HW week 12

w203: Statistics for Data Science

Kamaljeet Ghotra

Contents

```
## NB: REQUIRES INTERNET CONNECTION
if(!require(pacman)){

        install.packages("pacman")
}

## load required packages
pacman::p_load(tidyverse, ggplot2, sandwich, stargazer, skimr, ggthemes, GGally, kableExtra, corrplot, l

## Theme set
theme_set(theme_clean(base_size = 8))
```

```
d <- load_and_clean(input = 'videos.txt')
```

Regression analysis of YouTube dataset

In this project, I seek to to explain how much the quality of a video affects the number of views it receives on social media. In a world where people can now buy followers and likes, would such an investment increase the number of views that their content receives? **This is a causal question.**

I use a dataset created by Cheng, Dale and Liu at Simon Fraser University. It includes observations about 9618 videos shared on YouTube. Please see this link for details about how the data was collected.

The dataset has 10 variables. However, we focus on the following variables:

- `views`: the number of views by YouTube users.
- `average_rating`: This is the average of the ratings that the video received, it is a renamed feature from `rate` that is provided in the original dataset. (Notice that this is different from `cout_of_ratings` which is a count of the total number of ratings that a video has received.
- `length`: the duration of the video in seconds.

Exploratory Data Analysis (EDA)

a. I start by performing a brief exploratory data analysis on the data to discover patterns, outliers, or wrong data entries and summarize my findings.

```
## Visualizing the variables
d %>%

        select(views, average_rating, length) %>% ggpairs()
```

```
################################################
## Summary statistics
d %>% select(views, average_rating, length) %>%

        skim_without_charts() %>% select(-skim_type, -complete_rate) %>%

        kbl(., booktabs = TRUE, caption = "Summary Statistics") %>%

        kable_classic(full_width = FALSE,

        latex_options = "hold_position", font_size = 8)
```

```
d %>% filter(duplicated(.))
```

```
## # A tibble: 0 x 10
```

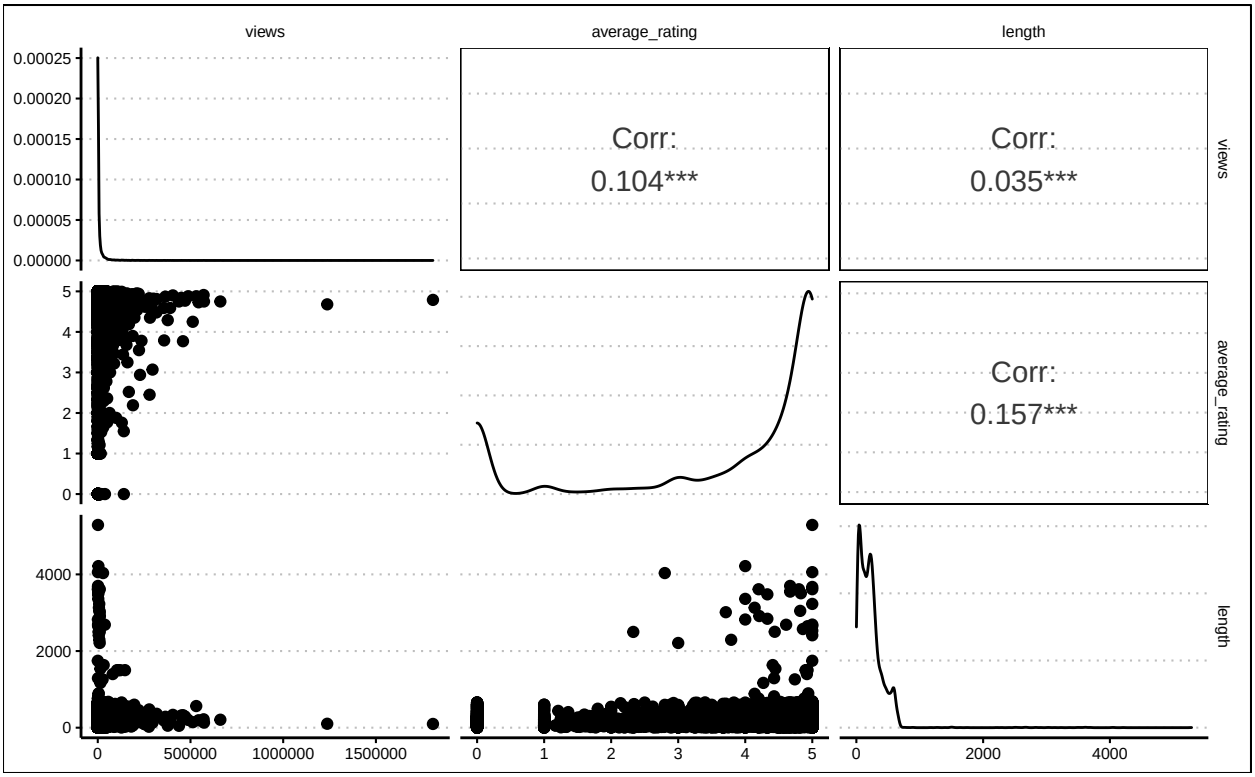

Figure 1: Pairs Plots

Table 1: Summary Statistics

skim_variable	n_missing	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100
views	9	9346.156936	37167.905412	3	348.0	1453.00	6179	1807640
average_rating	9	3.744057	1.788787	0	3.4	4.67	5	5
length	9	226.990009	238.788390	1	83.0	193.00	299	5289

```
## # ... with 10 variables: video_id <chr>, uploader <chr>, age <dbl>,
## #   category <chr>, length <dbl>, views <dbl>, average_rating <dbl>,
## #   count_of_ratings <dbl>, comments <dbl>, log_of_average_rating <dbl>
```

```r
d %>%
        ggplot(mapping = aes(x = views)) + geom_histogram(color = "black") +
        scale_x_log10(labels = scales::comma_format()) +
        labs(x = "Views (Log Scale)", y = "Count", title = "Distribution of Views (Log Scale)") +

d %>%
        ggplot(mapping = aes(x = average_rating)) +
        geom_histogram(color = "black") + labs(x = "Average Rating (Log Scale)", y = "Count",
             title = "Distribution of Average Rating, Log Scale")  +
d %>%
        ggplot(mapping = aes(x = length)) +
        geom_histogram(color = "black") + scale_x_log10(labels = scales::comma_format()) +
        labs(x = "Average Rating, Log Scale", y = "Count",
             title = "Distribution of Average Rating, Log Scale")
```

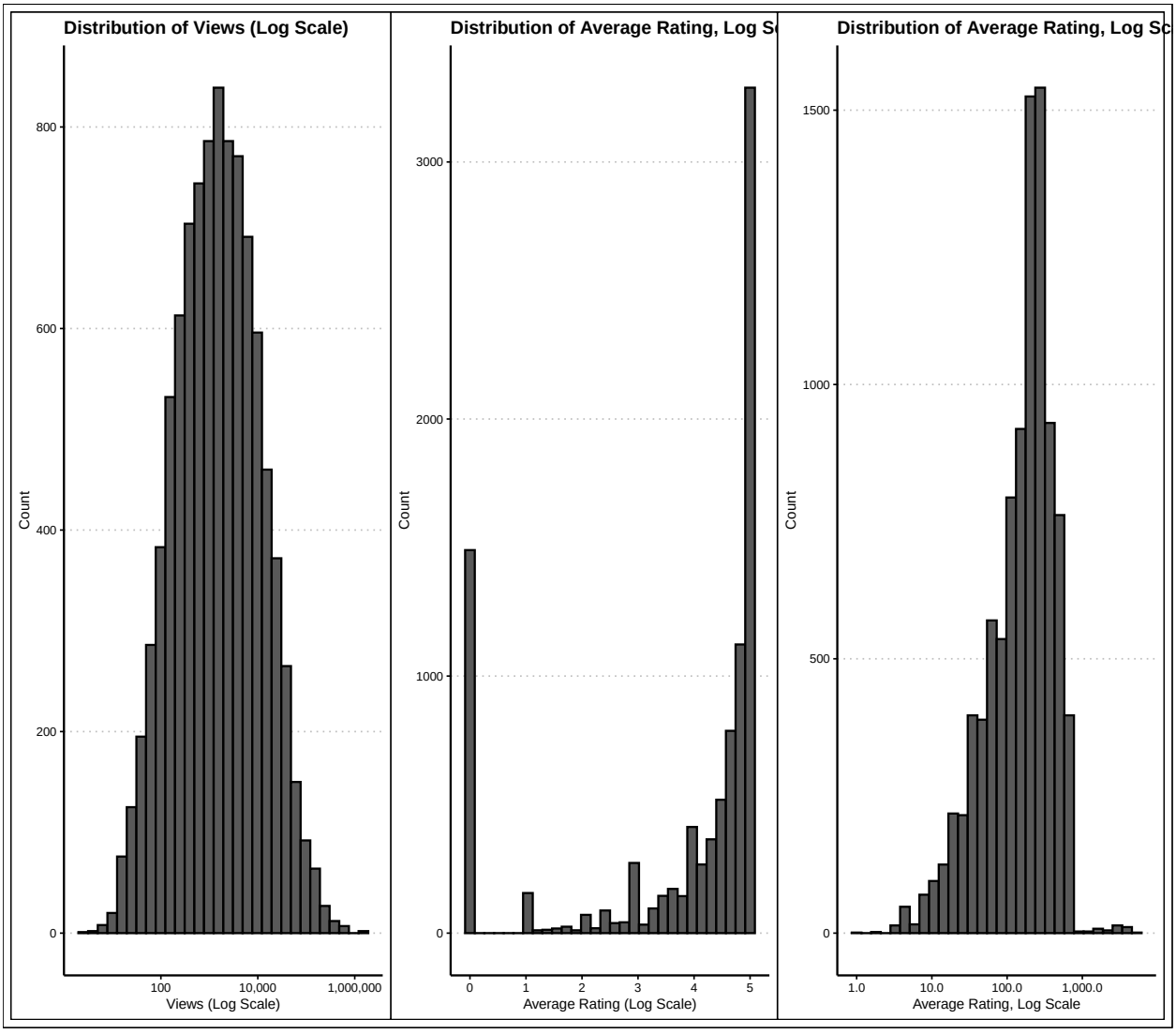

Figure 2: Distribution of Variables for the Regression Model

The EDA shows that the variables `views`, `average ratings` and `length` are highly skewed. Both views and length skew to the left while average rating skews to the right. The mean for views and length are hence substabtially larger than the median. On the contrary, the mean for average rating is substantially smaller than the median. There is also the presence of outliers in all cases, more so for views and length but which does not appear to be out of the ordinary as some movies could perform extraordinarily well while most attract medium or low views. Also, there is low correlation between the independent variables average ratings and length of YouTube videos.

Regression Model

b. Based on your EDA, I select an appropriate variable transformation (if any) to apply to each of my three variables. I will fit a model of the type,

$$f(\text{views}) = \beta_0 + \beta_1 g(\text{rate}) + \beta_3 h(\text{length})$$

Where f, g and h are sensible transformations, which might include making *no* transformation.

Due to the high skewness in the variables and the presence of outliers, I choose to log transform them. Note that in the case of average rating, we have some videos with zero rating. I add a small value (0.01) to each movie rating to eliminate these zero ratings to allow for the use of logarithm transformation. I next fit the model.

```
## NB: I add a small quantity to average rating as some videos have a
## zero average rating.
model <- lm(log(views) ~ log(average_rating + 0.01) + log(length), data = d)

stargazer(model, type = 'text', se = list(get_robust_se(model)))
```

```
##
## ===================================================
##                              Dependent variable:
##                          -------------------------
##                                  log(views)
## ---------------------------------------------------
## log(average_rating + 0.01)         0.397***
##                                     (0.007)
##
## log(length)                        0.127***
##                                     (0.017)
##
## Constant                           6.446***
##                                     (0.087)
##
## ---------------------------------------------------
## Observations                         9,609
## R2                                   0.210
## Adjusted R2                          0.210
## Residual Std. Error            1.775 (df = 9606)
## F Statistic                1,277.256*** (df = 2; 9606)
## ===================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

Regression Diagnostics

Using diagnostic plots, background knowledge, and statistical tests, I assess all five assumptions of the CLM. When an assumption is violated, I state what response you will take. As part of this process, I decide what transformation (if any) to apply to each variable. I Iterate against the model until I am satisfied that at least four of the five assumption have been reasonably addressed.

1. IID Assumption

IID Data: The data we have is a sample of the millions of videos uploaded on `YouTube`. For the IID assumption to hold, there are two conditions. First, if we were to pick another sample of videos from YouTube, this new sample should have the same distribution as our original sample data. Also, all samples must be mutually independent such that every video has an equal chance of selection into a sample. One way to ensure IID is met is to use random sampling or use the entire population. It is not clear how the sampling for the 9618 videos was done making it hard to assess the IID assumption. However, if the movies are a random selection from the population, then they meet this IID condition.

2. Multicollinerarity

No Perfect Collinearity: Figure 1 shows that the variables `average rating` and `length` have a correlation of 0.157 which, though significant, is far below the perfect collinearity values of +1 or -1. Hence our data meets this condition of "No Perfect Collinearity".

3. Linear Conditional Expectation

Linear Conditional Expectation: In linear regression, the assumption is that the dependent variable has a linear relationship with each of the independent/explanatory variables. Figure 2 clearly shows that the raw values of the variables do not have a linear relationship with the dependent variables. This is the reason we transform the variables. The transformed variables show a resonable linear relationship with the dependent variable. The model meets the assumption of linear conditional expectation with the logarithm transformation.

```
(d %>%

        ggplot(mapping = aes(x = average_rating + 0.01, y = views)) +

        geom_point() + labs(x = "Average Rating", y = "Views",

             title = "Views vs. Average Ratings- Raw Data") +

        scale_y_continuous(labels = scales::comma_format()) +

d %>%

        ggplot(mapping = aes(x = length + 0.01, y = views)) +

        geom_point() + labs(x = "Length", y = "Views",

             title = "Views vs. Length- Raw Data") +

        scale_y_continuous(labels = scales::comma_format())) /


(d %>%
```

```
        ggplot(mapping = aes(x = average_rating + 0.01, y = views)) +
        geom_point() + labs(x = "Average Rating", y = "Views",
             title = "Views vs. Average Ratings- Transformed Data") +
        scale_y_log10(labels = scales::comma_format()) +
d %>%
        ggplot(mapping = aes(x = length + 0.01, y = views)) +
        geom_point() + labs(x = "Length", y = "Views",
             title = "Views vs. Length- Transformed Data") +
        scale_y_log10(labels = scales::comma_format()))
```

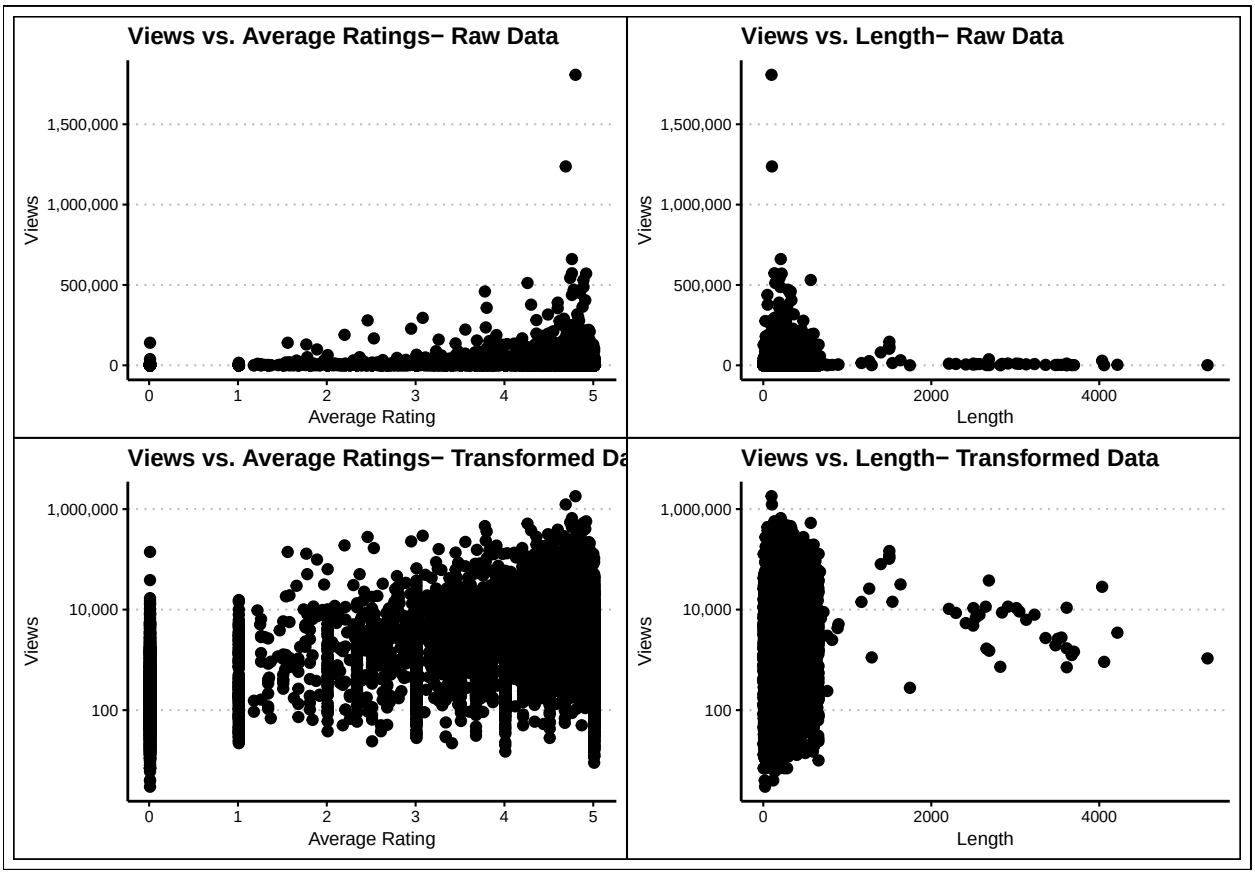

Figure 3: Assessing Linear Conditional Expectation

4. Homoscedasticity

Homoskedastic Errors: OLS regression assumes that the sample (and, hence, the residuals) comes from a family that has constant variance. Hence, we do not expect unequal scatter of

residuals over the range of measured values (Berenguer-Rico and Wilms 2021). In this case, I plot the standardized residuals against fitted values. For heteroscedasticity to be absent, I expect the resulting curve to be flat, meaning that the residuals do not vary with changes in observed values. In this case, the fitted line has a consistent upward slope meaning that there is heteroscedasticity. Hence, the regression model does not meet the requirements for homoscedasticity. However, OLS estimators remain unbiased and consistent but inefficient. ***To correct for heteroscedasticity, we present heteroscedasticity robust standard errors in the model.***

```
het_data <- augment(model)

het_data %>% ggplot(mapping = aes(x = .fitted, y = sqrt(.std.resid))) +

        geom_point() + geom_smooth(method = "lm", se = 0) +

        labs(x = "Fitted Values- Predictions",

             y = "Square Root Standardized Residuals",

             title = "Test for Heteroscedasticity")
```

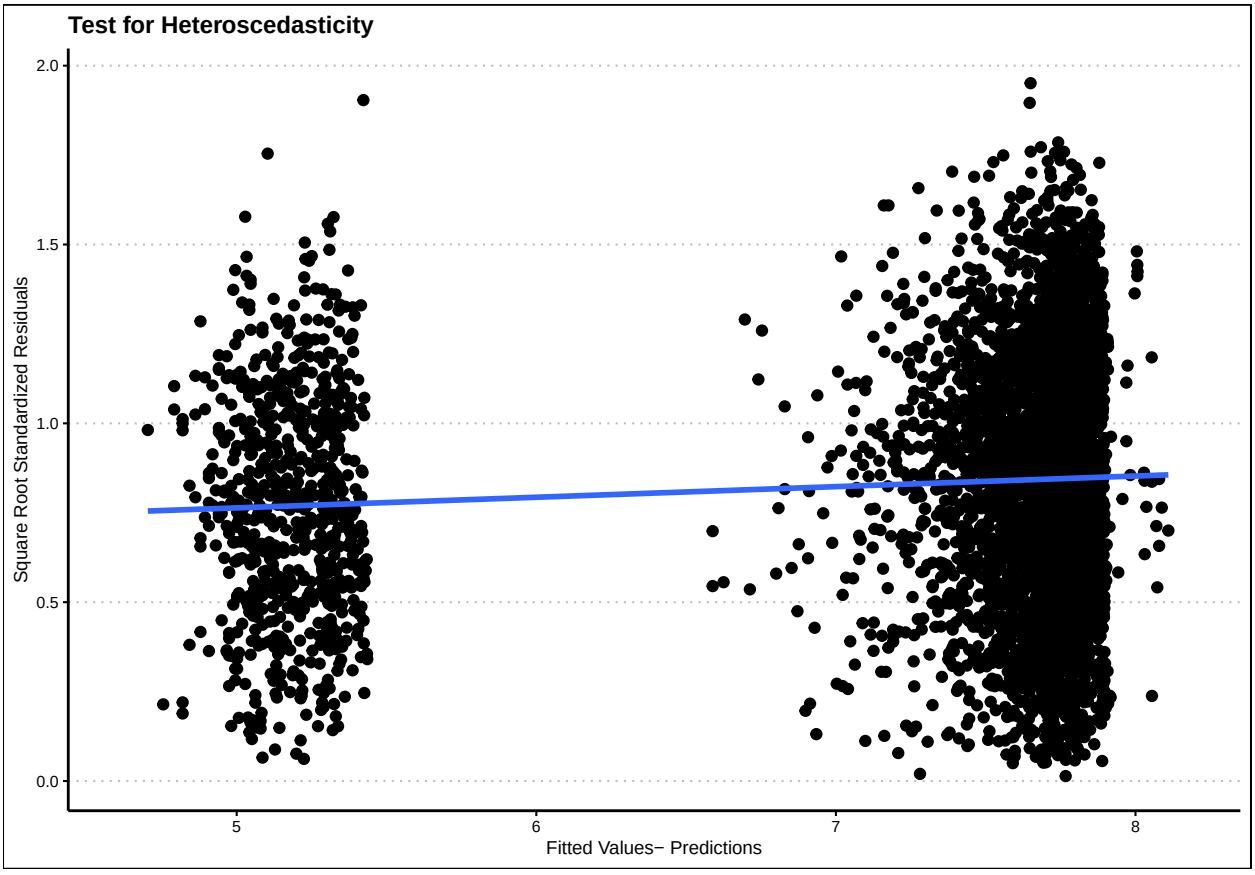

Figure 4: Assessing Error Variance

5. Normally Distributed Errors

Normally Distributed Errors: Here I examine whether the residuals are normally distributed. In this case, the scatter plot of the sample residuals quantiles against theoretical residual quantiles should fall roughly on a straight line. Figure 5 below shows that indeed, the error terms are normally distributed except for a few outlying observations. I am satisfied that in my model I have satisfied the assumption of normally distributed errors (Li et al. 2012).

```
qqnorm(het_data$`log(views)` - het_data$.fitted, pch = 1, frame = FALSE)

qqline(het_data$`log(views)` - het_data$.fitted, col = "steelblue", lwd = 2)
```

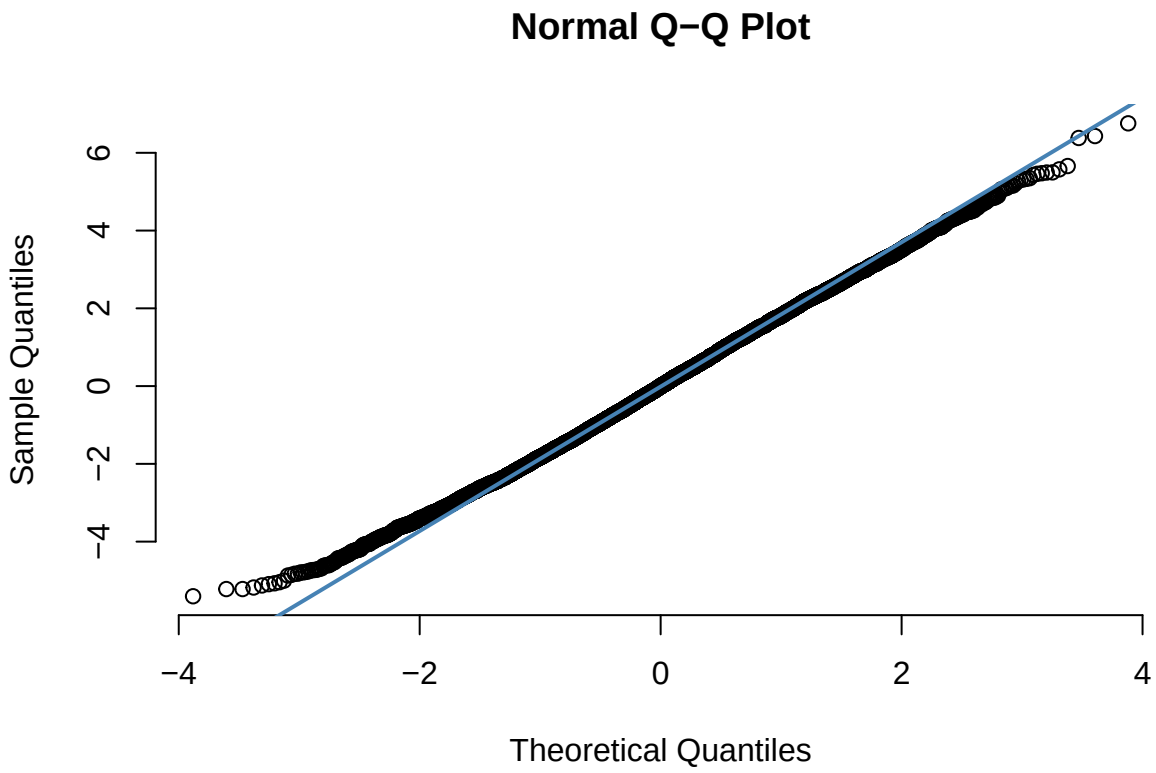

Figure 5: assessing normally distributed errors

References

Berenguer-Rico, Vanessa, and Ines Wilms. 2021. "Heteroscedasticity Testing After Outlier Removal." *Econometric Reviews* 40 (1): 51–85.

Li, Xiang, Wanling Wong, Ecosse L Lamoureux, and Tien Y Wong. 2012. "Are Linear Regression Techniques Appropriate for Analysis When the Dependent (Outcome) Variable Is Not Normally Distributed?" *Investigative Ophthalmology & Visual Science* 53 (6): 3082–83.