Unit 12 Homework

w203: Statistics for Data Science

July 26, 2022

Part 2 - CLM Practice

For the following questions, your task is to evaluate the Classical Linear Model assumptions. It is not enough to say that an assumption is met or not met; instead, present evidence based on your background knowledge, visualizations, and numerical summaries.

The file `videos.txt` contains 9618 observations of videos shared on YouTube. It was created by Cheng, Dale and Liu at Simon Fraser University. Please see this link for details about how the data was collected.

You wish to run the following regression:

$$ln(\text{views}) = \beta_0 + \beta_1 \text{rate} + \beta_3 \text{length}$$

The variables are as follows:

- `views`: the number of views by YouTube users.
- `rate`: This is the average of the ratings that the video received. You may think of this as a proxy for video quality. (Notice that this is different from the variable `ratings` which is a count of the total number of ratings that a video has received.)
- `length:` the duration of the video in seconds.

1. Evaluate the **IID** assumption.
2. Evaluate the **No perfect Colinearity** assumption.
3. Evaluate the **Linear Conditional Expectation:** assumption.
4. Evaluate the **Homoskedastic Errors:** assumption.
5. Evaluate the **Normally Distributed Errors:** assumption.

```r
df <- read.delim('videos.txt')
```

Solution

1.IID

(1) Independent There are several factors which makes our sample data not IID. First, Youtubers will often copy elements of other Youtube videos that are successful or have “gone viral.” This means that one video can impact other videos, breaking independence. Another factor is that videos are chosen based on recommendations from an initial set of videos. “Recommended” videos disproportionally favor videos from the same category and even same creator. This means that the videos are not picked randomly from the set of all videos.

(2) Identical distributed our sample is identically distributed because the algorithm referenced above affects all users.

In short, due to a lack of independence the sample violates this assumption.

2.No perfect Colinearity There is no perfect colinearity between the causal variables video length and average ratings. In essence, no amount of linear transformations will allow for one variable to be represented by the other.

3.Linear Conditional Expectation

First, let me remove the empty rows.

```
# check empty row
summary(df['rate'])
```

```
##       rate      
##  Min.   :0.000  
##  1st Qu.:3.400  
##  Median :4.670  
##  Mean   :3.744  
##  3rd Qu.:5.000  
##  Max.   :5.000  
##  NA's   :9
```

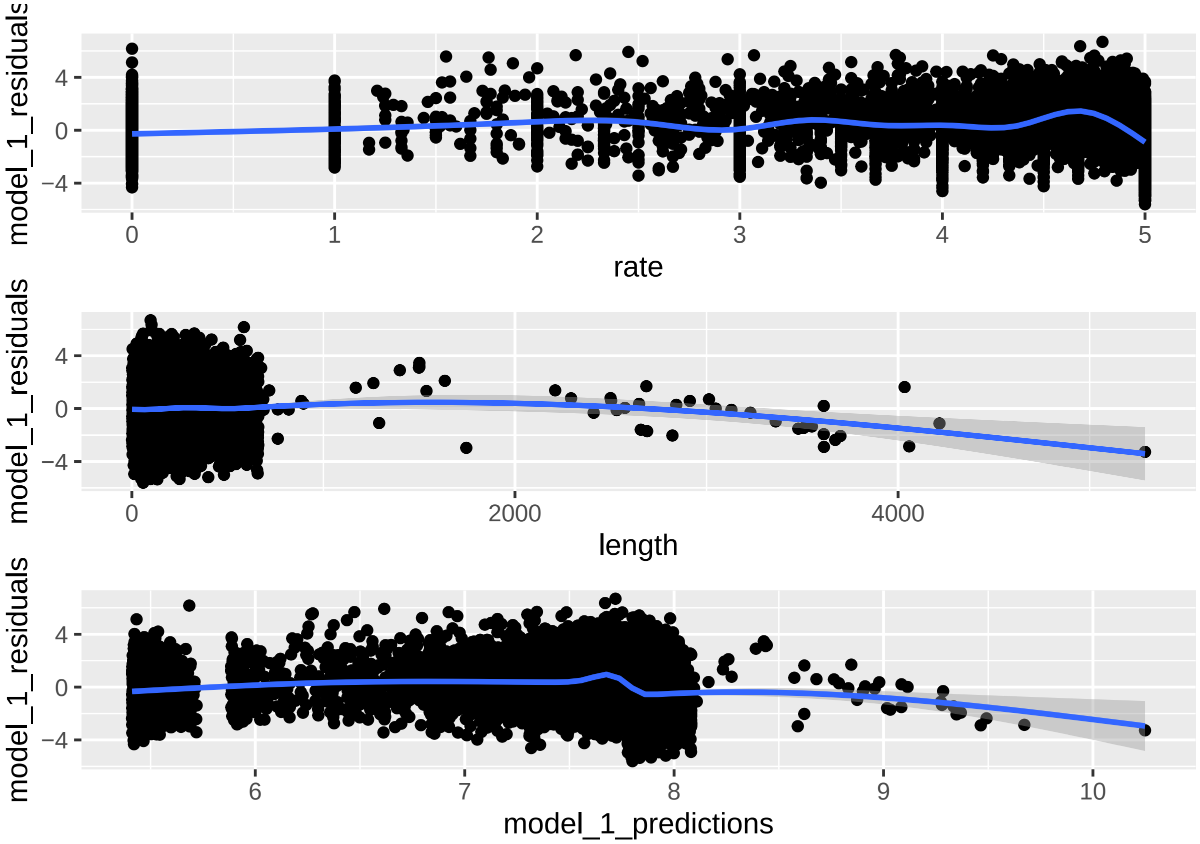
```
# drop empty rows
df <- df[!is.na(df$views),]
dim(df)
```

```
## [1] 9609    9
```

Second, let us check the outliers for rate, length and views.

```
model_1 <- lm(log(views) ~ rate + length, data = df)
# model_1


df <- df[!is.na(df$views),]

# df <- filter(df, views < 1000000)

# boxplot(df$views<1000000, data = df, subset = df$views)

df <- df %>%
  mutate(
    model_1_predictions = predict(model_1),
    model_1_residuals = resid(model_1)
  )
```

```
plot_model_1a <- df %>%
  ggplot(aes(x = rate, y = model_1_residuals)) +
  geom_point() + stat_smooth()

plot_model_1b <- df %>%
  ggplot(aes(x = length, y = model_1_residuals)) +
  geom_point() + stat_smooth()


plot_model_1c <- df %>%
  ggplot(aes(x = model_1_predictions, y = model_1_residuals)) +
  geom_point() + stat_smooth()

plot_model_1a / plot_model_1b / plot_model_1c
```



From above chart, there are lots of outliers for views, length and rate.

If the values in one column is out of the 1.5 standard error range of mean, then they are defined as outliers, and then the corresponding whole rows will be removed from the dateframe.

Remove the rate outliers and corresponding rows.

```
#find Q1, Q3, and interquartile range for values in column A
Q1 <- quantile(df$rate, .25)
Q3 <- quantile(df$rate, .75)
IQR <- IQR(df$rate)

#only keep rows in dataframe that have values within 1.5*IQR of Q1 and Q3
df <- subset(df, df$rate> (Q1 - 1.5*IQR) & df$rate< (Q3 + 1.5*IQR))

#view row and column count of new data frame
dim(df)
```

```
## [1] 8119   11
```

Remove the length outliers and corresponding rows.

```
#find Q1, Q3, and interquartile range for values in column A
Q1 <- quantile(df$length, .25)
Q3 <- quantile(df$length, .75)
IQR <- IQR(df$length)

#only keep rows in dataframe that have values within 1.5*IQR of Q1 and Q3
df <- subset(df, df$length> (Q1 - 1.5*IQR) & df$length< (Q3 + 1.5*IQR))
```

```r
#view row and column count of new data frame
dim(df)
```

```
## [1] 7968   11
```

Remove the views outliers and corresponding rows.

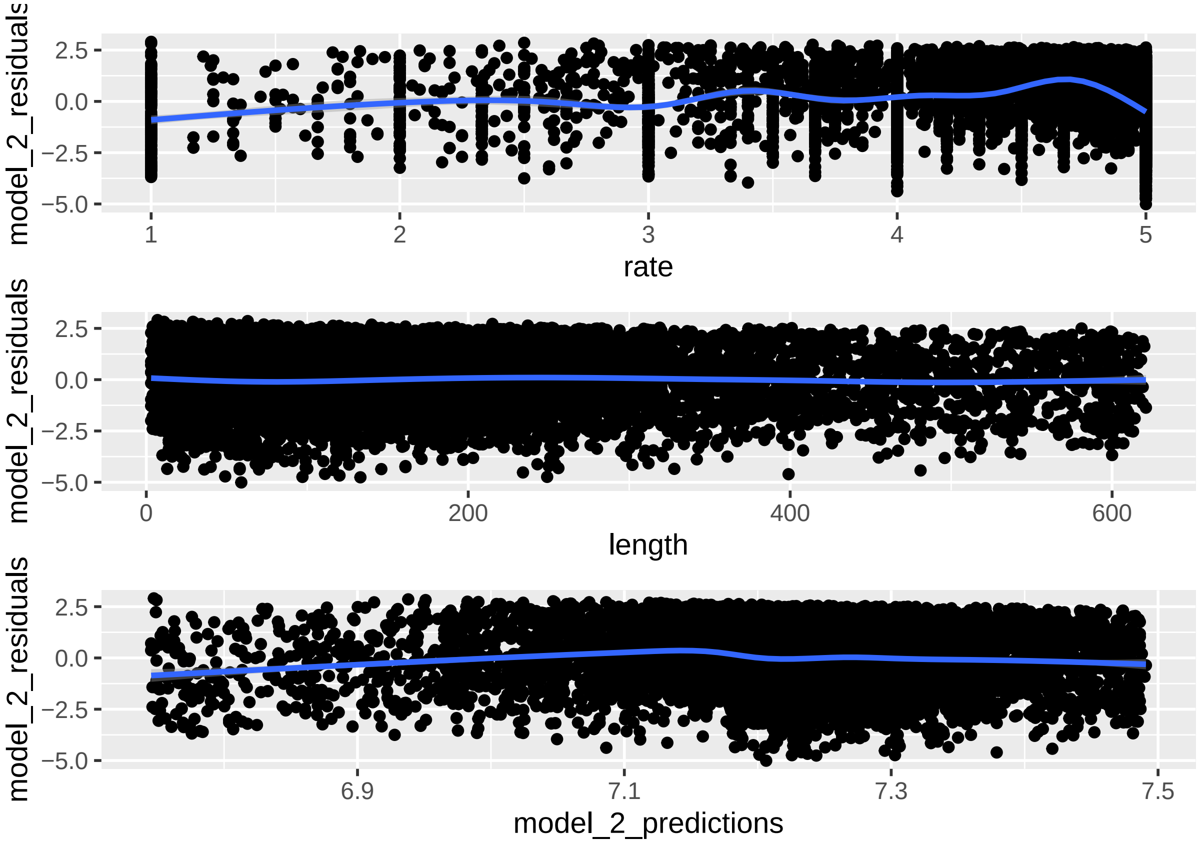
```r
#find Q1, Q3, and interquartile range for values in column A
Q1 <- quantile(df$views, .25)
Q3 <- quantile(df$views, .75)
IQR <- IQR(df$views)

#only keep rows in dataframe that have values within 1.5*IQR of Q1 and Q3
df <- subset(df, df$views> (Q1 - 1.5*IQR) & df$views< (Q3 + 1.5*IQR))

#view row and column count of new data frame
dim(df)
```

```
## [1] 6962   11
```

After the outliers are removed, the plots (residiual vs rate, residiual vs length and residiual vs predictions) are drawn.

```r
model_2 <- lm(log(views) ~ rate + length, data=df)

# df_2 <- df[!is.na(df$views),]
# df <- filter(df, views < 1000000)
# boxplot(df$views<1000000, data = df, subset = df$views)

df <- df %>%
  mutate(
    model_2_predictions = predict(model_2),
    model_2_residuals = resid(model_2)
  )
```

```r
plot_model_2a <- df %>%
  ggplot(aes(x = rate, y = model_2_residuals)) +
  geom_point() + stat_smooth()

plot_model_2b <- df %>%
  ggplot(aes(x = length, y = model_2_residuals)) +
  geom_point() + stat_smooth()


plot_model_2c <- df %>%
  ggplot(aes(x = model_2_predictions, y = model_2_residuals)) +
  geom_point() + stat_smooth()

plot_model_2a / plot_model_2b / plot_model_2c
```

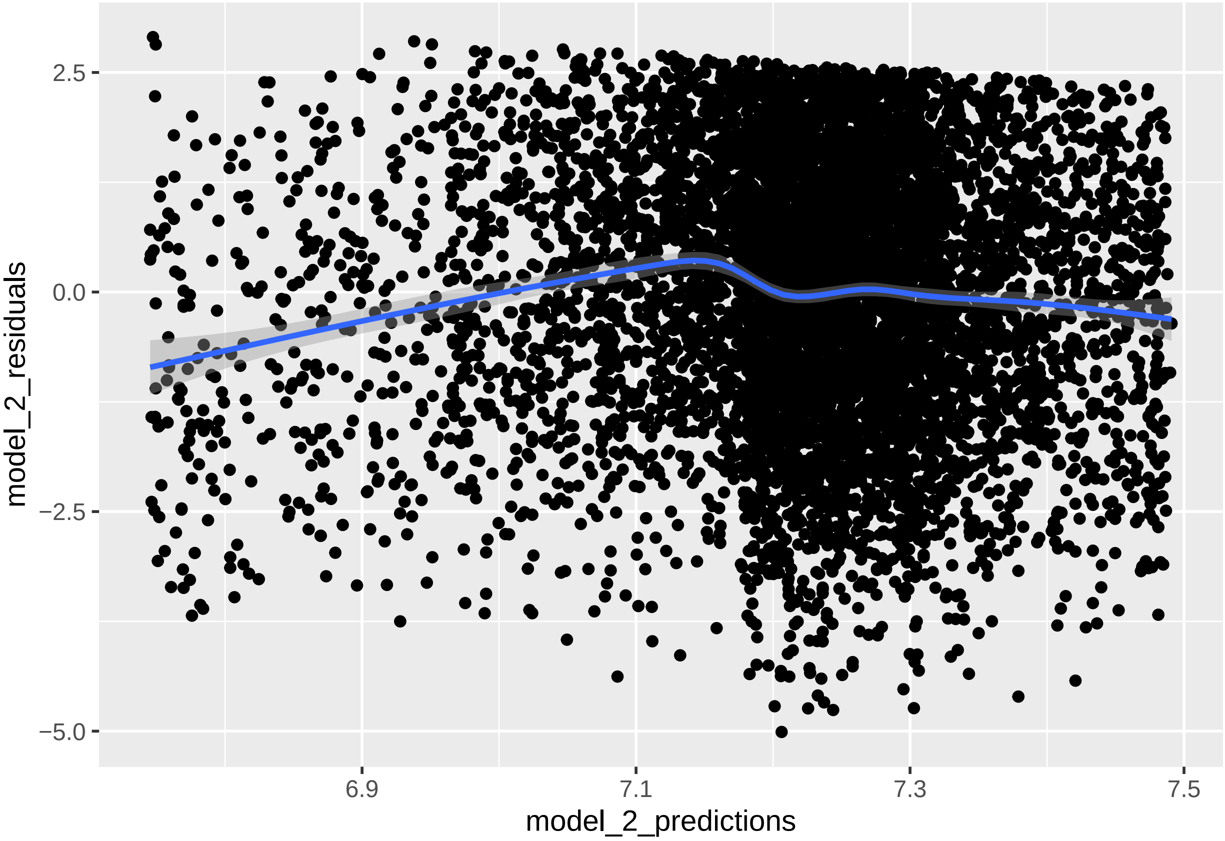


As seen in said chart above, there is only one place in the middle of our sample where we don't have a near perfect linear-fit. Overall, the model suffices to not violate this assumption.

4.Homoskedastic Errors

```
plot_model_2c
```

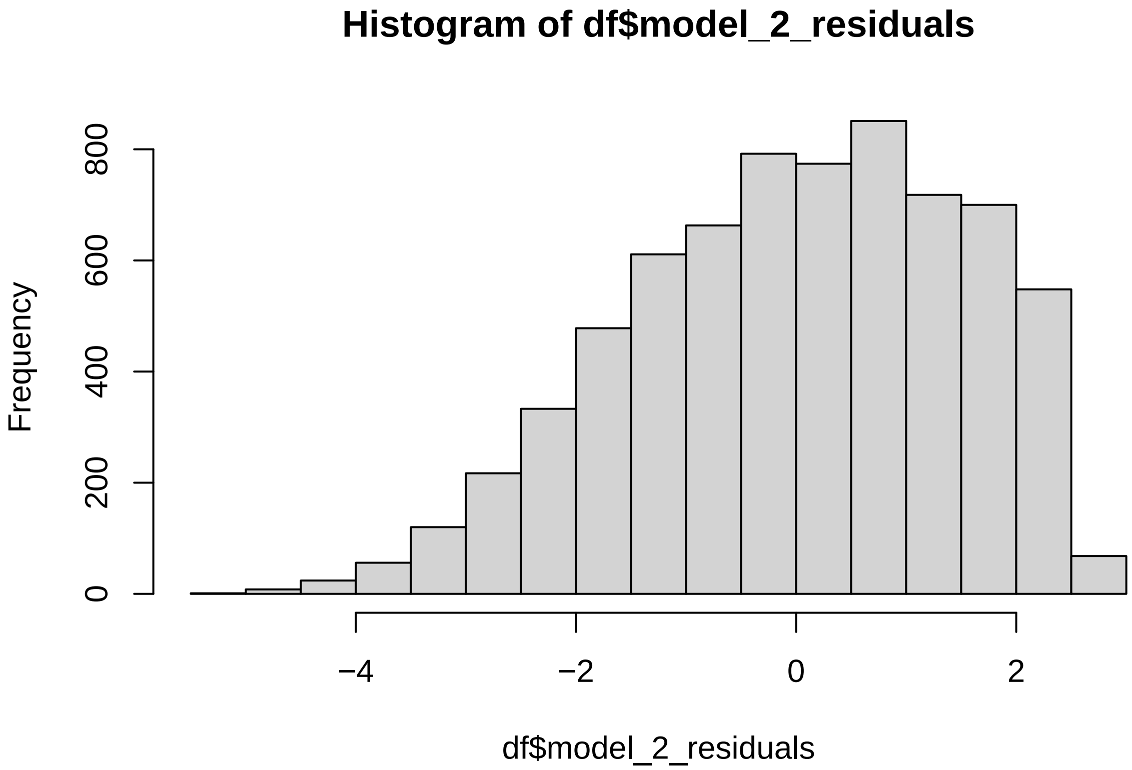
```
## ‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```



As seen in said chart above, variance is very consistent throughout the model, the difference on the y-axis between positive and negative data points only slightly changes. Therefore, the model is homoskedastic and does not violate this assumption.

5.Normally Distributed Errors

```
hist(df$model_2_residuals)
```



As seen in said histogram above, there is a left-tail. Thus, the sample is left-skewed and has normally distributed errors, which violates this assumption.