Assignment 2 of STAT5320

Feng Gu(T00751197), Anti Li(T00751339), Yuzhuo Ye(T00751492)

March 10, 2025

1 Preamble

1.1 Declaration of Using Generative Artificial Intelligence

The ideas and work are designed by the authors, with the use of ChatGPT in some parts of coding and grammar checking.

1.2 Support Materials for This Assignment

All relevant code and results are stored in the GitHub repository:
Click here to access the repository.

1.3 Solid text less than 3 pages

We have checked that after moving all tables and figures to the appendix, the main text is less than 3 pages.

2 Explore the relationship between X's and Y's

Before we build the linear model with MLE method, we can explore the relationship between X's and Y's by plotting the scatter plot.

The figure 1 shows the scatter plot of Anscombe's 4 different data set.

Based on the scatter plot, it is suitable to build linear models for some of the data sets, but others are not. Let us build simple linear models for each of them with MLE in estimating the coefficients.

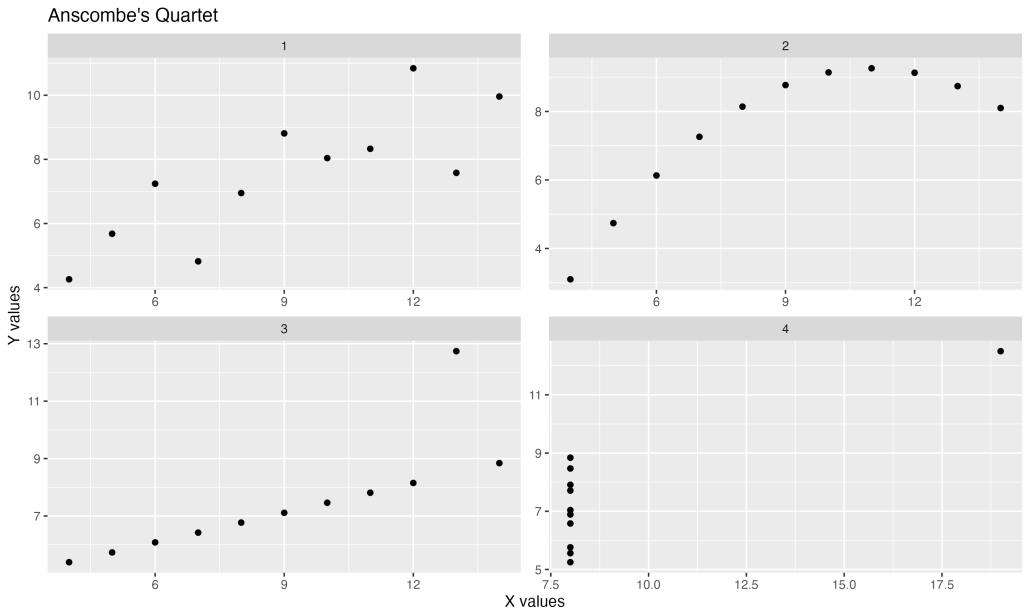

Figure 1: Anscombe's Quartet

3 Fit a linear model to each dataset

3.1 What the MLE method results are supposed to be

In each data set, we assume X probabilistically determines Y by assume the following:

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Given a data set, we can estimate the coefficients by maximizing the log-likelihood function, assuming $X's$ and $\epsilon's$ are 'iid'.

$$argmax_{\beta_0,\beta_1} \sum_{i=1}^{n} \log f(y_i|x_i, \beta_0, \beta_1)$$

If our assumption about the model and distribution of ϵ is correct, the residuals should be normally distributied with mean 0 and constant variance σ^2 on each level of X.

3.2 Model Fitting Results

The table 1 shows the linear coefficients, p-values for each data set with the MLE method we set in the previous section.

model	term	coefficients	std_errors	t_values	p_values	R^2
data 1	β_0	3.00	1.12	2.67	0.03	0.67
	β_1	0.50	0.12	4.24	0.00	
data 2	β_0	3.00	1.13	2.67	0.03	0.67
	β_1	0.50	0.12	4.24	0.00	
data 3	β_0	3.00	1.12	2.67	0.03	0.67
	β_1	0.50	0.12	4.24	0.00	
data 4	β_0	3.00	1.12	2.67	0.03	0.67
	β_1	0.50	0.12	4.24	0.00	

Table 1: Model Fitting Results

From the regression results, we found that all models have the same coefficients and R^2 value, even though the relationship between X's and Y's are different in each data set(shown in fig 1).

The figure 2 shows the fitted lines for each data set. We found the lines are the same but the relationship between X's and Y's are not always linear.

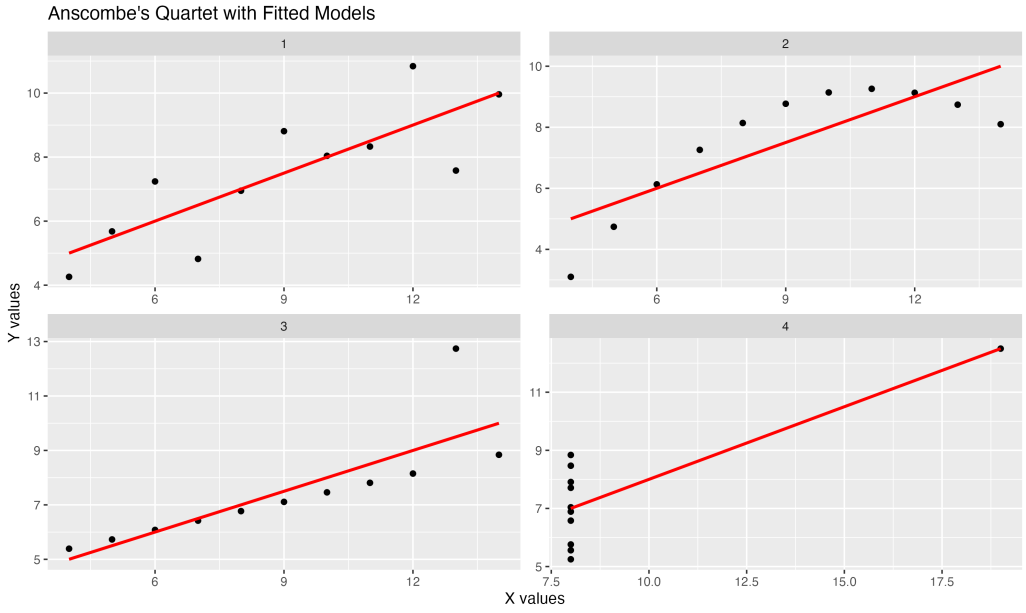

Figure 2: Anscombe's Quartet with Fitted Lines

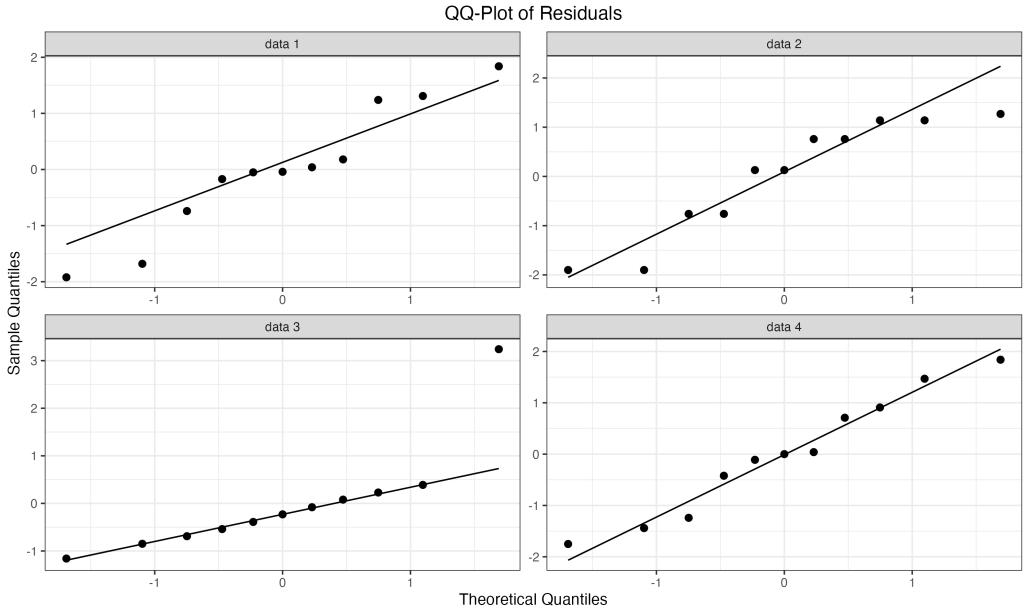

Figure 3: Q-Q Plot of Residuals

4 Residual Analysis for each data set

According to the assumption we made in the MLE method, if the residuals distribution were correctly spcified, the distribution of residuals from the models should be normal with mean 0 and constant variance σ^2, on each level of X, due to $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

We draw the residuals against each fitted values(which are corresponding to levels of X), the histogram of residuals and Q-Q plot of residuals for each data set, shown in figures 4, 5(appendix) and 3.

From these residual distribution plots, if we believe that these ϵ_i exhibit homoscedasticity with $E(\epsilon_i) = 0$, and they come from the same distribution, and are independent, we found that the residuals are not exactly normally distributed.

- For data set 1:

  The residuals exhibit a slight right skewness, indicating they are not perfectly normally distributed. However, this residual distribution is acceptable.

- For data set 2: The residuals have a heavy right tail, indicating that residuals are more likely to have higher values.

- For data set 3: The residuals appear very close to a normal distribution, but there is one extreme outlier in the regression, which affects the normality.
- For data set 4: The residuals appear closest to a normal distribution, but all except one are from a single level of X, which is not ideal for establishing the relationship between X and Y.

5 What we can glean about model fits from the residuals

From the simple linear models we built for each data set, it is clear that even though the relationship between X's and Y's can be different in levels of X, or the residuals distribution properties are different, we can still get the same coefficients and R^2 value for them.

Personally speaking, part of this problem might be caused by the assumption that ϵ_i are 'iid'. This assumption determines that all residuals are from a same distribution, regardless of the levels of X and the true relationship between X and Y. It casues that we equally consider each observation with the same weight and apply this regression relationship to other levels. However it should not be the trueth in real world.

Except the assumptions about reisudals, **the correct setting about the partially deterministic relationship between X and Y is also crucial in modeling.** Reviewing back to what we have fitted for data set 2(fig 2), the relationship between X and $\hat{Y}$ should be quadratic rather than linear. But we still use linear pattern to capture this relationship, which is lack of capability to capture the true relationship, causing the R^2 not high enough.

References

[1] Marek Hlavac. "stargazer: Well-Formatted Regression and Summary Statistics Tables," R package version 5.2.3, 2022. Available at: `https://CRAN.R-project.org/package=stargazer`.

6 Appendix

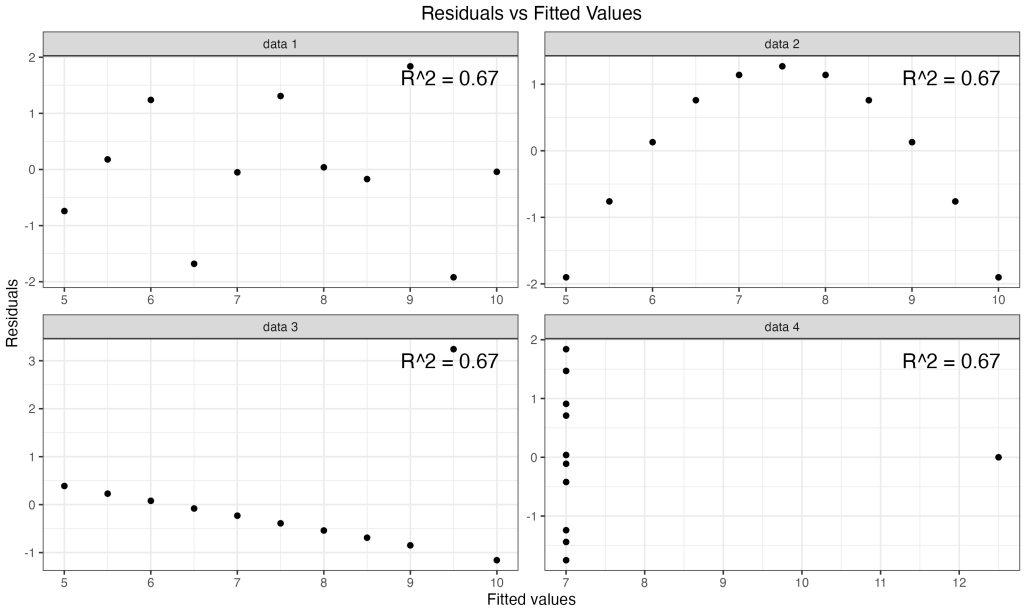

Figure 4: Residuals vs Fitted Values

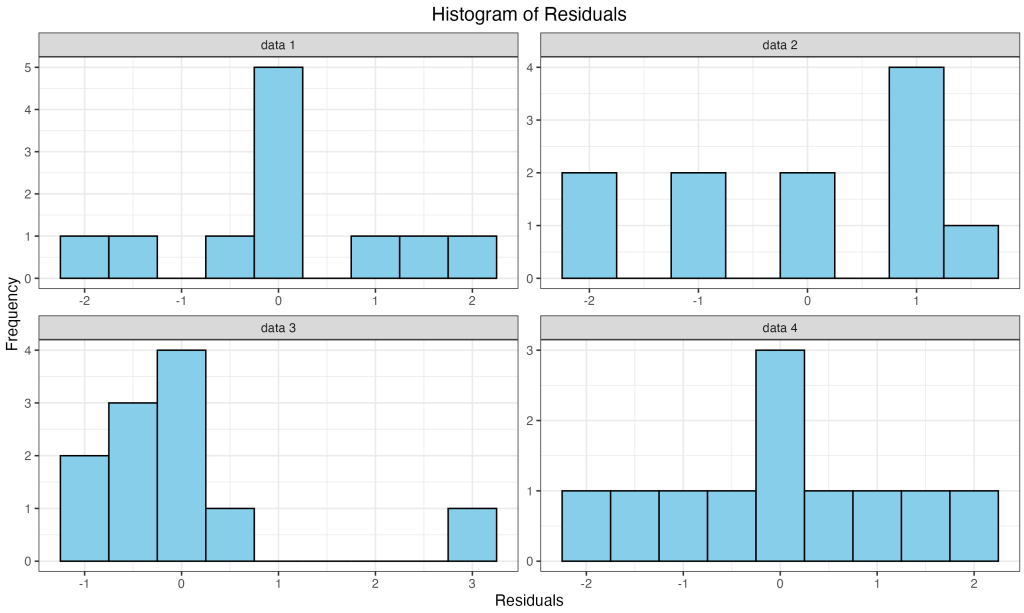

Figure 5: Histogram of Residuals