Project 3

Purpose:

The purpose of this analysis is to build a linear regression model and make sure it satisfies all regression conditions so as to be able to predict the gross earnings and examine the relationship between gross earnings and film budget.

Data:

The analysis uses a dataset of 100 randomly selected movies from the past 10 years. Our response variable of interest is USGross or the gross earnings of the movie in millions of USD. The explanatory variable is the budget of the movie in millions of USD.

Methods:

A simple linear regression model was employed to analyze the relationship between the budget and gross earnings of the movies as well as utilization of scatter plots, and possible transformations for the most optimal results.

$$\text{The population model is: } y_{\text{USGross}} = \beta0 + \beta1^{*}{}_{\text{Budget}} + \varepsilon$$

$$\text{The estimated model is: } \hat{y}_{\text{USGross}} = 14.12 + 0.91^{*}{}_{\text{Budget}}$$

According to Figure 1 as well as Table 1, with a p-value of less than 0.001, (t-test = 9.174, df=98), there is a strong, positive, linear relationship between a movie's budget and its gross earnings in the US.

Assessing the diagnostic assumptions using the diagnostic plots (Appendix C). First, independence can be assumed to be met because it was mentioned that the dataset contains information on 100 movies that were randomly chosen from the past 10 years. Linearity was also met as looking at Figure 2, at the residuals vs fitted plot, the line is fairly flat when the response is at zero.

Unfortunately, based on the residuals vs fitted plot, equal variance is not satisfied as the variability of the data points are inconsistent throughout fitted values. Normality is also violated as though most points fall along the dashed line, there are really large deviations on the two tails, especially the right tail along with a few potential outliers.

Since we do not have equal variance as well as normality satisfied, I have decided to perform log transformation on both our response and explanatory variables. As a result, according to Figure 2, we now have all four assumptions satisfied.

Results:

With log transformation, we have a new estimated model from the data from Table 2 (Appendix C):

$$\text{The estimated model is: } \hat{y}_{\text{USGross}} = 1.51 + 0.62^{*}{}_{\text{Budget}}$$

This can be interpreted as when the film budget is increased by one percent, it is estimated that the median value of the gross earning increases by 0.62%.

As a result, according to Figure 2, we now have all four assumptions satisfied and we will be able to use this model to predict the gross earnings of movies and quantify the relationship between the gross earnings and the film budgets.

Appendix A: Variable Descriptions

Budget: This variable represents the total expenditure on the movie’s production in millions of US dollars

USGross: This variable represents the total revenue generated by the movie from sales in the US in millions of dollars.

Appendix B: Scatter plot and estimated statistics for relationship between a movie's budget and its gross earnings in the US dataset before transformation.

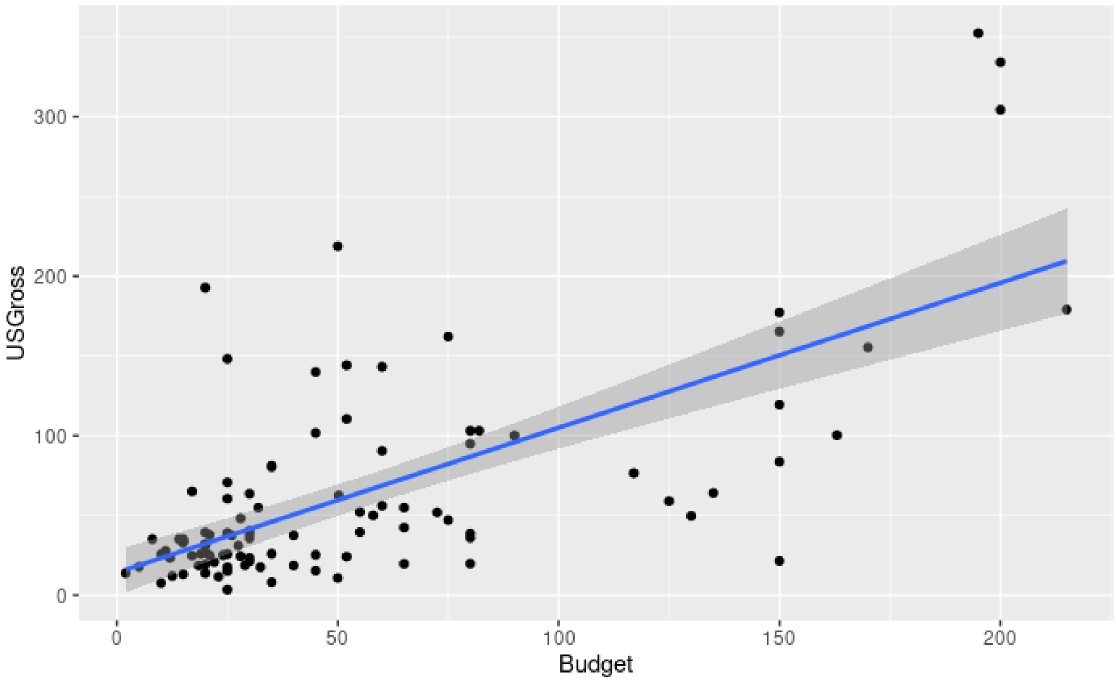

Figure 1: Scatter plot for the relationship between a movie's budget in millions of USD and its gross earnings in the US in millions of USD.

	Estimated values	Standard Error	T-statistic	P-value
(Intercept)	14.11976	7.34098	1.923	0.0573
Budget	0.90857	0.09904	9.174	7.52e-15

Table 1: Linear model statistics for the relationship between a movie's budget and its gross earnings in the US.

Appendix C: Diagnostic plots before dataset’s log transformation

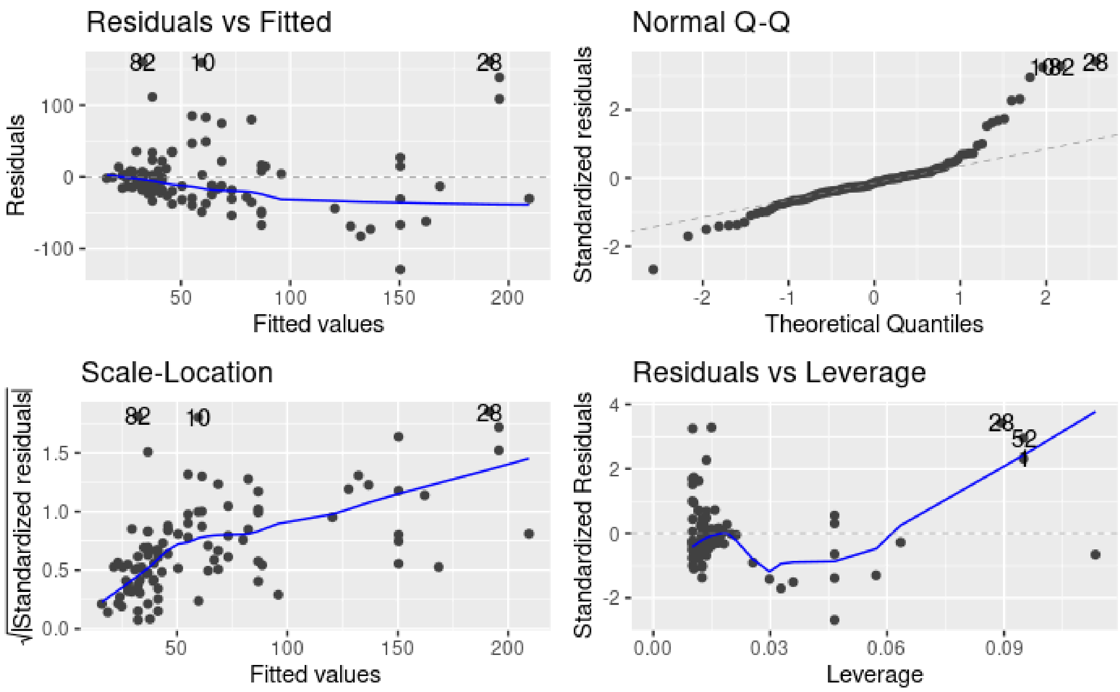

Figure 2: Diagnostic plots for the initial linear model (without transformations)

According to Figure 2, while linearity is met, normality as well as equal variances are questionable as according to the residuals vs fitted plot, the variability is not consistent across fitted values and on the normal Q-Q plot, the deviations are large on the two tails with a few potential outliers.

Appendix C: Diagnostic plots and estimated summary statistics after dataset’s log transformation

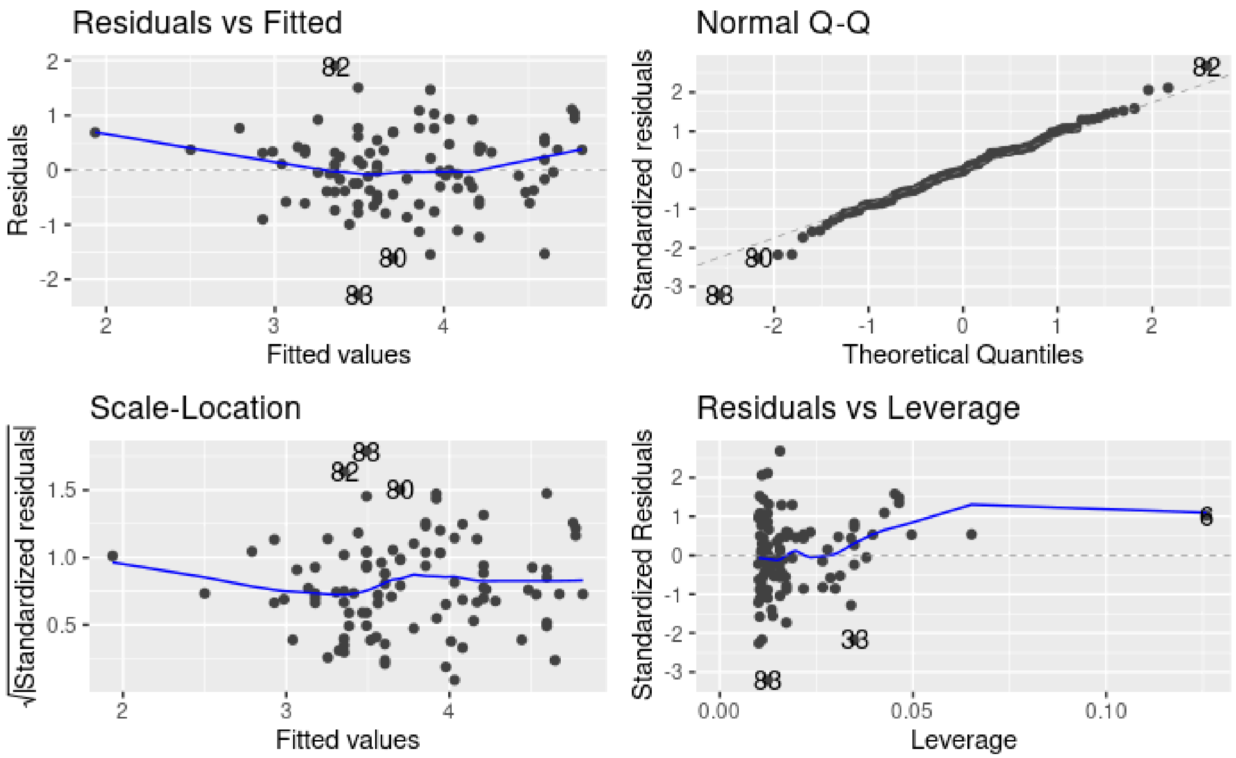

Figure 3: Diagnostic plots for the final model (after log transformations)

After trying log transforming individually on the response variable, explanatory variable, I found out it is the best to transform both variables. Figure 3 represents the final results where all four diagnostic assumptions are met.

	Estimated values	Standard Error	T-statistic	P-value
(Intercept)	1.50880	0.30976	4.871	4.27e-06
Budget	0.61606	0.08268	7.451	3.69e-11

Table 2: Linear model statistics for the relationship between a movie's budget and its gross earnings in the US after log transformation.