model predict the aggregate well, and the prediction interval has the least to do with this criterion.

6. Use the dataset "diamonds.txt" to assess the relationship between carat size of a diamond and its price.

```
diamonds <- read.delim("diamonds.txt",
                       sep ="", header = TRUE, dec =".")
mod1 <- lm(Price ~ Carat, data = diamonds)
standard_res <- rstandard(mod1)
```

(a) Which variable is the explanatory and which is the response?

In this context, the explanatory variable is the carat size and the response variable is the price of the diamond.

(b) Assess the assumptions of the SLR model. Show whatever is necessary to justify your answer.

First, we consider if each diamond was independently, randomly selected from the total global population of diamonds. This information is not given, though we consider it safe to assume the carat size of one diamond does not affect the carat size of another. Likewise, the prices, other than economic equilibrium market price, should be independent of one another.

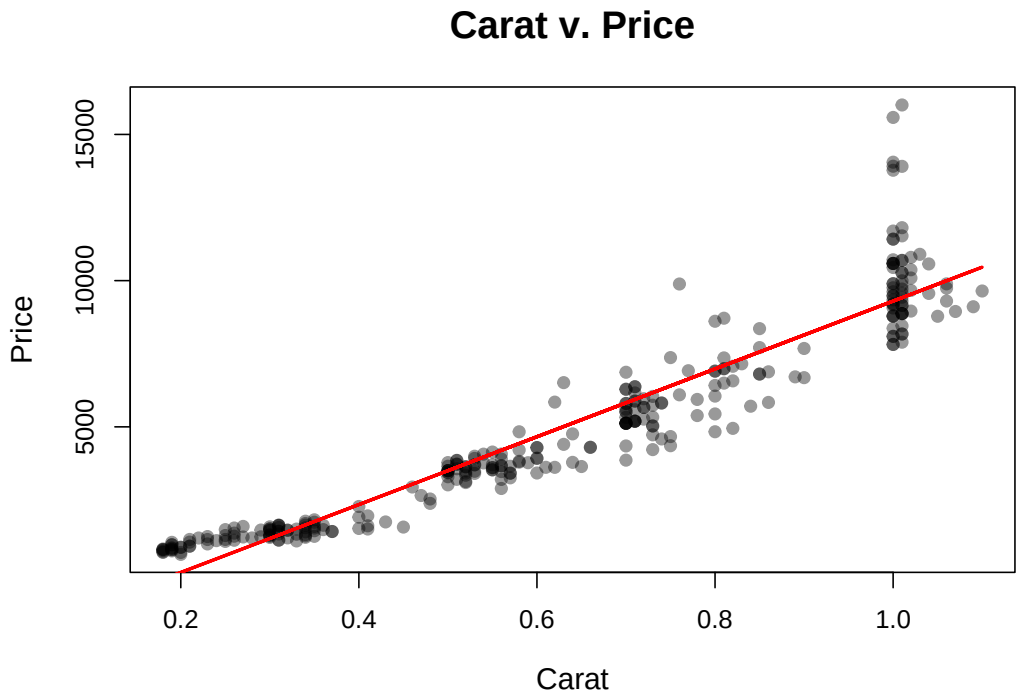


Figure 1: The relationship between carat size and the price of a diamond. There appears to be a positive non-linear relationship, and there appears to be heteroscedasticity, as the variance of the price appears to increase as the carat size appears to increase.

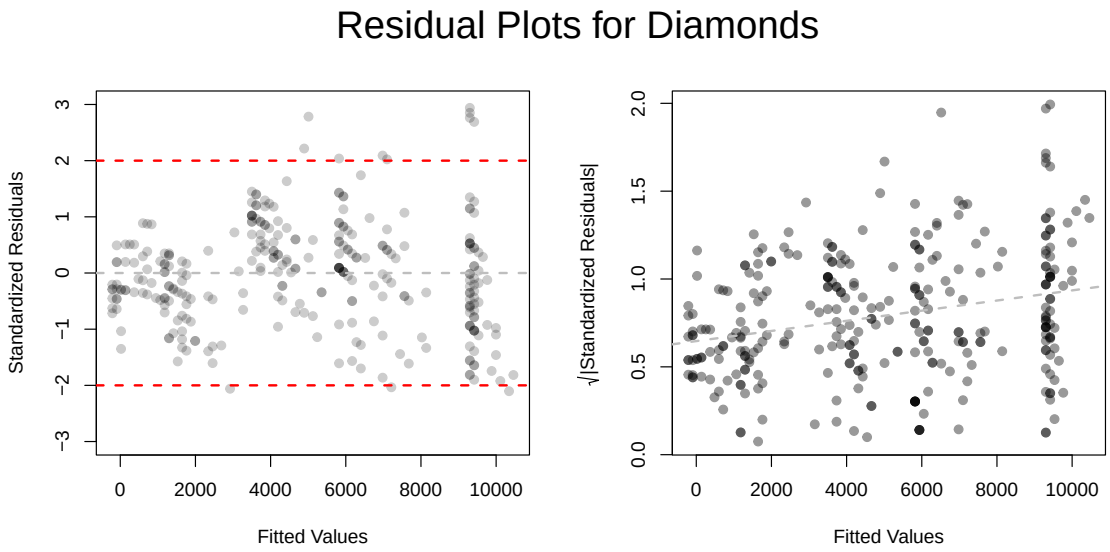


Figure 2: Both residual plots affirm the suspicion that there exists some heteroscedasticity in the price of diamonds.

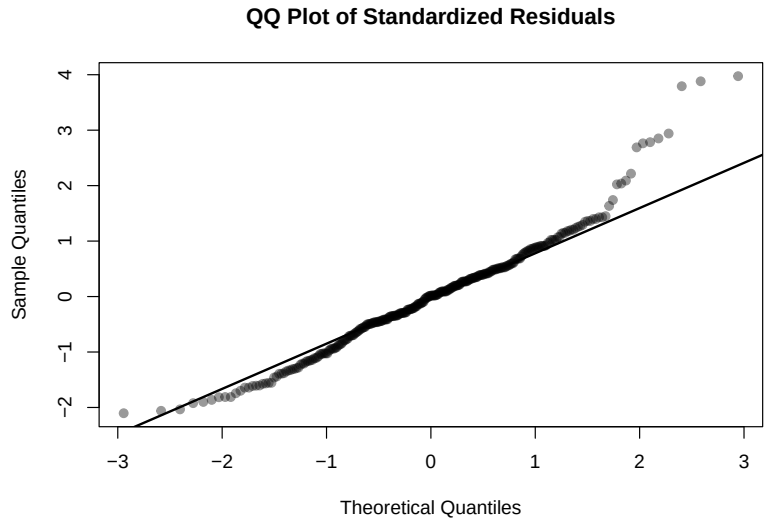


Figure 3: For the most part, this QQ-plot indicates the standardized residuals resembles a normal distribution. This can be more objectively checked with a Shapiro-Wilk test for normality.

```
lmtest::bptest(mod1)

##
##  studentized Breusch-Pagan test
##
## data:  mod1
## BP = 18.053, df = 1, p-value = 2.149e-05
```

In this case, we can conclusively assume that observations are independent, and that

the errors are normally distributed, as the number of observations is sufficiently large and the QQ-plot indicates the distribution of standardized residuals looks fairly normal. Likewise the Breusch-Pagan test for heteroscedasticity yielded a significant p-value, meaning we should reject the null hypothesis that the data are homoscedastic. The assumption of a linear relationship and constant variance of the response variable are not good assumptions to make based on these data.

(c) What power transformation for the response is indicated may help by the Box-Cox family? Does this transformation solve the problem(s) listed in part (b)?

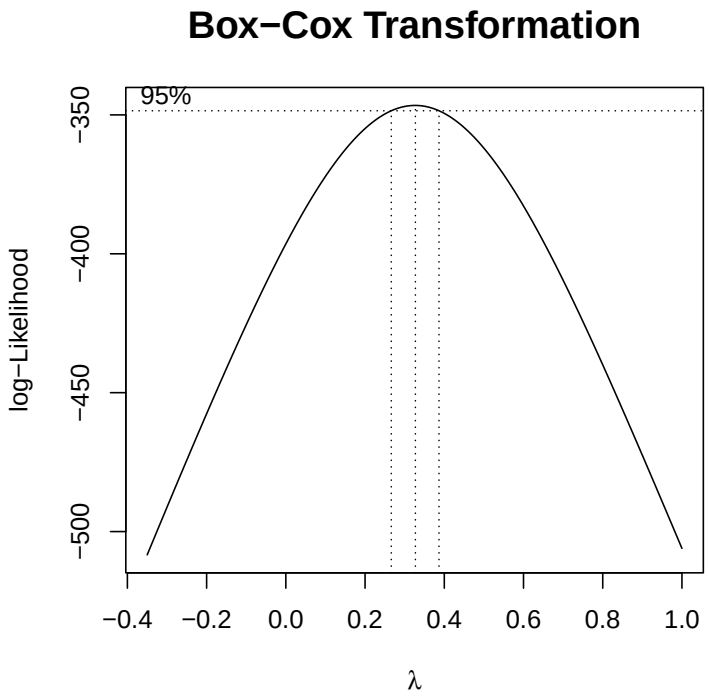


Figure 4: The Box-Cox transformations plot of the linear model of price using carats.

```
(lambda <- boxcox$x[which.max(boxcox$y)])

## [1] 0.3434343
```

The appropriate Box-Cox transformation is $\lambda \approx 0.3434$. This appears to resolve the apparent violations of the linear fit assumption; however, it does not appear to resolve the heteroscedasticity assumption.

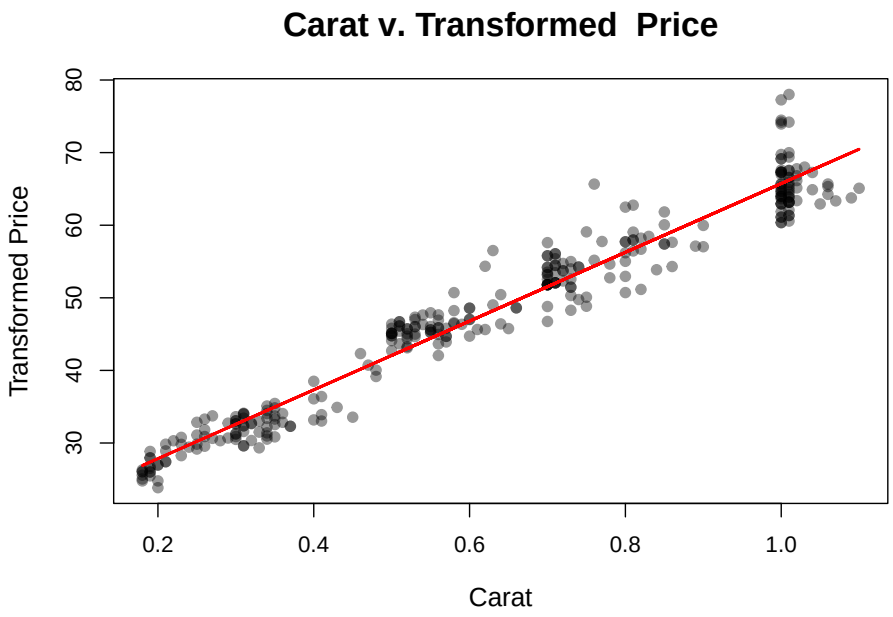


Figure 5: The relationship between transformed carat size and the price of a diamond. There appears to be a positive linear relationship, but there appears to be heteroscedasticity, as the variance of the price appears to increase as the carat size appears to increase.

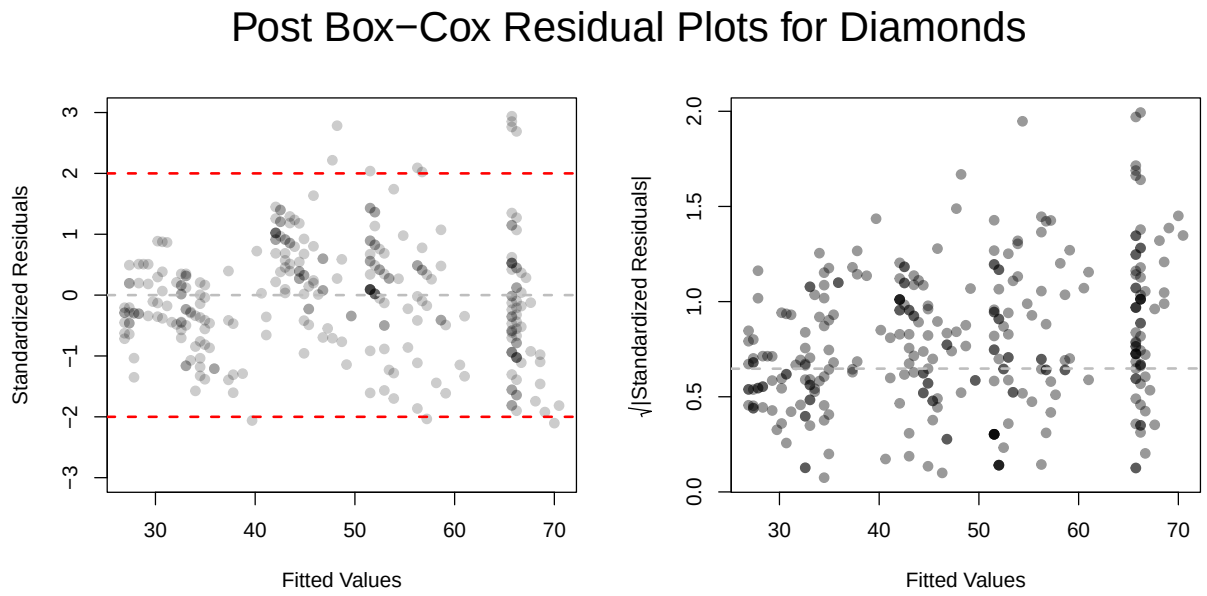


Figure 6: Although the new residual plots indicate there now exists a linear relationship between the carat and transformed price variables, both residual plots still affirm the suspicion that there exists some heteroscedasticity in the price of diamonds.

```
diamonds_bc <- transform(diamonds, "Price" = ((diamonds$Price)^(lambda)-1)/lambd
mod2 <- lm(Price ~ Carat, data = diamonds_bc)
lmtest::bptest(mod2)

##
##  studentized Breusch-Pagan test
##
## data:  mod2
## BP = 21.161, df = 1, p-value = 4.222e-06
```

Here we observe a low p-value for any reasonable choice for α, so for the Breusch-Pagan test, we reject the null hypothesis that the data is homoscedastic. Hence, the Box-Cox transformation did not eliminate the heteroscedasticity in the data.

(d) What set of transformations should be used simultaneously for both the explanatory and the response variables? Does this set of transformations solve the problem(s) listed in part (b)?

```
multi_bc <- summary(powerTransform(cbind(diamonds$Carat, diamonds$Price) ~ 1))
multi_bc[[3]]

##                                 LRT df       pval
## LR test, lambda = (0 0)  62.43443  2 2.7756e-14
## LR test, lambda = (1 1) 380.27761  2 < 2.22e-16
```

Both the transformations for $\boldsymbol{\lambda} = (0, 0)$ and $\boldsymbol{\lambda} = (1, 1)$ yield significant p-values; however, we observed that a transformation is needed based on the result in part (b). So we can consider the $\boldsymbol{\lambda} = (0, 0)$, or pairwise logarithmic transformation. For the sake of keeping the assignment concise after further investigation, a log-square root transformation for Y and X respectively seemed to fix problems in the model.

```
diamonds_log <- transform(diamonds,
                          "sqrt_Carat" = sqrt(Carat),
                          +"log_Price" = log(Price))
log_mod <- lm(log_Price ~ sqrt_Carat, data= diamonds_log)
standard_res_log <- rstandard(log_mod)

## Error:  <text>:3:40:  unexpected '='
## 2:                            "sqrt_Carat" = sqrt(Carat),
## 3:                            +"log_Price" =
##                                            ^
```

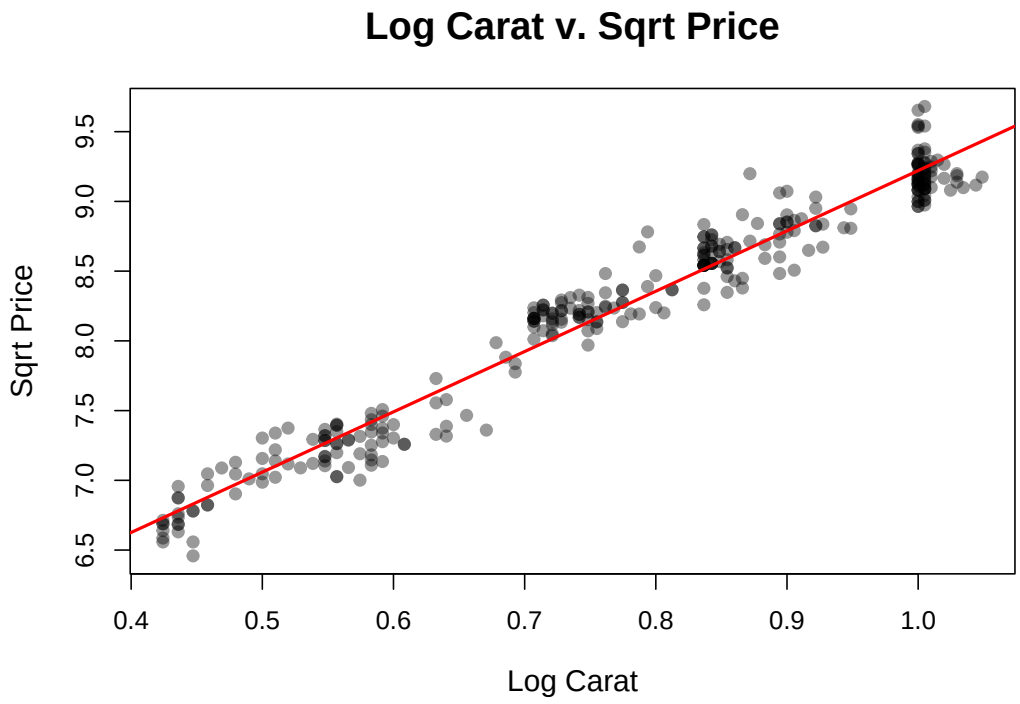


Figure 7: The relationship between log carat size and the log price of a diamond. There appears to be a positive more linear relationship than without the transformations, but there still appears to be a little heteroscedasticity, as the variance of the price appears to increase as the carat size appears to increase.

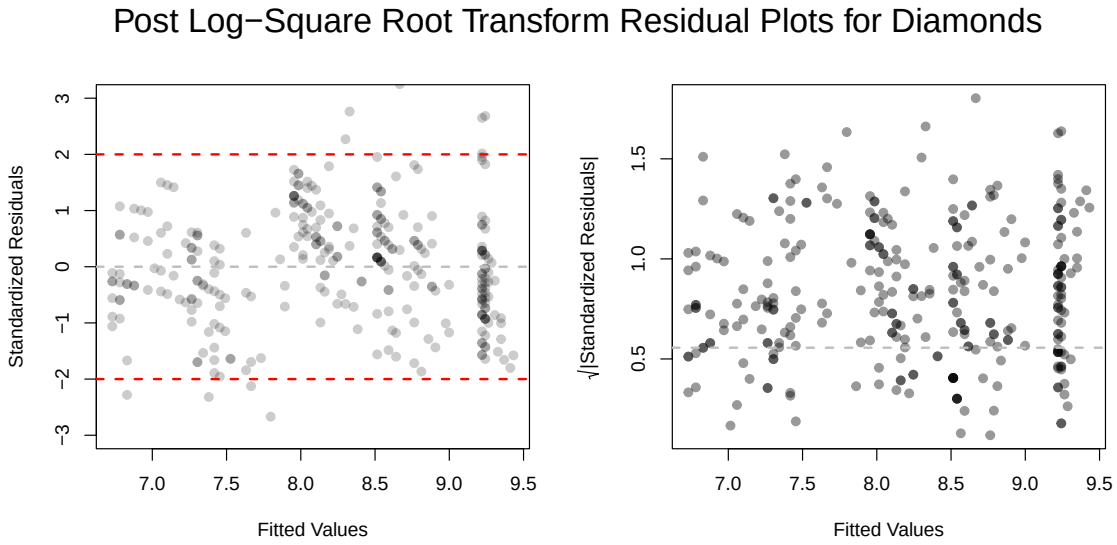


Figure 8: Both residual plots seem to indicate there is no more heteroscedasticity in the data.

```
lmtest::bptest(log_mod)

## Error in lmtest::bptest(log_mod):  object 'log_mod' not found
```

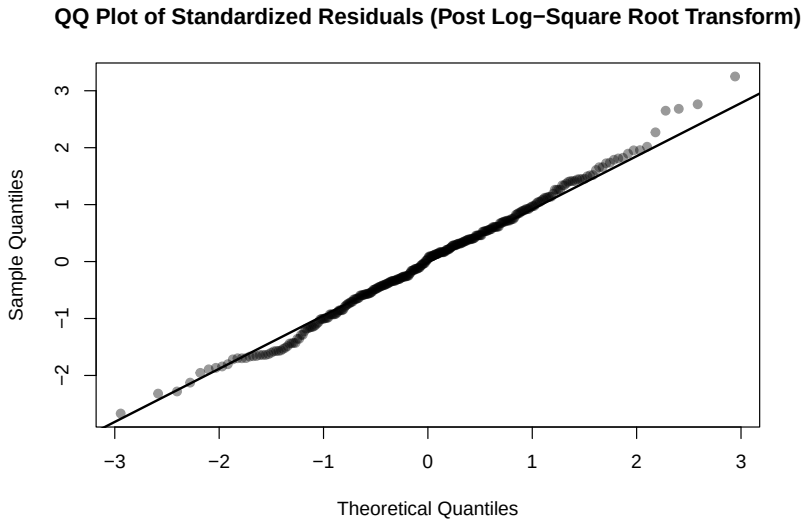
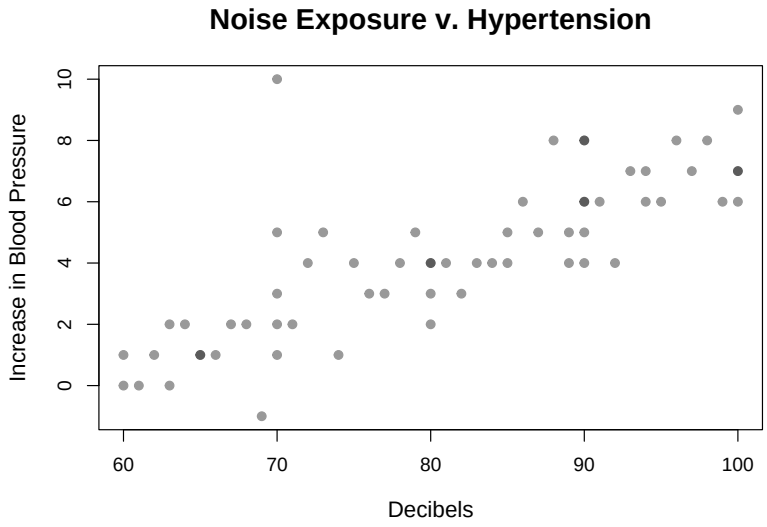


Figure 9: For the most part, this QQ-plot indicates the standardized residuals of the log-square root model resembles a normal distribution.

This Breusch-Pagan test for the log-log data transformation yields a much more ambiguous p-value that is not significant at the $\alpha = 0.1$ level, which would mean that by that decision rule, we would fail to reject the notion that the transformed predictor and response data is homoscedastic. This does not necessarily mean that the data is homoscedastic, but it is an objective tool we have to evaluate this.

7. An article in the *Journal of Sound and Vibration* described a study investigating the relationship between noise exposure and hypertension. The data in "sound.txt" contains two columns. The first is "db" or sound pressure level in decibels, and the second is "bp" or blood pressure rise in millimeters of mercury measured on a sample of 62 individuals.

   (a) Plot the rise in blood pressure versus the sound.



   (b) Obtain the leverage plot. Are there any points of high leverage?

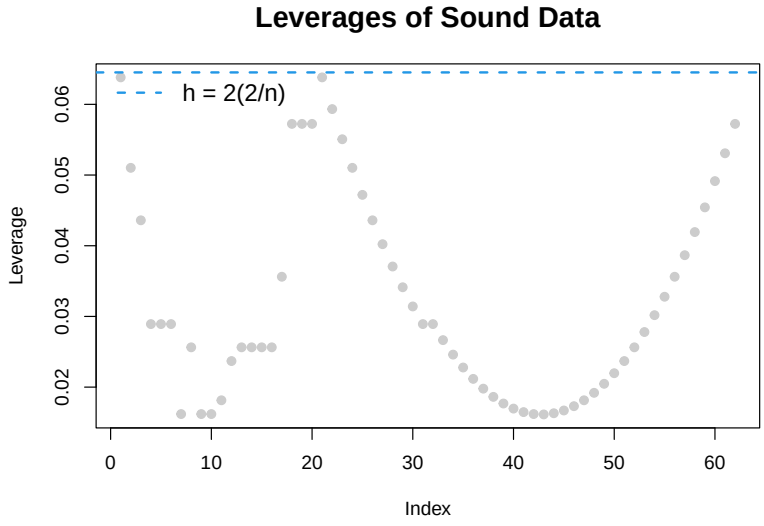


Figure 10: Leverage plot of the decibels of sound. There appears to be no points of high leverage.

(c) Obtain the values of Cook's distance. What value should these be compared to in order to determine if any are unusually large values? Are any of the values large? If so, which ones?

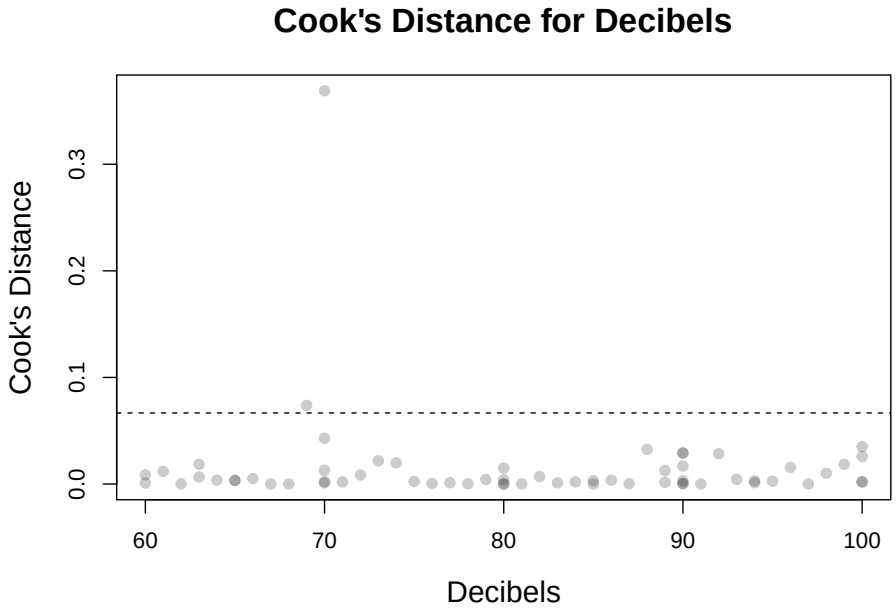


Figure 11: The Cook's distance measurement for the complete sound data with a horizontal line plotted at $\frac{4}{n-2}$.

```
cd <- cooks.distance(lm(bp ~ db, data = sound))
n <- nrow(sound)
cd |> order(decreasing = TRUE) |> head()

## [1] 32 30  5 18 50 14

cd[c(32, 30,  5, 18, 50, 14)]; 4/(n-2)

##         32         30          5         18         50         14
## 0.36891840 0.07372797 0.04290985 0.03500338 0.03250965 0.02909547
## [1] 0.06666667
```

A general rule of thumb is to flag observations whose Cook's distance is greater than $4/(n-2)$, where n is the number of observations in the data. Here, we see that observations 32 and 30 exceed this threshold, as seen in Figure 11.

(d) Fit the SLR both with and without any unusual observations you may have found. What are the differences in the two fits?

```
sound2 <- sound[-c(30, 32),]
lm(bp ~ db, data = sound) |> summary(); lm(bp ~ db, data = sound2) |> summary()

##
```

```
## Call:
## lm(formula = bp ~ db, data = sound)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2464 -0.7451 -0.2338  0.6028  7.5878
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.19589    1.32958  -6.916 3.47e-09 ***
## db           0.16583    0.01629  10.182 1.07e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.547 on 60 degrees of freedom
## Multiple R-squared:  0.6334,Adjusted R-squared:  0.6273
## F-statistic: 103.7 on 1 and 60 DF,  p-value: 1.074e-14
##
## Call:
## lm(formula = bp ~ db, data = sound2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0440 -0.6182 -0.1763  0.7237  2.7173
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.68507    0.99588  -9.725 8.63e-14 ***
## db           0.17097    0.01214  14.079  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 58 degrees of freedom
## Multiple R-squared:  0.7736,Adjusted R-squared:  0.7697
## F-statistic: 198.2 on 1 and 58 DF,  p-value: < 2.2e-16
```

Looking at the output for both models, aside from marginal differences between the coefficient estimates and decreases in residual standard error, the only great change was the jump in R^2 the model recieved when fitting it without the unusual observations (which rose by 0.1402).

(e) Fit a robust linear regression model with both of the Huber and Tukey's bi-square weights to all of the observations. Overlay both lines on the scatterplot with the least squares line and the line with any outliers removed; give their estimated regression lines; and comment on the comparison of the fits among these four lines.

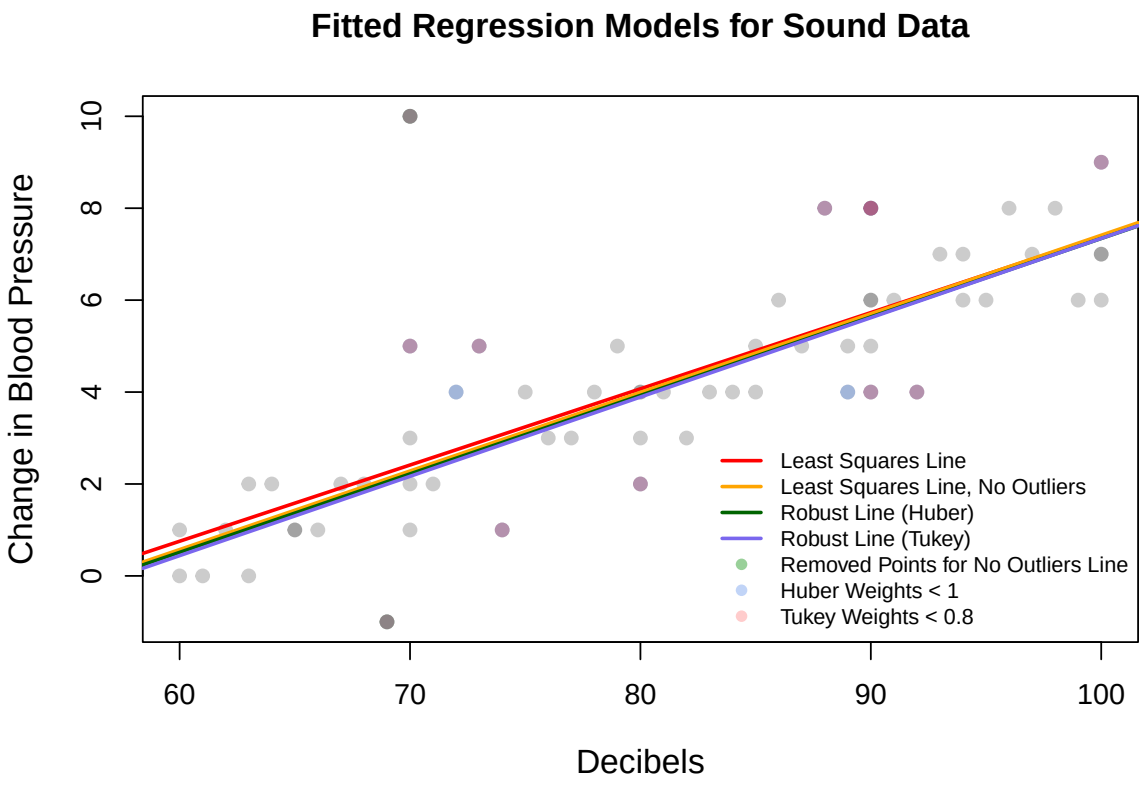


Figure 12: Four models of the relationship between noise exposure and hypertension.

All of these models essentially appear the same. The only marginal difference is that the intercept of the robust lines are marginally lower, probably due to the lower weight allocated to the outlier above the lines at 70 decibels.

(f) Compare the Huber and Tukey weights of each observation. Which tend to be larger– the Huber or Tukey bi-square weights?

```
summary(weights$sound_huber_w); summary(weights$sound_tukey_w)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1800  1.0000  1.0000  0.9198  1.0000  1.0000
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.8350  0.9550  0.8666  0.9900  1.0000
```

In this case, the Tukey weights tend to be smaller than the Huber weights, as the Tukey weights even throw out at least one observation entirely (see summaries).

8. In this exercise, you will generate data to show (1) how residual plots should look when the assumptions are satisfied and (2) how QQ-normal plots look when the residuals are not normally distributed.

   (a) Generate 5 uncorrelated variables with the following code to ensure that everyone has the same data:

   ```
   set.seed(1128)
   n <- 30
   data <- matrix(rnorm(5*n), byrow = F, ncol =5)
   ```

   Find all 10 pairwise correlations, and plot every combination, i.e., `pairs(data)`. Does every plot look as if the points are randomly scattered? Should they all look randomly scattered?

```r
set.seed(1128)
n <- 30
data <- matrix(rnorm(5*n), byrow = F, ncol =5)
```

```r
pairs(data)
```

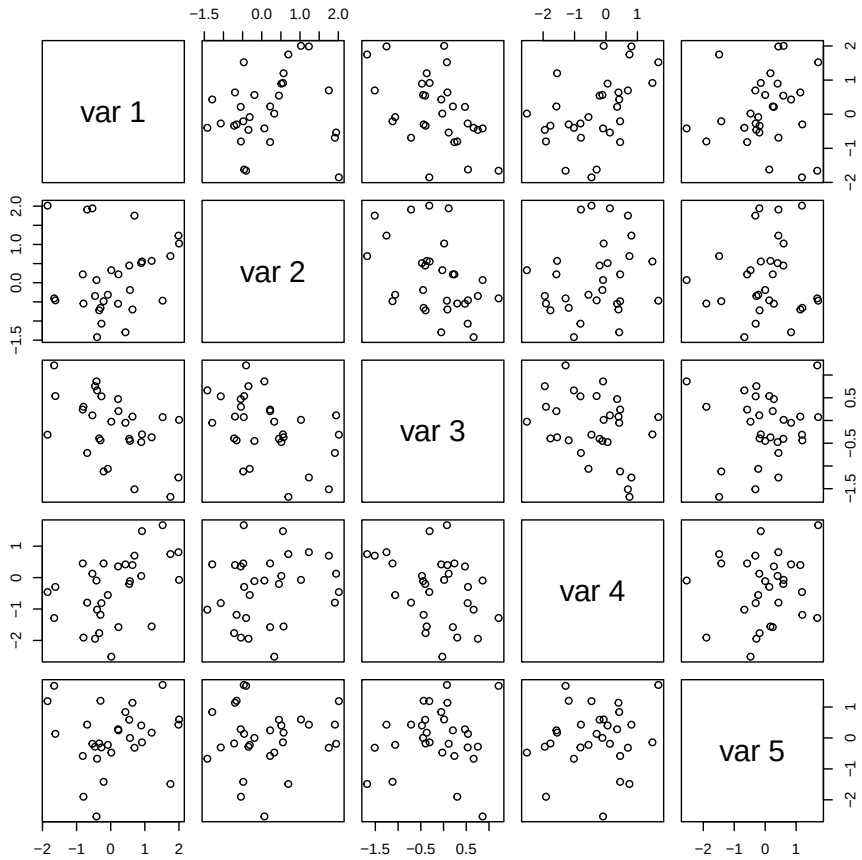


Figure 13: Pairs plot for the uncorrelated data.

```
data |> as.data.frame() |> corrr::correlate() |> knitr::kable()
```

Term	V1	V2	V3	V4	V5
V1	·	0.1316506	-0.4911861	0.4545818	0.0658970
V2	0.1316506	·	-0.4429776	0.1910418	0.0505549
V3	-0.4911861	-0.4429776	·	-0.3482814	0.0749569
V4	0.4545818	0.1910418	-0.3482814	·	0.1154437
V5	0.0658970	0.0505549	0.0749569	0.1154437	·

All of the pairwise plots look uncorrelated, and they should all look uncorrelated, as they are generated from a standard normal random number generator. None of the pairwise sample correlations are noticeably large.

(b) Now, generate data from the following model $Y_i = 3.2 - 0.5x_i + \epsilon_i$ where x_i can be any number between 100 and 200. First, let $\epsilon_i \sim \text{Unif}(0, 1)$, and generate data for $n = 50$, 100, 300, and 500. Plot the QQ-normal plot for the standardized residuals at each sample size. Describe what you see. Repeat for $\epsilon_i \sim \text{Exp}(1)$.

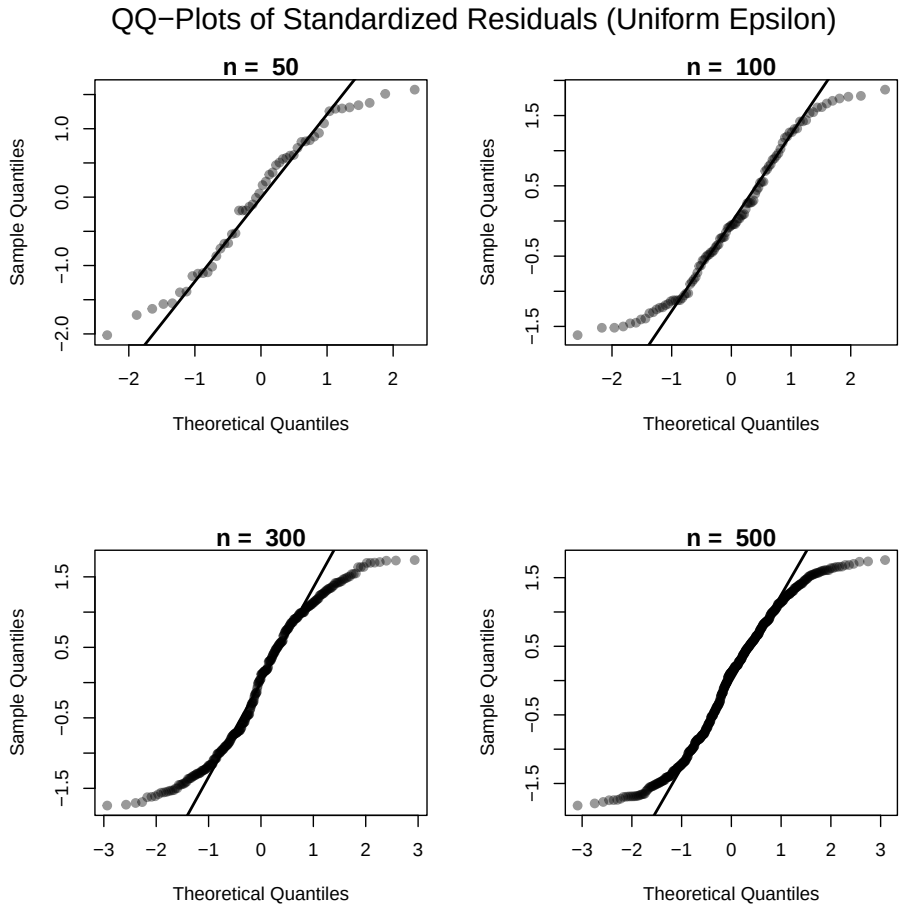


Figure 14: QQ-plots at respective sample sizes for the standardized residuals at $Y_i = 3.2 - 0.5x_i + \epsilon_i$, where $\epsilon_i \sim \text{Unif}(0, 1)$.

In Figure 14, the QQ-plots indicate that the standardized residuals mostly resemble a standard normal distribution with lighter tails.

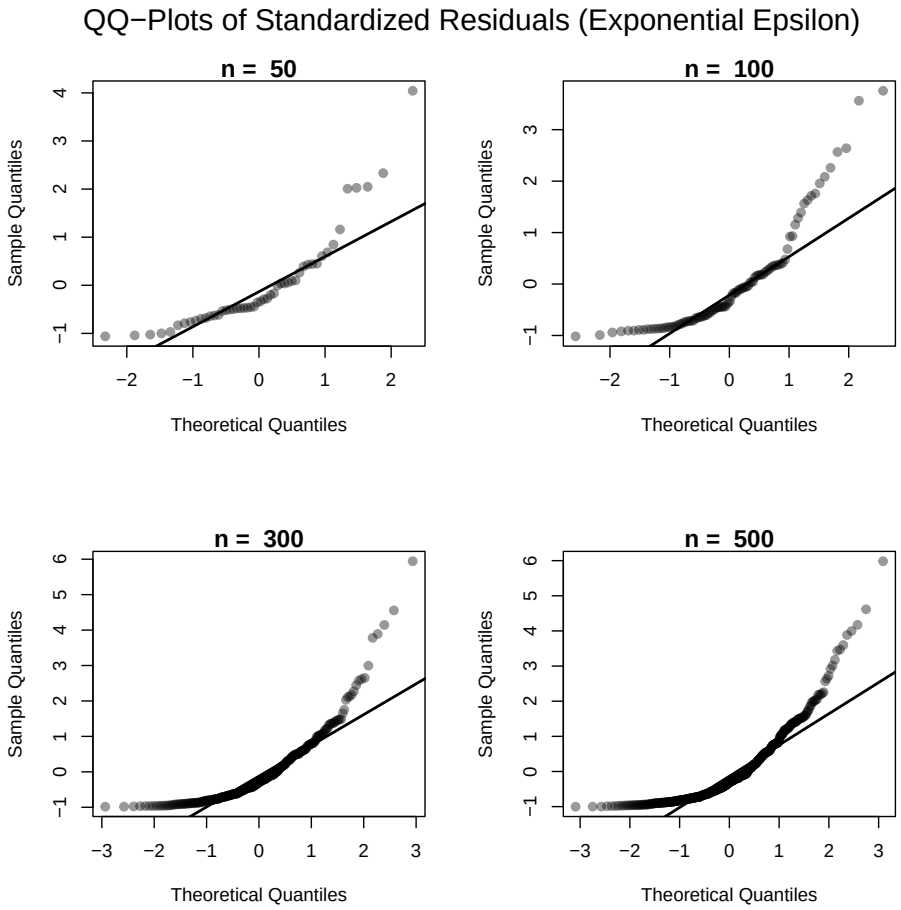


Figure 15: QQ-plots at respective sample sizes for the standardized residuals at $Y_i = 3.2 - 0.5x_i + \epsilon_i$, where $\epsilon_i \sim \text{Exp}(1)$.

In Figure 15, the QQ-plots indicate that the standardized residuals appear to be very right-skewed, and do not resemble a standard normal distribution.

9. A study recorded the age of truck tractors (in years) and the cost (in dollars) of maintaining them over a six month period. The data are given in "tractor.txt."

```
tractor <- read.delim("tractor.txt",
                      sep ="", header = TRUE, dec =".")
mod3 <- lm(Cost ~ Age, data = tractor)
mod3_standard_res <- rstandard(mod3)
```

   (a) Plot cost versus age and fit a SLR regression of cost on age. Give the estimated regression line.

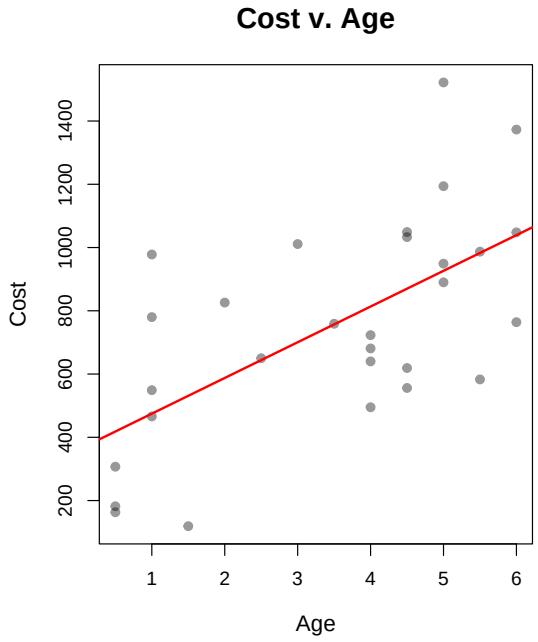


Figure 16: Scatter plot showing relationship between the age of a truck tractors and the cost of maintenance every six month period.

(b) Assess the assumptions, and check for outliers.

We assume that each truck tractor was randomly selected, yielding independent observations. There are no apparent reasons each observation would be dependent on each other.

Both of the residual plots exhibits linear relationships amongst the residuals, and constant spread over the fitted values of the model. Hence, this relationship appears to be correctly specified as a linear model. Additionally, since the variance appears to remain constant amongst the residuals, we can assume there exists no heteroscedasticity. We can check with the Breusch-Pagan test.

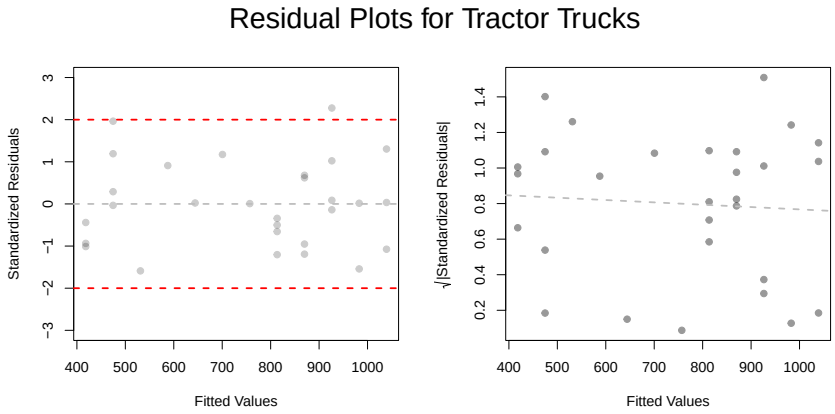


Figure 17: Residual plots for the standardized residuals of the fitted model of tractor truck age on cost for repair every six months.

```
lmtest::bptest(mod3)

##
##  studentized Breusch-Pagan test
##
## data:  mod3
## BP = 0.00012973, df = 1, p-value = 0.9909
```

From our Breusch-Pagan test, we obtain a very high p-value of 0.9909, which would not be significant for any reasonable α value. Hence, we fail to reject the null hypothesis which states the data is homoscedastic.

```
shapiro.test(mod3_standard_res)

##
##  Shapiro-Wilk normality test
##
## data:  mod3_standard_res
## W = 0.96453, p-value = 0.4226
```

From the Shapiro-Wilks test for normality, we fail to reject the null hypothesis that our standardized model residuals are normally distributed. Likewise, the QQ-plot

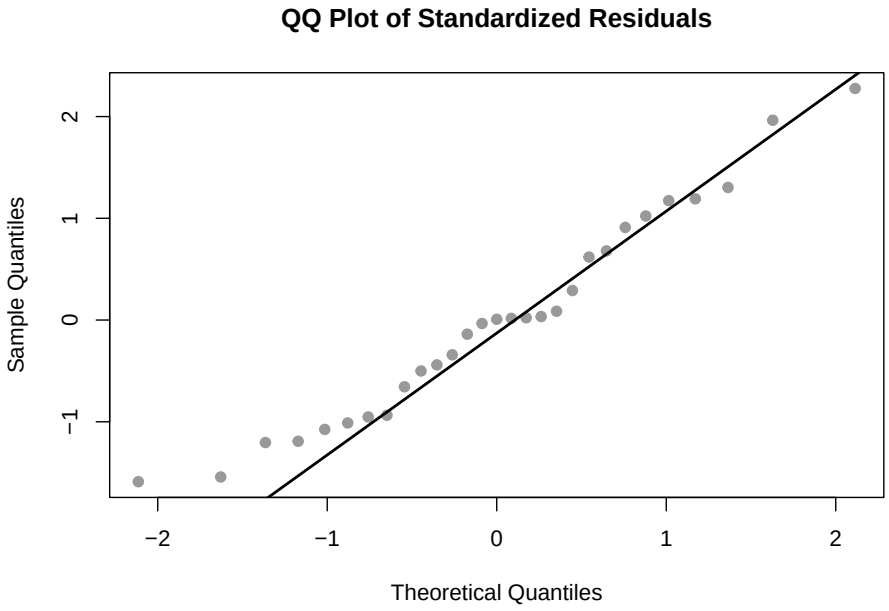


Figure 18: QQ-plot for the standardized residuals of the fitted model of tractor truck age on cost for repair every six months.

of the standardized residuals indicates they not to severely deviate from a normal distribution. Hence, we conclude the observations are independent, the relationship can be correctly specified using simple linear regression, the data are homoscedastic, and the residuals of the model are normally distributed.

(c) Is there a significant linear relationship between cost and age? Justify your answer.

```
mod3 |> summary()

##
## Call:
## lm(formula = Cost ~ Age, data = tractor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -412.17 -236.26    2.02  179.11  595.66
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   361.80     106.28   3.404 0.002087 **
## Age           112.91      26.92   4.194 0.000264 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 269.5 on 27 degrees of freedom
## Multiple R-squared:  0.3945,Adjusted R-squared:  0.3721
## F-statistic: 17.59 on 1 and 27 DF,  p-value: 0.0002642
```

```r
# p-value for age coefficient
summary(mod3)$coefficients[2,4]

## [1] 0.0002642406
```

For all reasonable values of α, the p-value for the coefficient on age is significant, which indicates there exists a significant relationship between cost and age

(d) Give 95% confidence intervals for the slope and intercept.

```r
n <- nrow(tractor)
b0hat <- mod3$coef[1]
Sxx   <- sum((tractor$Age - mean(tractor$Age))^2)
sighat_sq <-  270^2

b0CI_lower <- b0hat - qt(0.975, n - 2)*sqrt(sighat_sq*(1 / n + mean(tractor$Age)
b0CI_upper <- b0hat + qt(0.975, n - 2)*sqrt(sighat_sq*(1 / n + mean(tractor$Age)

# CI for intercept
cat("CI for intercept: (", b0CI_lower, ",", b0CI_upper, ")")

## CI for intercept: ( 143.3545 , 580.2544 )

b1hat <- mod3$coef[2]

b1CI_lower <- b1hat - qt(0.975, n - 2)*sqrt(sighat_sq/Sxx)
b1CI_upper <- b1hat + qt(0.975, n - 2)*sqrt(sighat_sq/Sxx)
# CI for slope
cat("CI for slope: (", b1CI_lower, ",", b1CI_upper, ")")

## CI for slope: ( 57.57495 , 168.2403 )
```

The 95% CI for the intercept is (143.35, 580.25); the 95% CI for the intercept is (57.57, 168.24).

(e) Give a 99% confidence interval for the mean cost of maintaining tractors that are 2.5 years old. Interpret this interval in the context of the problem.

```r
# CI
sighat <- sqrt(sighat_sq)
ciME <- qt(1 - (0.01 / 2), n - 2)*sighat*sqrt(1/n + (2.5 - mean(tractor$Age))^2/

estMU <- b0hat + b1hat*2.5
estMU

## (Intercept)
##    644.0735

cat("99% CI for Cost: (", round(estMU - ciME, 3), ",", round(estMU + ciME, 3), "

## 99% CI for Cost: ( 486.944 , 801.203 )
```

We are 99% confident that the mean 6-month cost of maintaining tractor trucks that are 2.5 years old is between \$486.94 and \$801.203.

(f) Give a 99% prediction interval for the cost of maintaining a particular tractor that is 2.5 years old. Interpret this interval in the context of the problem.

```
# PI
piME <- qt(1 - 0.01 / 2, n - 2)*sighat*sqrt(1 + 1/n + (2.5 - mean(tractor$Age))^
cat("99% PI for Cost: (", round(estMU - piME, 3), ",", round(estMU + piME, 3), "

## 99% PI for Cost: ( -120.335 , 1408.482 )
```

The 99% PI for the 6-month cost of maintaining a 2.5 year-old tractor truck is (0, 1408.482). Hence, we would expect with 99% confidence that the $(n+1)^{\text{th}}$ 2.5 year old tractor trucks costs between \$0 and \$1408.48 to maintain for six months.

(g) Overlay 95% CI and PI bounds for the entire range of tractor ages on the scatterplot.

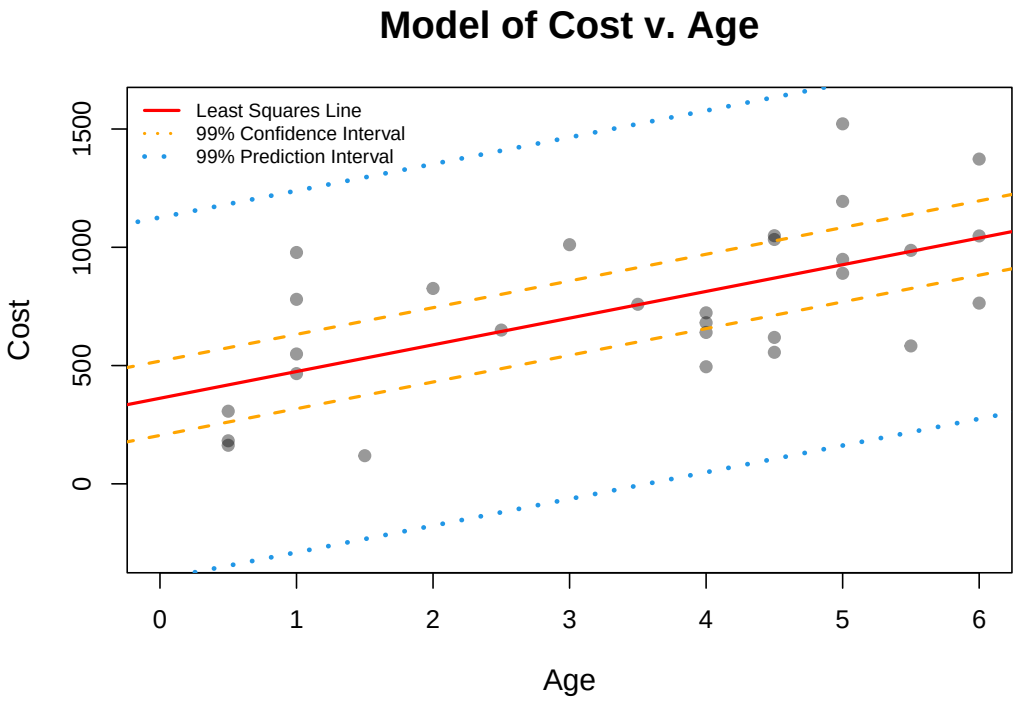


Figure 19: Tractor truck cost per age, and linear model with confidence and prediction intervals.

Sheather Problem, Chapter 3 # 7

We have that $E[Y] = \mu$, and that $\text{Var}[Y] = \mu^2$. Consider a transformation of Y, $f(Y) = Z$ such that it makes the variance of $f(Y) \approx 1$, and consider its Taylor Series expansion around $E[Y]$.

$$\begin{aligned}
f(Y) &= \underbrace{f(E[Y])}_{\text{A Constant}} + f'(E[Y])(Y - E[Y]) \\
&= \underbrace{f(E[Y])}_{\text{A Constant}} - \underbrace{f'(E[Y])E[Y]}_{\text{A Constant}} + \underbrace{f'(E[Y])Y}_{\text{A R.V.}} \\
\implies 1 \approx \text{Var}[f(Y)] &= \text{Var}\left[\underbrace{f(E[Y])}_{\text{A Constant}} - \underbrace{f'(E[Y])E[Y]}_{\text{A Constant}} + \underbrace{f'(E[Y])Y}_{\text{A R.V.}}\right] \\
&\approx f'(E[Y])^2 \,\text{Var}\,[Y] \\
\implies f'(E[Y])^2 &= \frac{1}{\mu^2} \quad (\text{Since } \text{Var}[Y] = \mu^2) \\
\implies f'(\mu) &= \frac{1}{\mu}. \quad (\text{Since } E[Y] = \mu)
\end{aligned}$$

We can find $f(\cdot)$ by integrating both sides with respect to μ.

$$\begin{aligned}
\int f'(\mu)d\mu &= \int \frac{1}{\mu} d\mu \\
&= \log(\mu) + C \\
&= f(\mu).
\end{aligned}$$

So then the transformation $f(Y) = \log(Y) + C$, where C is a constant of integration. Hence the appropriate transformation of Y for stabilizing variance is the log transformation.

Code Appendix

The following is a copy of the entire R script used to do all analyses in this assignment. It is carefully commented to label which code corresponds to each problem.

```
# 3 ----------------------------------------------------------------------

SSReg <- 13599
SST <- 39957

(rsq <-  (SSReg/(SSReg + SST)) |> round(digits = 4))

# 6 ----------------------------------------------------------------------

diamonds <- read.delim("diamonds.txt",
                       sep ="", header = TRUE, dec =".")


# linearly related shown by residuals
mod1 <- lm(Price ~ Carat, data = diamonds)
standard_res <- rstandard(mod1)
mod1_pred <- predict(mod1, diamonds)

# 6b ---------------------------------------------------------------------
# linearly related
pdf(file = "Figures/diamonds_scatter.pdf", height = 5, width = 7)
plot(diamonds$Carat, diamonds$Price,
     pch = 19, col = alpha("black", 0.4),
     xlab = "Carat", ylab = "Price",
     main = "Carat v. Price",
     cex.lab = 1.2,
     cex.main = 1.5)
lines(diamonds$Carat, mod1_pred, lwd = 2, col = "red")
dev.off()

# residual plot
pdf(file = "Figures/diamond_res.pdf", width = 10, height = 5)
par(mfrow = c(1,2))
plot(mod1$fitted, standard_res,
     xlab = "Fitted Values",
     ylab = "Standardized Residuals",
     pch = 19,
     col = alpha("black", 0.2),
     ylim = c(-3, 3))
abline(h = 0, lty = 2, col = "gray", lwd = 2)
abline(h = -2, lty = 2, col = "red", lwd = 2)
abline(h = 2, lty = 2, col = "red", lwd = 2)
```

```r
# alternative residual plot
plot(mod1$fitted, sqrt(abs(standard_res)),
     xlab = "Fitted Values",
     ylab = latex2exp::TeX(r"(\sqrt{|Standardized Residuals|})"),
     pch = 19,
     col = alpha("black", 0.4))
abline(lm(sqrt(abs(standard_res)) ~ mod1$fitted), lwd = 2, col = "gray", lty = 2)
mtext("Residual Plots for Diamonds", side = 3, line = - 2, outer = TRUE, cex = 2)
dev.off()

# qq-plot of standardized residuals
pdf(file = "Figures/diamond_qq.pdf", width = 7, height = 5)
qqnorm(standard_res, pch = 19, main = "", col = alpha("black", 0.4))
qqline(standard_res, lwd = 2)
title("QQ Plot of Standardized Residuals")
dev.off()

shapiro.test(standard_res)

# 6c ----------------------------------------------------------------------
library("MASS")

pdf(file = "Figures/diamond_boxcox.pdf", width = 5, height = 5)
bc <- boxcox(mod1, lambda=seq(-0.35, 1, len = 1000))
title("Box-Cox Transformation", cex.main = 1.5)
dev.off()

boxcox <- boxcox(mod1)
lambda <- boxcox$x[which.max(boxcox$y)]

diamonds_bc <- transform(diamonds, "Price" = ((diamonds$Price)^(lambda)-1)/lambda)

mod2 <- lm(Price ~ Carat, data = diamonds_bc)
standard_res_bc <- rstandard(mod2)
mod2_pred <- predict(mod2, diamonds_bc)

# post Box-Cox transform scatter plot
pdf(file = "Figures/diamonds_scatter_bc.pdf", height = 5, width = 7)
plot(diamonds_bc$Carat, diamonds_bc$Price,
     pch = 19, col = alpha("black", 0.4),
     xlab = "Carat", ylab = "Transformed Price",
     main = "Carat v. Transformed  Price",
     cex.lab = 1.2,
     cex.main = 1.5)
lines(diamonds_bc$Carat, mod2_pred, lwd = 2, col = "red")
dev.off()
```

```r
# residual plot post Box-Cox transform
pdf(file = "Figures/diamond_res_bc.pdf", width = 10, height = 5)
par(mfrow = c(1,2))
plot(mod2$fitted, standard_res_bc,
     xlab = "Fitted Values",
     ylab = "Standardized Residuals",
     pch = 19,
     col = alpha("black", 0.2),
     ylim = c(-3, 3))
abline(h = 0, lty = 2, col = "gray", lwd = 2)
abline(h = -2, lty = 2, col = "red", lwd = 2)
abline(h = 2, lty = 2, col = "red", lwd = 2)

# alternative residual plot post Box-Cox transform
plot(mod2$fitted, sqrt(abs(standard_res_bc)),
     xlab = "Fitted Values",
     ylab = latex2exp::TeX(r"(\sqrt{|Standardized Residuals|})"),
     pch = 19,
     col = alpha("black", 0.4))
abline(lm(sqrt(abs(standard_res)) ~ mod1$fitted), lwd = 2, col = "gray", lty = 2)
mtext("Post Box-Cox Residual Plots for Diamonds", side = 3, line = - 2, outer = TRUE, cex
dev.off()

# qq-plot of standardized residuals post Box-Cox transform
pdf(file = "Figures/diamond_qq_bc.pdf", width = 7, height = 5)
qqnorm(standard_res_bc, pch = 19, main = "", col = alpha("black", 0.4))
qqline(standard_res_bc, lwd = 2)
title("QQ Plot of Standardized Residuals (Post BC)")
dev.off()

# 6d ----------------------------------------------------------------------

library("car")
#Gives the multivariate normal transformation
multi_bc <- summary(powerTransform(cbind(diamonds$Carat, diamonds$Price) ~ 1))
multi_bc[[3]]


#Gives the univariate transformation for Y
summary(powerTransform(diamonds$Carat ~ diamonds$Price))

diamonds_log <- transform(diamonds, "sqrt_Carat" = sqrt(Carat), "log_Price" = log(Price))

log_mod <- lm(log_Price ~ sqrt_Carat, data= diamonds_log)
standard_res_log <- rstandard(log_mod)
```

```r
# post Box-Cox transform scatter plot
pdf(file = "Figures/diamonds_scatter_log.pdf", height = 5, width = 7)
plot(diamonds_log$sqrt_Carat, diamonds_log$log_Price,
     pch = 19, col = alpha("black", 0.4),
     xlab = "Log Carat", ylab = "Sqrt Price",
     main = "Log Carat v. Sqrt Price",
     cex.lab = 1.2,
     cex.main = 1.5)
abline(log_mod, lwd = 2, col = "red")
dev.off()

# residual plot post Box-Cox transform
pdf(file = "Figures/diamond_res_log.pdf", width = 10, height = 5)
par(mfrow = c(1,2))
plot(log_mod$fitted, standard_res_log,
     xlab = "Fitted Values",
     ylab = "Standardized Residuals",
     pch = 19,
     col = alpha("black", 0.2),
     ylim = c(-3, 3))
abline(h = 0, lty = 2, col = "gray", lwd = 2)
abline(h = -2, lty = 2, col = "red", lwd = 2)
abline(h = 2, lty = 2, col = "red", lwd = 2)

# alternative residual plot post Box-Cox transform
plot(log_mod$fitted, sqrt(abs(standard_res_log)),
     xlab = "Fitted Values",
     ylab = latex2exp::TeX(r"(\sqrt{|Standardized Residuals|})"),
     pch = 19,
     col = alpha("black", 0.4))
abline(lm(sqrt(abs(standard_res)) ~ mod1$fitted), lwd = 2, col = "gray", lty = 2)
mtext("Post Log-Square Root Transform Residual Plots for Diamonds", side = 3, line = - 2,
dev.off()

# qq-plot of standardized residuals post Box-Cox transform
pdf(file = "Figures/diamond_qq_log.pdf", width = 7, height = 5)
qqnorm(standard_res_log, pch = 19, main = "", col = alpha("black", 0.4))
qqline(standard_res_log, lwd = 2)
title("QQ Plot of Standardized Residuals (Post Log-Square Root Transform)")
dev.off()

# 7 ----------------------------------------------------------------------

sound <- read.delim("sound.txt",
                    sep ="", header = TRUE, dec =".")

#  - scatter plot
```

```r
pdf(file = "Figures/sound_scatter.pdf", height = 5, width = 7)
plot(sound$db, sound$bp,
     pch = 19, col = alpha("black", 0.4),
     xlab = "Decibels", ylab = "Increase in Blood Pressure",
     main = "Noise Exposure v. Hypertension",
     cex.lab = 1.2,
     cex.main = 1.5)
dev.off()


n <- nrow(sound)
hi <- hat(sound$db, intercept = TRUE)

# b - leverage plot
pdf(file = "Figures/lev_sound.pdf", width = 7, height = 5)
plot(hi, pch = 19, ylab = "Leverage", col = alpha("black", 0.20))
abline(h = 2*2/n, lty = 2, lwd = 2, col = 4)
# identify(hi) #4 27 47 79
legend("topleft", c("h = 2(2/n)"), lty = c(2, 2), lwd = c(2, 2), col = c(4), bty = "n", c
title("Leverages of Sound Data", cex.main=1.5)
dev.off()

# c - Cook's distance
cd <- cooks.distance(lm(bp ~ db, data = sound))
n <- nrow(sound)
cd |> order(decreasing = TRUE) |> head()
cd[c(32, 30,  5, 18, 50, 14)]
4/(n-2)

pdf(file = "Figures/cd_sound.pdf", width = 7, height = 5)
plot(sound$db, cd, xlab = "Decibels", ylab = "Cook's Distance",
     pch = 19, col = alpha("black", 0.2),
     main = "Cook's Distance for Decibels",
     cex.main = 1.5,
     cex.lab = 1.3)
abline(h = 4 / (n - 2), lty = 2, col = "red")
dev.off()

# d - linear modeling
sound2 <- sound[-c(30, 32),]
lm(bp ~ db, data = sound) |> summary(); lm(bp ~ db, data = sound2) |> summary()

sound_ols <- lm(bp ~ db, data = sound)
sound_ols_no <- lm(bp ~ db, data = sound2)
sound_huber <- rlm(bp ~ db, data = sound, psi = psi.huber)
sound_bisq  <- rlm(bp ~ db, data = sound, psi = psi.bisquare)
```

```r
weights <- data.frame(x.db = sound$db,
                      y.bp = sound$bp,
                      # sound_ols_resid = round(sound_ols$residuals, 2),
                      # sound_ols_no_resid = round(sound_ols_no$residuals, 2),
                      sound_huber_w = round(sound_huber$w, 2),
                      sound_tukey_w = round(sound_bisq$w, 2))
huber_downweights <- which(weights$sound_huber_w <  1)
tukey_downweights <- which(weights$sound_tukey_w <  0.8)

pdf(file = "Figures/sound_mods.pdf", width = 7, height = 5)
plot(sound$db, sound$bp, pch = 19,
     col = alpha("black", 0.2),
     xlim = c(60, 100),
     ylim = c(-1, 10),
     xlab = "Decibels",
     ylab = "Change in Blood Pressure",
     cex.lab = 1.2)
abline(sound_ols, lwd = 2, col = "red")
abline(sound_ols_no, lwd = 2, col = "orange")
abline(sound_huber, lwd = 2, col = "darkgreen")
abline(sound_bisq, lwd = 2, col = "slateblue2")
points(sound[c(30, 32),], col = alpha("green4", 0.4), pch = 19)
points(weights[huber_downweights, c(1,2)], col = alpha("cornflowerblue", 0.4), pch = 19)
points(weights[tukey_downweights, c(1,2)], col = alpha("red", 0.2), pch = 19)
title(main = "Fitted Regression Models for Sound Data", cex = 1.5)
legend("bottomright", c("Least Squares Line",
                        "Least Squares Line, No Outliers",
                        "Robust Line (Huber)",
                        "Robust Line (Tukey)",
                        "Removed Points for No Outliers Line",
                        "Huber Weights < 1",
                        "Tukey Weights < 0.8"),
       col = c("red", "orange", "darkgreen", "slateblue2",
               alpha("green4", 0.4), alpha("cornflowerblue", 0.4),
               alpha("red", 0.2)),
       lwd = c(2, 2, 2, 2, NA, NA, NA),
       pch = c(NA, NA, NA, NA, 19, 19, 19),
       bty = "n",
       cex = 0.75)
dev.off()

mean(weights$sound_huber_w); mean(weights$sound_tukey_w)
summary(weights$sound_huber_w); summary(weights$sound_tukey_w)

# 8 ------------------------------------------------------------------
```

```r
# a
set.seed(1128)
n <- 30
data <- matrix(rnorm(5*n), byrow = F, ncol =5)

pdf(file = "Figures/data_pairs.pdf", width = 7, height = 7)
pairs(data)
dev.off()

data |> as.data.frame() |> corrr::correlate() |> knitr::kable(format = "latex")

# b

generate_data_unif <- function(n = 50) {
  set.seed(1128)
  x <- runif(n, min = 100, max = 200)
  epsilon <- runif(n)
  y <- 3.2 - 0.5*x + epsilon
  cbind(y, x, epsilon) |> as.data.frame()
}

generate_data_exp <- function(n = 50) {
  set.seed(1128)
  x <- runif(n, min = 100, max = 200)
  epsilon <- rexp(n, rate = 1)
  y <- 3.2 - 0.5*x + epsilon
  cbind(y, x, epsilon) |> as.data.frame()
}

make_qq <- function(model) {
  standard_res <- rstandard(model)
  qqnorm(standard_res, pch = 19, main = "", col = alpha("black", 0.4))
  qqline(standard_res, lwd = 2)
  title(paste("n = ", n), cex.main = 1.3, line = 0.2)
}

n_vec <- c(50, 100, 300, 500)

pdf(file = "Figures/unif_rng_qq.pdf", width = 7, height = 7)
par(mfrow = c(2,2))
for (n in c(50, 100, 300, 500)) {
  data <- generate_data_unif(n = n)
  mod <- lm(y ~ x, data = data)
  make_qq(model = mod)
}
mtext("QQ-Plots of Standardized Residuals (Uniform Epsilon)", side = 3, line = -2, outer
dev.off()
```

```r
pdf(file = "Figures/exp_rng_qq.pdf", width = 7, height = 7)
par(mfrow = c(2,2))
for (n in c(50, 100, 300, 500)) {
  data <- generate_data_exp(n = n)
  mod <- lm(y ~ x, data = data)
  make_qq(model = mod)
}
mtext("QQ-Plots of Standardized Residuals (Exponential Epsilon)", side = 3, line = -2, ou
dev.off()


# 9 ----------------------------------------------------------------------

tractor <- read.delim("tractor.txt",
                      sep ="", header = TRUE, dec =".")

# a
pdf(file = "Figures/tractor_scatter.pdf", width = 5, height = 6)
plot(tractor$Age, tractor$Cost,
     pch = 19, col = alpha("black", 0.4),
     xlab = "Age", ylab = "Cost",
     main = "Cost v. Age",
     cex.lab = 1.2,
     cex.main = 1.5)
abline(lm(Cost ~ Age, data = tractor), lwd = 2, col = "red")
dev.off()

# 9b

mod3 <- lm(Cost ~ Age, data = tractor)
mod3_standard_res <- rstandard(mod3)

# residual plot
pdf(file = "Figures/tractor_res.pdf", width = 10, height = 5)
par(mfrow = c(1,2))
plot(mod3$fitted, mod3_standard_res,
     xlab = "Fitted Values",
     ylab = "Standardized Residuals",
     pch = 19,
     col = alpha("black", 0.2),
     ylim = c(-3, 3))
abline(h = 0, lty = 2, col = "gray", lwd = 2)
abline(h = -2, lty = 2, col = "red", lwd = 2)
abline(h = 2, lty = 2, col = "red", lwd = 2)

# alternative residual plot
```

```r
plot(mod3$fitted, sqrt(abs(mod3_standard_res)),
     xlab = "Fitted Values",
     ylab = latex2exp::TeX(r"(\sqrt{|Standardized Residuals|})"),
     pch = 19,
     col = alpha("black", 0.4))
abline(lm(sqrt(abs(mod3_standard_res)) ~ mod3$fitted), lwd = 2, col = "gray", lty = 2)
mtext("Residual Plots for Tractor Trucks", side = 3, line = - 2, outer = TRUE, cex = 2)
dev.off()

  lmtest::bptest(mod3)

# qq-plot of standardized residuals
pdf(file = "Figures/tractor_qq.pdf", width = 7, height = 5)
qqnorm(mod3_standard_res, pch = 19, main = "", col = alpha("black", 0.4))
qqline(mod3_standard_res, lwd = 2)
title("QQ Plot of Standardized Residuals")
dev.off()

shapiro.test(mod3_standard_res)

# 9c

summary(mod3)$coefficients[2,4]

# 9d

n <- nrow(tractor)
b0hat <- mod3$coef[1]
Sxx   <- sum((tractor$Age - mean(tractor$Age))^2)
sighat_sq <-  270^2

b0CI_lower <- b0hat - qt(0.975, n - 2)*sqrt(sighat_sq*(1 / n + mean(tractor$Age)^2/Sxx))
b0CI_upper <- b0hat + qt(0.975, n - 2)*sqrt(sighat_sq*(1 / n + mean(tractor$Age)^2/Sxx))

cat("(", b0CI_lower, ",", b0CI_upper, ")")

b1hat <- mod3$coef[2]

b1CI_lower <- b1hat - qt(0.975, n - 2)*sqrt(sighat_sq/Sxx)
b1CI_upper <- b1hat + qt(0.975, n - 2)*sqrt(sighat_sq/Sxx)

cat("(", b1CI_lower, ",", b1CI_upper, ")")

# 9e

sighat <- sqrt(sighat_sq)
```

```r
ciME <- qt(1 - 0.01 / 2, n - 2)*sighat*sqrt(1/n + (2.5 - mean(tractor$Age))^2/Sxx)

# CI
estMU <- b0hat + b1hat*2.5
cat("99% CI for Cost: (", round(estMU - ciME, 3), ",", round(estMU + ciME, 3), ")")

# 9f

# PI
piME <- qt(1 - 0.01 / 2, n - 2)*sighat*sqrt(1 + 1/n + (2.5 - mean(tractor$Age))^2/Sxx)
cat("99% PI for Cost: (", round(estMU - piME, 3), ",", round(estMU + piME, 3), ")")

# 9g

xstar <- seq(min(tractor$Age), max(tractor$Age), len = 500)

upCI <- mod3$coef[1] + mod3$coef[2]*xstar + ciME
loCI <- mod3$coef[1] + mod3$coef[2]*xstar - ciME

upPI <- mod3$coef[1] + mod3$coef[2]*xstar + piME
loPI <- mod3$coef[1] + mod3$coef[2]*xstar - piME

pdf(file = "Figures/tractor_scatter_ints.pdf", width = 7, height = 5)
plot(tractor$Age, tractor$Cost,
     pch = 19, col = alpha("black", 0.4),
     xlab = "Age", ylab = "Cost",
     main = "Model of Cost v. Age",
     cex.lab = 1.2,
     cex.main = 1.5,
     ylim = c(-100,1500),
     xlim = c(0,6))
abline(lm(Cost ~ Age, data = tractor), lwd = 2, col = "red")
lines(xstar, upCI, lwd = 2, lty =2, col = "orange")
lines(xstar, loCI, lwd = 2, lty =2, col = "orange")
lines(xstar, upPI, lwd = 3, lty =3, col = 4)
lines(xstar, loPI, lwd = 3, lty =3, col = 4)
legend("topleft", c("Least Squares Line",
                    "99% Confidence Interval",
                    "99% Prediction Interval"),
       col = c("red", "orange", 4),
       lwd = c(2, 2, 3),
       lty = c(1, 3, 3),
       bty = "n",
       cex = 0.75)
dev.off()
```