Homework-3-Q1

BIOS507 Spring 2025 | Dr Lukemire | Elizabeth Nemeti Due: March 3 2025

Problem 1.

The data were logged in catalyst_dataset.csv and contain the following columns:

- Day: A column logging which day of the study the new technician performed the analysis.
- Intermediate Concentration: A column logging the concentration of the intermediate generated during the morning phase of the analysis.
- Yield A: column logging yield of the novel compound obtained by the technician in the afternoon using the intermediate species obtained in the morning.

predictor variable (x) -> Intermediate Concentration response variable (y) -> Yield

```r
data_path = "/Users/elizabethnemeti/Documents/GitHub/BIOS507-Coursework/Homeworks/Homework-3/"
data_file <- file.path(data_path, "catalyst_dataset.csv")
catalyst_data <- read.table("catalyst_dataset.csv", header = TRUE, sep = "")

head(catalyst_data)
```

```
##      Day.Intermediate_Concentration.Yield
## 1  1,-2.302378232761063,-5.10406563699301
## 2 2,-0.1508874474163997,6.568837091565295
## 3   3,9.29354157074562,3.5330812153762636
## 4    4,2.35254195712288,4.524574006022668
## 5 5,3.146438675804731,0.48381432734984386
## 6 6,11.575324934416406,11.549722751910798
```

```r
str(catalyst_data) # data is in 1 var atm
```

```
## 'data.frame':    100 obs. of  1 variable:
##  $ Day.Intermediate_Concentration.Yield: chr  "1,-2.302378232761063,-5.10406563699301" "2,-
0.1508874474163997,6.568837091565295" "3,9.29354157074562,3.5330812153762636" "4,2.35254195712288,4.5245
```

```r
catalyst_data <- read.csv(text = catalyst_data$Day.Intermediate_Concentration.Yield,
                          header = FALSE,
                          col.names = c("Day", "Intermediate_Concentration", "Yield"))

str(catalyst_data) # now it's in 3 vars
```

```
## 'data.frame':    100 obs. of  3 variables:
##  $ Day                       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Intermediate_Concentration: num  -2.302 -0.151 9.294 2.353 3.146 ...
##  $ Yield                     : num  -5.104 6.569 3.533 4.525 0.484 ...
```

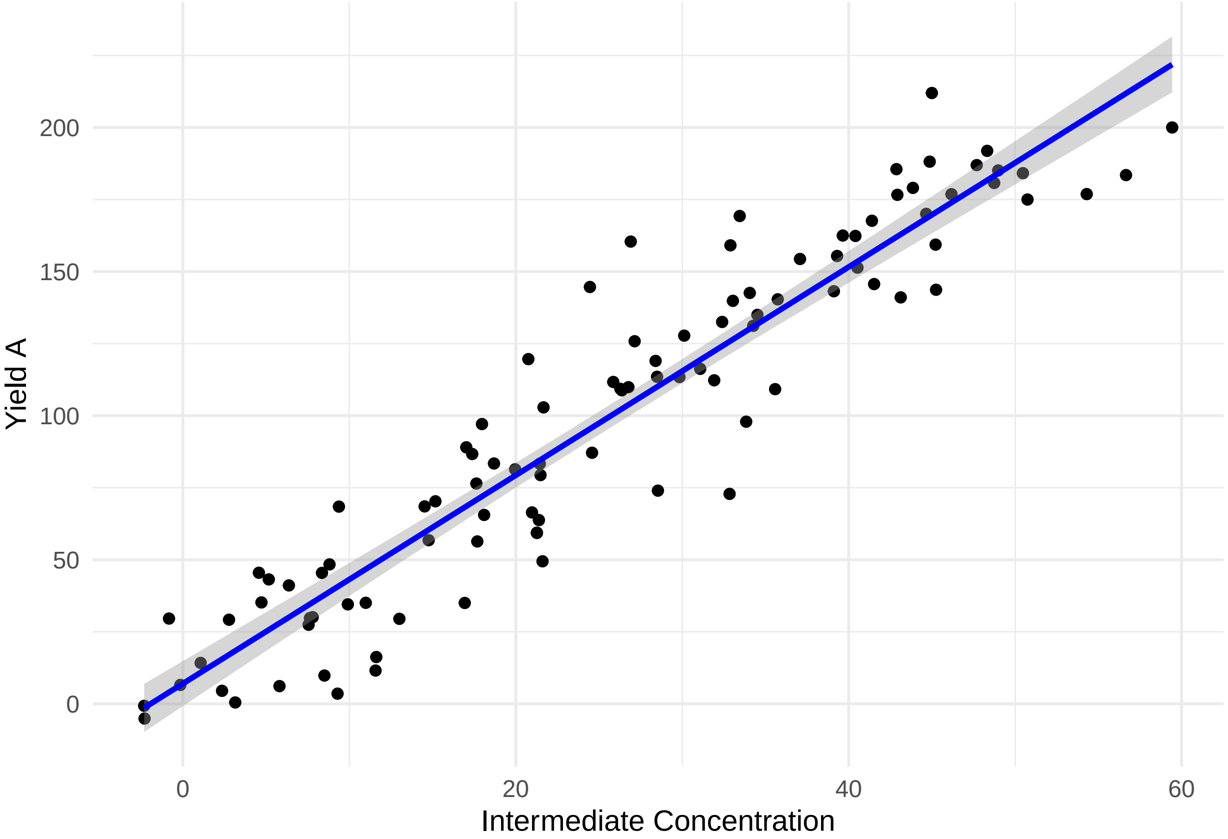
1. Ignore any potential “problems” in the data and answer the researcher team’s questions using a simple linear regression model. Draw conclusions about the association between concentration of the intermediate species and total yield of the organic compound. Be sure to carry out all of the usual steps in the analysis.

```r
summary(catalyst_data)
```

```
##       Day         Intermediate_Concentration     Yield
##  Min.   :  1.00   Min.   :-2.325             Min.   : -5.104
##  1st Qu.: 25.75   1st Qu.:12.660             1st Qu.: 47.656
##  Median : 50.50   Median :26.071             Median :105.842
##  Mean   : 50.50   Mean   :25.702             Mean   : 99.925
##  3rd Qu.: 75.25   3rd Qu.:39.160             3rd Qu.:152.120
##  Max.   :100.00   Max.   :59.437             Max.   :211.972
```

```r
catalyst_data_scatterplot <- ggplot(
  catalyst_data,
  aes(x = Intermediate_Concentration, y = Yield)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  theme_minimal() +
  xlab("Intermediate Concentration") +
  ylab("Yield A")
catalyst_data_scatterplot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Does a linear relationship appear appropriate? -> Yes

Assumed regression model for Y -> Yield = ß0 + ß1(Intermediate Concentration) +   (epsilon)

Assumptions about   (epsilon): - average erorr should be 0 - errors should folow a normal dist. (bell curve) - there should be no pattern in the errors, we should see homoscedasticity

```r
catalyst_model <- lm(Yield ~ Intermediate_Concentration, data = catalyst_data)
summary(catalyst_model)
```

```
##
## Call:
## lm(formula = Yield ~ Intermediate_Concentration, data = catalyst_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.890 -11.527   1.894  11.373  56.130
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  7.0189     3.9383   1.782   0.0778 .
## Intermediate_Concentration   3.6147     0.1313  27.535   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.31 on 98 degrees of freedom
## Multiple R-squared:  0.8855, Adjusted R-squared:  0.8844
## F-statistic: 758.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

ß0 (intercept) = 7.0189 (aka Y when Intermediate concentration is 0) ß1 (slope) = 3.6147 (aka for each unit increase in Intermediate_Concentration, Yield increases by 3.6147)

Estimated model -> E[Yield] = (7.0189) + (3.6147) x (Intermediate_Concentration)

```r
anova(catalyst_model)
```

```
## Analysis of Variance Table
##
## Response: Yield
##                            Df Sum Sq Mean Sq F value    Pr(>F)
## Intermediate_Concentration  1 312801  312801  758.19 < 2.2e-16 ***
## Residuals                  98  40431     413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSreg = 312801 SS total = 353232

```r
# manual check
# R2 = SSreg/SStotal
R2 = 312801/353232
R2
```

```
## [1] 0.8855398
```

```r
# R check
summary(catalyst_model)$r.squared
```

```
## [1] 0.88554
```

R squared being 0.8855398 means 88.55% of the variation in our Yield variable can be explained by the Intermediate Concentration and it's therefore a good predictor, remaining % would have to be explained by other factors beyond Intermediate Concentration.

H0: 1 = 0 -> Intermediate Concentration does not affect Yield HA: 1!= 0 -> Intermediate Concentration does affect Yield   = 0.05

Two tailed test, as we're testing if the slope is different from 0 in either direction, not just one. Therefore, (alpha)/2 = 0.025, and the test statistic is the t-statistic.

df residual for t test = 98

```
# t-value via manual t-test
# t = ß1hat/standard error of ß1hat
t = 3.6147/0.1313
t
```

```
## [1] 27.53008
```

```
# t-value via R table
summary(catalyst_model)$coefficients
```

```
##                             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                 7.018895  3.9382516  1.782236 7.780768e-02
## Intermediate_Concentration 3.614720  0.1312757 27.535331 6.383880e-48
```

```
# = 27.53008 matches our t value from the summary table (small diff probably due to internal rounding)

# critical t-value
qt(0.025, df = 98, lower.tail = TRUE)  # get -ve val
```

```
## [1] -1.984467
```

```
qt(0.025, df = 98, lower.tail = FALSE)  # get +ve val
```

```
## [1] 1.984467
```

```
# = ±1.984
```

To decide on whether to reject the null hypothesis, we need to check if |t| > critical value.

|27.53008| > 1.984467, therefore we reject the null hypothesis.

Now, to get our 95% CI:

```
# manual check for CI

# ß1hat +- critical value * SE_ß1hat

ß1_hat <- 3.614720
SE_ß1_hat <- 0.1312757
t_crit_val <- 1.984467

upper_bound = ß1_hat + t_crit_val * SE_ß1_hat
upper_bound
```

```
## [1] 3.875232
```

```
lower_bound = ß1_hat - t_crit_val * SE_ß1_hat
lower_bound
```

```
## [1] 3.354208
```

```
# R check for CI
confint(catalyst_model)
```

```
##                                 2.5 %    97.5 %
## (Intercept)                -0.7964372 14.834227
## Intermediate_Concentration  3.3542075  3.875232
```

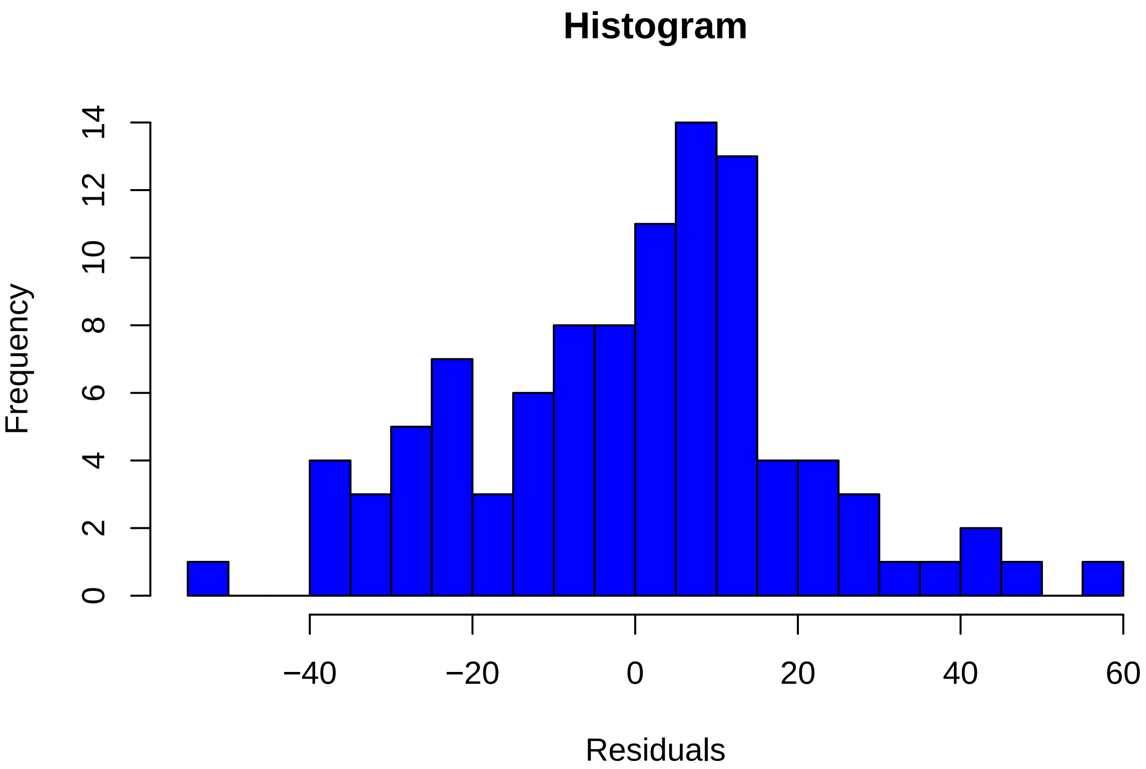
To answer the researchers' question: "The team hypothesized that higher concentrations of the intermediate species during the reaction would correlate with increased yields of the final product."

In the context of the study, since the p-value (2.2e-16) is much smaller than = 0.05, we have strong evidence to reject the H0, and that the Intermediate Concentration influences Yield. The R squared value being 0.886, further supports that Intermediate Concentration influences Yield. We are 95% confident that the interval 3.35-3.872 units contains the true value for ß1, where for each additional unit increase in Intermediate Concentration, Yield increases by 3.6147 units.

2. Perform diagnostics. Is anything wrong? If so, what might be the reason for this?

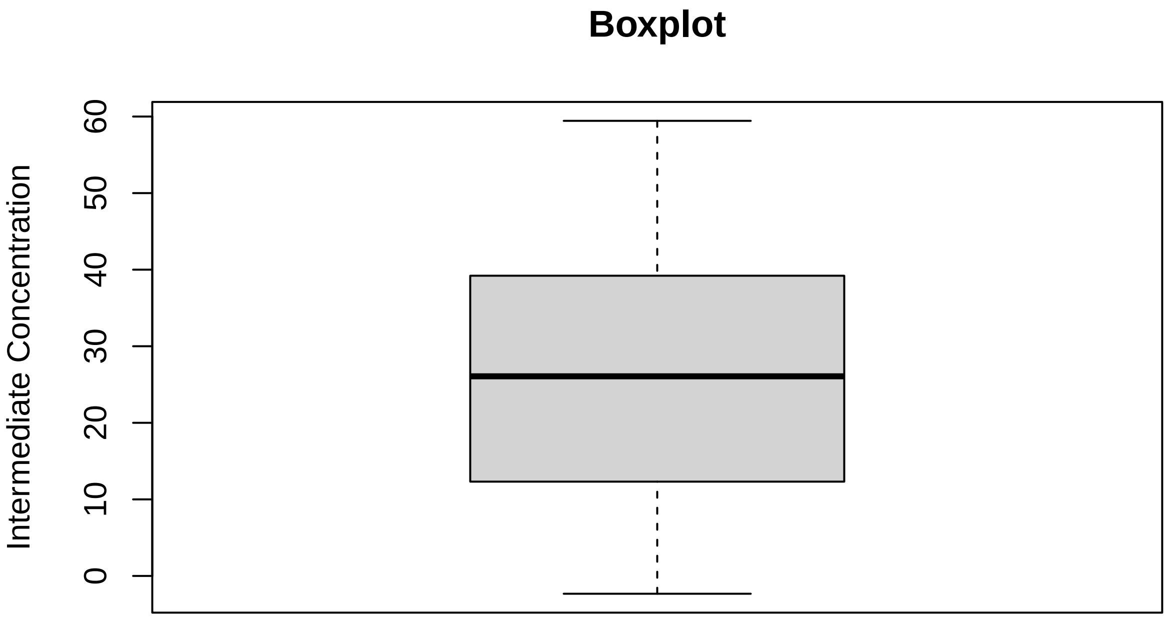
First, we can use a histogram to check if the data is normally distributed. It appears that the data follows a bell curve and is mostly centered around 0, indicating normal distribution. The only concern is somehwat of a right skew, to be checked with the QQ plot.

```r
hist(residuals(catalyst_model), breaks = 30, col = "blue",
     main = "Histogram", xlab = "Residuals")
```



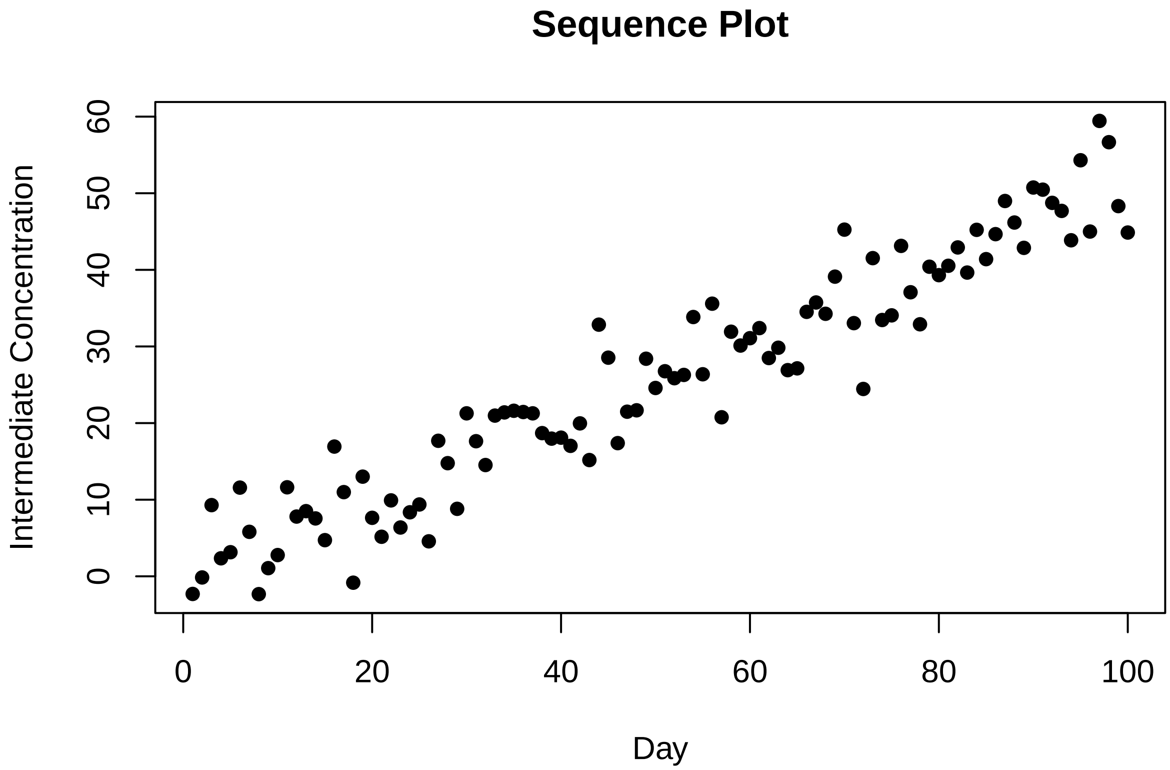
Checking for extreme values with a boxplot. No outliers.

```r
boxplot(catalyst_data$Intermediate_Concentration,
        main = "Boxplot",
        ylab = "Intermediate Concentration")
```



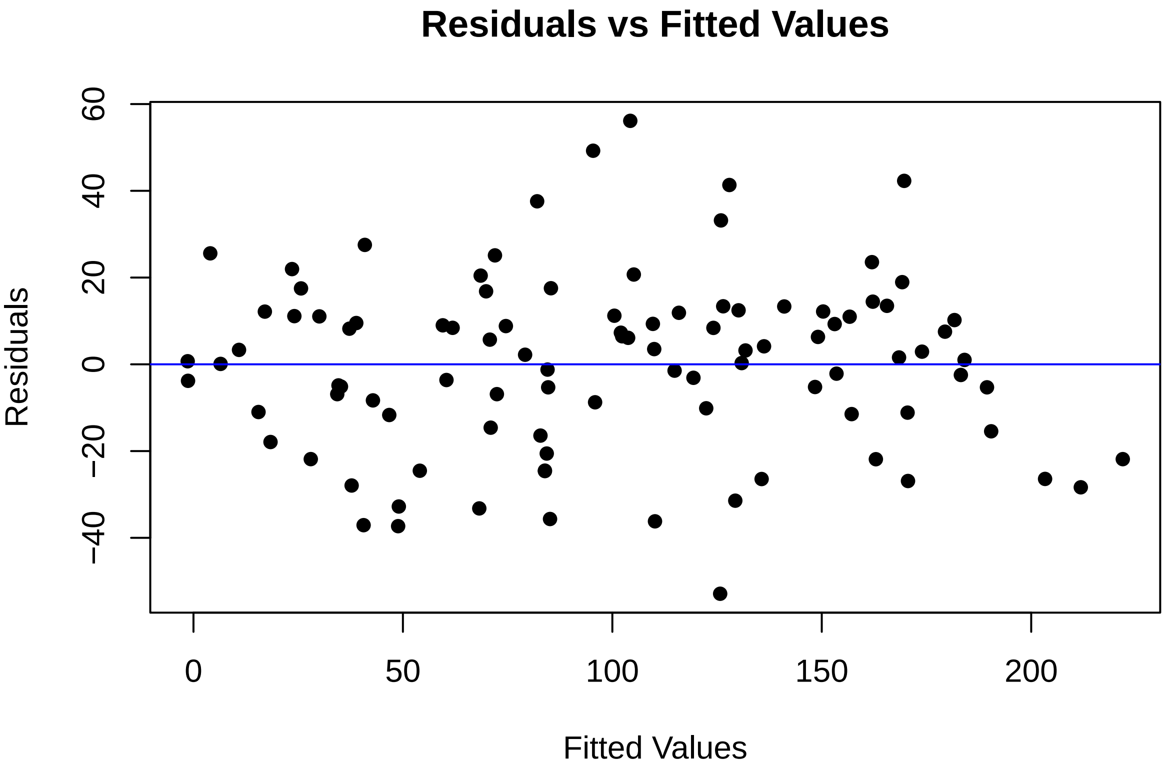
Next, we check the sequence plot to see if there's inconsistencies in the experiment being done over the course of the 100 days. There is a clear upwards trend, indicating the experimenter may be improving in creating the concentration as one explanation. However, this would mean that the variable Day is influencing Yield unintentionally, and should be factored in as a predictor too. Additionally, there are some "jumps" in the observations (e.g. 2 larger jumps between days 40-60), that could indicate some kind of inconsistency during measurements.

```
plot(catalyst_data$Day, catalyst_data$Intermediate_Concentration,
     pch = 16, # this lets us use black filled in points instead
     xlab = "Day",
     ylab = "Intermediate Concentration",
     main = "Sequence Plot")
```



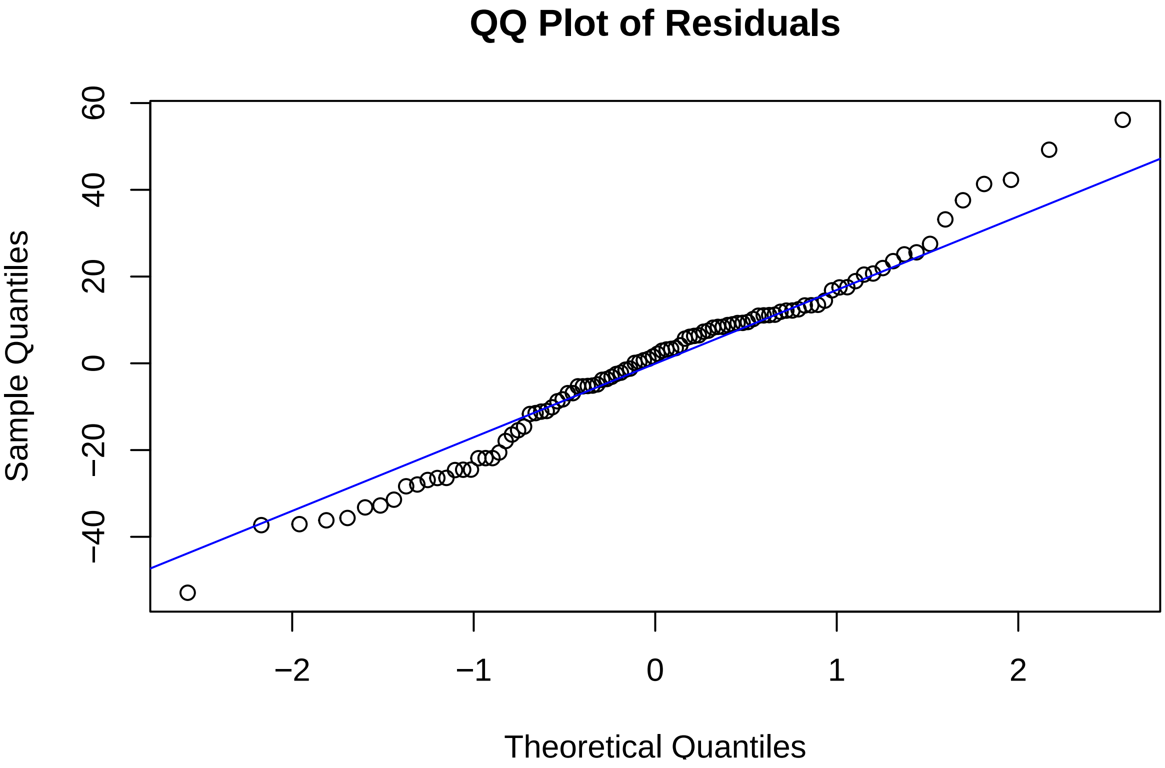
Next, we'll check the residuals vs fitted values to see if our assumptions on linearity and homoscedasticity are met. Below, it appears that the scatter around 0 is mostly random and no curves are apparent. There are instances of points being quite far from 0 though, and more towards the right of the plot it appears the variance may be increasing, which indicates a chance of heteroscedasticity.

```r
plot(catalyst_model$fitted.values, residuals(catalyst_model),
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs Fitted Values",
     pch = 16)
abline(h = 0, col = "blue")
```



We check the QQ plot to see from another perspective whether the data follows a normal distribution. While most of the residuals follow the diagonal, both tails deviate away, indicating that the data has more extreme values that aren't accepted under normality, aka "heavy tails".

```
qqnorm(residuals(catalyst_model), main = "QQ Plot of Residuals")
qqline(residuals(catalyst_model), col = "blue")
```


QQ Plot of Residuals
Sample Quantiles
Theoretical Quantiles