**Stats 271**

**A Statistical Analysis of Movie Metadata**

**Project271\_G7**

University of the Fraser Valley

Ali Rezah Fotouhi, Ph.D. Applied Statistics

11/27/19

Kajal Dhaliwal (300140508)

Logan Beck (300154293)

Joseph Charles (300166618)

Sehajveer Singh (300166729)

**Table of Contents**

1. Goals of the project p (2)
2. Description and exploratory analysis of the data p (2-3)
3. Methodology p (3)
4. Summary of the computer results p (3-8)
5. Interpretation of the computer output p (8-10)
6. Conclusion p (10-11)
7. Appendix p (12-21)

**Goals of the project**

* Determine correlation, and impact of the respective quantitative and qualitative explanatory variables: number of user reviews, number of critics, duration of movie, budget, number of cast Facebook likes, number of director Facebook likes, country of origin, and content rating on the dependent variables Success (Revenue - Budget) and IMDB rating
* Using tests of hypothesis, make determinations of whether the explanatory variables are significant in the outcome of success, and the IMDB ratings and also the overall significance of the model.
* Make predictions on IMDB\_scrore based on linear regression model, formed by methods: manual-backwards elimination, and adjusted R2, etc. and also use logistic regression to determine the significant variables that have an impact on the response variable, success or failure of the movie.

**Description and exploratory analysis of the data**

IMDB has been a reliable source for ratings of TV shows, movies and documentaries (Watson, 2018). The film industry is one of the major industries around the globe and forecasted to be a 50 Billion US Dollar Industry by 2020 (Watson, 2018). Understanding what factors contribute to the success of a movie has been a challenge for many. Movies that are highly successful can be rated low by critics and vice-versa, usually leading to confusion among viewers. The success of a film can sometimes solely be due to the great fan following of the main actor or could also be due to the theme of the movie, which may be relatable to all. In this project, we will analyze all the different aspects that could be helpful in predicting if a movie will be a success or not, as well as predicting the imbd\_score. We will be investigating the effect of six quantitative explanatory variables: number of user reviews, number of critics, duration, budget, number of cast Facebook likes, number of director Facebook likes, and two qualitative variables: country of origin, and content rating on the two dependent variables discussed earlier.

For our project, we will be using the imdb movie data set published on data.world (<https://data.world/popculture/imdb-5000-movie-dataset/workspace/file?filename=movie_metadata.csv%2Fmovie_metadata.csv>), providing 28 variables for 5043 movies spanning across 100 years in 66 countries. After filtering and sorting the data to include movies from 2010 to 2016 and limiting the countries to the top five, we narrowed our data from 5043 to 1340 observations. We also removed a few observations that were missing some key information which could have altered the outcome of the project. Later we randomly selected 25% of the filtered dataset using SAS Enterprise Guide application to get the final dataset for analysis with 245 observation.

**Methodology**

For our project, we use linear regression to fit a model to predict the imbd\_score for movies. This will include the six quantitative explanatory variable and two qualitative variables discussed above. We have used transformation method to make sure it satisfies the model assumption. Also reduce the model by fitting variable that are significant at 10% level. Based on AIC, Adjusted R-square, Mean Square Error, F-Value and its overall significance we will choose the best model that will fit this dataset.

We will also be using logistic regression to predict if a movie will be successful or not. For each of the 245 observations, the cost of the film was subtracted from the revenue and if the resulting number was positive, in a separate column, we made success = 1. If the resulting value was 0 or less than 0 we made success = 0, concluding that the movie, in fact, was not successful. Using the same explanatory variables, a full logistic regression model will be fitted. After testing the overall significance of the model as well as the significance of each of the predictors, another model will be fitted to include only the explanatory variables that are significant at the 5% mark.

**Summary of the computer results**

**Multiple Linear Regression**

**Model 1:**

y = βo + β1X1 + β2X2 + β3X3 + β4X4 + β5X5 + β6X6 + β7X7 + β8X8 + β9X9 + β10X10  + β11X11 + β12X12 + β13X13 + β14X14 + ↋  ↋ ~ iid N (0, σ2)

 X1:number of user views             X8: Country of origin = 1 if France, 0 otherwise

X2: number of critics     X9: Country of origin = 1 if UK,  0 otherwise

X3:duration of movie                 X10: Country of origin = 1 if Canada, 0 otherwise

X4: budget                 X11: Content rating = 1 if G, 0 otherwise

X5: number of cast Facebook likes                     X12: Content rating = 1 if PG, 0 otherwise

X6:  number of director Facebook likes     X13: Content rating = 1 if PG13, 0 otherwise

X7:Country of origin = 1 if USA, 0 otherwise   X14: Content rating = 1 if R, 0 otherwise

**See figure 1.1**

ŷ = 6.27137 - 0.00038349X1 + 0.00419X2 + 0.01347X3 + 0.0000X4 + 0.0000X5 + 0.00000967X6 - 1.19579X7 -1.32538X8 - 0.79771X9 -1.00079X10 + 0.26436X11 - 0.73806X12 -1.20629X13 -1.05099X14

**Significance Tests:**

Ho: β1= β2=.... βi= 0

Ha: at least one βi ≠ 0

Furthermore, we can do a hypothesis test on the normality of distribution of residuals.

Ho: Residuals have normal distribution

Ha: Residuals doesn’t have normal distribution

Refer Figure 1.4. Goodness-of-Fit Test for Normal Distribution for result.

**Data Transformed for Model 2**

Based on above test we couldn’t satisfy error assumption of the model. So, we try to fit a model with response variable IMDB score transformed to (IMDB\_score) ^2

**Model 2:**

Square of imdb\_score = ty (transformed)

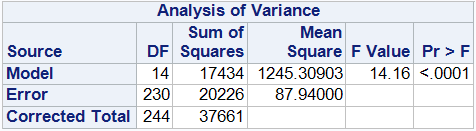
tŷ = 44.31024- 0.00242X1 + 0.05056X2 + 0.16101X3 -.0.0000X4+ 0.00001145X5+ 0.00013769X6 -18.22511X7 - 19.73169X8 - 13.18316X9 - 15.43887X10 + 4.91786X11 - 8.64454X12- 14.92158X13- 13.25245X14

ŷ = √44.31024- 0.00242X1 + 0.05056X2 + 0.16101X3 -.0.0000X4+ 0.00001145X5+ 0.00013769X6 -18.22511X7 - 19.73169X8 - 13.18316X9 - 15.43887X10 + 4.91786X11 - 8.64454X12- 14.92158X13- 13.25245X14

**Significance Tests:**

Ho: β1= β2= βi= 0

Ha: at least one βi ≠ 0



**Figure 2.1. Analysis of Variance**

Furthermore, we can do a hypothesis test on the normality of distribution of residuals.

**H0**: Residuals have normal distribution

**Ha**: Residuals doesn’t have normal distribution

**Figure 2.2 Goodness-of-Fit Test for Normal Distribution**

**Model 3:**

**Reduced Transformed Linear Model:**

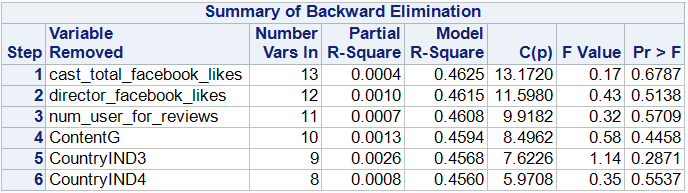
X2: number of critics X8: Country of origin = 1 if France, 0 otherwise X3:duration of movie X12: Content rating = 1 if PG, 0 otherwise X4: budget X13: Content rating = 1 if PG13, 0 otherwise X7:Country of origin = 1 if USA, 0 otherwise X14: Content rating = 1 if R, 0 otherwise

Square of imdb\_score = ty

ŷ =32.16048+0.04775X2+ 0.17122X3 -3.14135E-8X4- 5.11935X7- 6.62157X8- 10.33990X12- 16.81251X13- 15.09183X14

ŷ = √32.16048+0.04775X2+ 0.17122X3 -3.14135E-8X4- 5.11935X7- 6.62157X8- 10.33990X12- 16.81251X13- 15.09183X14

After backward elimination we can see that the  total of six variables are removed there is decrease in MSE. By using backward elimination we eliminated below variables which had higher p-value at 10% significance level.

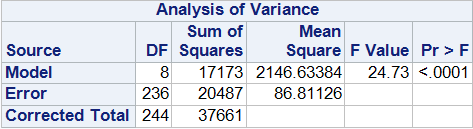


**Figure 3.1. Backwards Elimination Summary**

**Overall Significance Tests for Reduced Model:**

Ho: β1= β2= βi= 0

Ha: at least one βi ≠ 0



**Figure 3.2. Analysis of Variance**

Further we can do a hypothesis test on the normality of distribution of residuals.

**H0**: Residuals have normal distribution

**Ha**: Residuals doesn’t have normal distribution

**Figure 3.5 Goodness-of-Fit Test for Normal Distribution**

Logistic Regression

**Model 4:** logit (**) = o+ 1**X1 + **2**X2  + **3**X3 + **4**X4 + **5**X5 + **6**X6  + **7**X7  + **8**X8 + **9**X9 + **10**X10  + **11**X11 + **12**X12 + **13**X13  +**14**X14

= P (Y=1)         P=(Success =1)

X1:number of user views             X8: Country of origin = 1 if France, 0 otherwise

X2: number of critics     X9: Country of origin = 1 if UK,  0 otherwise

X3:duration of movie                 X10: Country of origin = 1 if Canada, 0 otherwise

X4: budget                 X11: Content rating = 1 if G, 0 otherwise

X5: number of cast Facebook likes                     X12: Content rating = 1 if PG, 0 otherwise

X6:  number of director Facebook likes     X13: Content rating = 1 if PG13, 0 otherwise

X7:Country of origin = 1 if USA, 0 otherwise   X14: Content rating = 1 if R, 0 otherwise

**Fitted Model:** logit ()= -13.294+ 0.00327 X1 + 0.00244 X2 + 0.00640 X3 - 0.0000000206 X4 -0.0000152 X5 -0.00005 X6 + 11.527 X7 + 10.562 X8 + 11.43 X9 +1.180 X10 + 13.888 X11 + 1.878X12 + 0.908 X13 +0.370X14

**Significance Tests:**

Overall significance:  Ho: **1= 2=** i= 0

Ha: at least one **i0**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **Degrees of Freedom** | **Pr > ChiSq** |
| **Likelihood Ratio** | 56.124 | 14 | <0.0001 |
| **Score** | 46.489 | 14 | <0.0001 |
| **Wald** | 31.099 | 14 | 0.0054 |

**Table 4.1. Testing Global Null Hypothesis: BETA=0**

Significance of each predictor

Ho: i= 0 vs. Ha: at least onei**0** where i = 1,2,3,...13

|  |  |
| --- | --- |
| **Variables significant at =10%** | **Variables Significant at =5%** |
| Number of user reviews | Number of user reviews |
| Budget | budget |
| Content rating PG | Content rating PG |
| Content rating PG-13 |  |

**Table 4.2. Significant predictors at 5 and 10%**

**Refer to Figure 4.2 in appendix.**

|  |  |  |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| AIC | 341.442 | 313.318 |
| SC | 344.943 | 365.837 |
| -2 Log L | 339.442 | 283.318 |

**Table 4.3. Model Fit Statistics**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criterion** | **Value** | **DF** | **Value/DF** | **Pr > ChiSq** |
| Deviance | 283.3180 | 227 | 1.2481 | 0.0065 |
| Pearson | 236.3133 | 227 | 1.041 | 0..3219 |

**Table 4.4. Goodness-of-Fit Statistics**

**Model 5:** logit (**) = o+ 1**X1 + **4**X4 + **7**X7

-**See table 5.4 in the appendix.**

Reduced Model: logit () = -1.4000 + 0.00346X1 - 10^-8X4 + 1.1850X7

Overall significance:  Ho: **1= 2=** i= 0

Ha: at least one **i0**

**See table 5.5 in the appendix.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **Degrees of Freedom** | **Pr > ChiSq** |
| **Likelihood Ratio** | 36.1376 | 3 | <0.0001 |
| **Score** | 29.0533 | 3 | <0.0001 |
| **Wald** | 25.0028 | 3 | <0.0001 |

**Table 5.1. Testing Global Null Hypothesis: BETA=0**

Significance of each predictor

Ho: i= 0 vs. Ha: i**0** where i = 1,2,3,...13

|  |  |  |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| AIC | 341.442 | 311.304 |
| SC | 344.943 | 325.309 |
| -2 Log L | 339.442 | 303.304 |

**Table 5.2. Model Fit Statistics**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criterion** | **Value** | **DF** | **Value/DF** | **Pr > ChiSq** |
| Deviance | 303.3045 | 238 | 1.2744 | 0.0027 |
| Pearson | 247.6860 | 238 | 1.0407 | 0.3197 |

**Table 5.3. Goodness-of-Fit Statistics**

**Interpretation of computer output**

**Linear Regression**

**Model 1:**

Based on Figure 1.1. Analysis of Variance Analysis of Variance test reported F-value is 12.75 and p-value of <.0001 and at 5% significance level we fail to accept the null hypothesis and we conclude the fitted linear regression model is overall significant

**The constancy of variance assumption**

Based on Residual by Predicted for imdb\_score **figure 1.2**, the constancy of variance is not perfect to satisfy the model assumption.  Hence the Fitted model is not good.

**The normality of the error term assumption**

Based on Q-Q plot in **figure 1.3** of Residuals normality assumption has been violated as the residual doesn’t seem to follow the straight-line pattern.  Hence the Fitted model is not good.

Based on Goodness-of-Fit test **(Figure 1.4)** at 10 % significance level except Kolmogorov-Smirnov Test both Cramer-von Mises and Anderson-Darling has reported p-value lower than significance level 10%. Hence, we fail to accept the null hypothesis and conclude residuals are not normally distributed.

**Model 2:**

Based on Figure 2.1. Analysis of Variance table reported F-value is 14.16 and p-value of <.0001 and at 5% significance level we fail to accept the null hypothesis and we conclude the fitted linear regression model is overall significant

**The constancy of variance assumption**

Based on the Residual by Predicted for imdb\_score (**figure** **2.2**), constancy of variance is satisfied as the residual spread seems constant along the graph.  Hence the Fitted model is good.

**The normality of the error term assumption**

Based a Q-Q plot of Residuals (**figure** **2.3)**, the assumption of normality does seem to follow the straight-line pattern and satisfy the normality assumption.  Hence the Fitted model is good.

The Goodness-of-Fit **Figure 3.5** test shows at 10 % significance level reported p-value higher than significance level. Hence, we accept the null hypothesis and conclude residuals are normally distributed.

**Model 3:**

Based on Figure 3.2 Analysis of Variance table report F-value=24.73 and p-value <.0001 at significance level 5% we fail to accept the null hypothesis and we conclude that the model is overall significant.

**The constancy of variance assumption**

Based on the Residual by Predicted for imdb\_score (**Figure 3.3**), constancy of variance is satisfied as the residual spread seems constant along the graph.  Hence the Reduced Fitted model is good.

**The normality of the error term assumption**

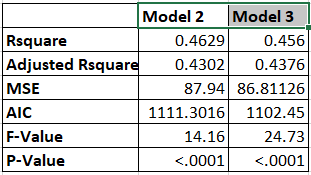
Based on a Q-Q plot of the residuals (**Figure 3.4**), the assumption of normality for the residuals does seem to follow the straight-line pattern and satisfy the normality assumption.  Hence the Reduced Fitted model is good.

The Goodness-of-Fit **Figure 2.2** test shows at 10 % significance level reported p-value higher than significance level. Hence, we accept the null hypothesis and conclude residuals are normally distributed.

**Overall comparison of Model 2 and 3**

Model 2 ~ Transformed Full Linear Model

Model 3 ~ Reduced Transformed Linear Model

Based on the table MSE value for Model 3 which is a reduced model is lesser than full model justifying that reduced fitted model is better than Model 2. Comparison of Adjusted R Square shows Model 3 can explain the variance of response variable better than Model2. Comparison of AIC value shows lesser the AIC value better the Model in favor of Model 3. For hypothesis test the F-value is higher for reduced model and p-value is <0.0001 at 10% significance level giving us more evidence that the model is overall significance of model is better.

For IMDB Score prediction reported in Model 3 **Figure 3.6** number of critics, duration, budget, Country of production USA and UK, Content Rating PG, PG13 and R are modest significant predictor at 10% significance level.

**Logistic Regression: Model 4**

As shown in table 4.1, the Likelihood Ratio, Score, and Wald test produce p-values <0.0001, <0.0001, and 0.0054 respectively. At a significance level of 5%, the three tests support the alternative hypothesis because the p-values are small, which means that the logistic regression is overall significant.

When testing for the significance of each predictor, Wald test gives p-values 0.0244, 0.2124, 0.4668, <0.0001, 0.8312, 0.3782, 0.9651, 0.9689, 0.9664, 0.9970, 0.0070, 0.0855, 0.9285 for number of user reviews, number of critics, duration of movie, budget, number of cast Facebook likes, number of director Facebook likes, USA country of origin, France country of origin, UK country of original, Canada country of origin, content rating G, content rating PG, content rating PG13, respectively (as shown in Figure 4.2). At a significance level of 10%, the number of user reviews, budget, content rating PG and content rating PG13 were significant predictors because of the small p-values supporting Ha. At a significant level of 5%, only the number of user reviews, budget, and content PG are significant because of the small p-values which support the alternative hypothesis.

**Reduced Logistic Regression: Model 5**

-An attempt was made to add interactions of the model, however, with the addition of variable interactions all interactions and variables were insignificant at any reasonable significance level.

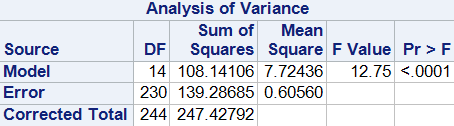
The Likelihood Ratio, Score, and Wald test produce p-values <0.0001, <0.0001, and <0.0001 respectively. At a significance level of 5%, the three tests support the alternative hypothesis because the p-values are small, which means that the three tests report that this logistic regression is overall significant.

Wald test gives p-values <.0001, 0.0044, and 0.0062 for number of user reviews, budget, USA country of origin, respectively. At a significance level of 5%, the number of user reviews, budget, and the USA as a country of origin were significant predictors because of the small p-values supporting Ha.

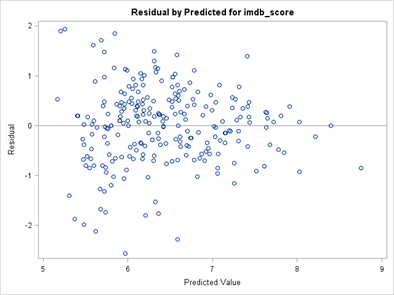
It is determined that the reduced model is preferred over the full model, as it maintains similar model fit statistics (**5.2,4.2**) while reducing the number of independent variables from 14 down to 3.  Furthermore, a deviance test may be conducted to determine goodness-of-fit if necessary.

**Conclusion:** Hence we can conclude that after reduced fitted linear regression model only eight variables are significant factors contributing in prediction of IMBD\_score, giving us the best fitted model with 10% of significance level. As seen in figure **3.7** we have predicted some IMBD\_score giving us a close prediction to its final predicted value proving that it is the best fitted model that we can use to predict the IMBD\_score.  For the logistic regression on a binary variable success/failure, 14 variables were introduced in the full-fitted model. Of the variables introduced, only 4 were significant at a significance level of 10%, **see figure 4.2**.  Using backwards-manual elimination, the full logistic regression was reduced to 3 independent variables, all of which were significant at a significance level of 5%.  Based on model-fit statistics (**4.2,5.2**), it is determined that the reduced model is preferred to the full-fitted model.  Overall, the only variables significant in either reduced model were: budget (x4), and USA origin (x7).  From the analysis, we can determine that the success of a movie is majorly dependent on the budget, as you have to generate more revenue in order to be profitable, and the USA as the country of origin (significantly more movie goers than other countries).

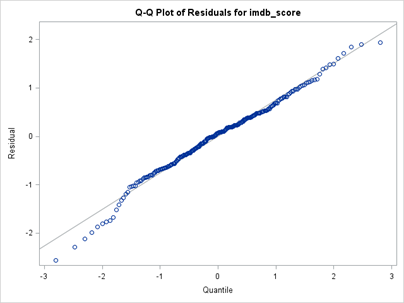
**Appendix**



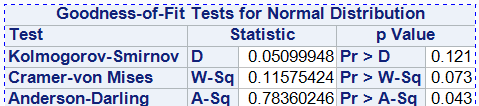
**Figure 1.1. Analysis of Variance**



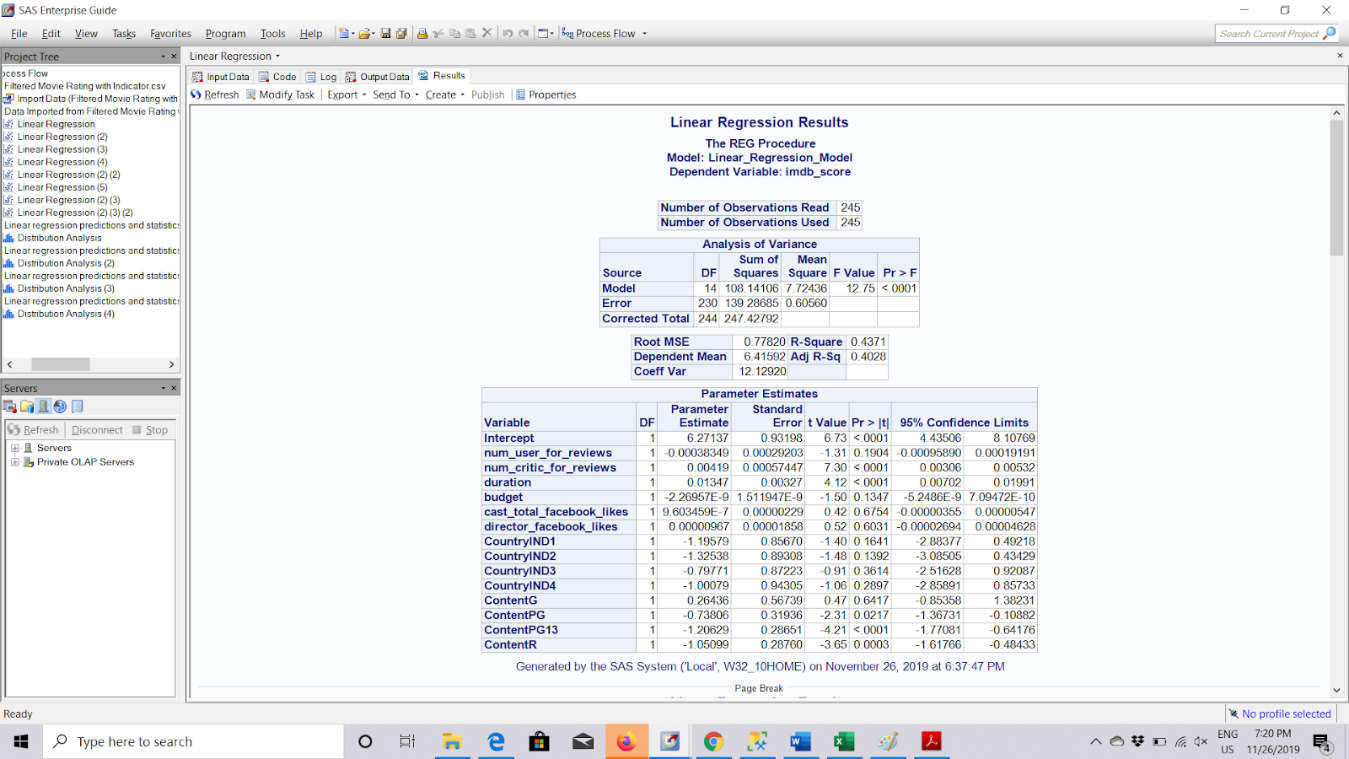
**Figure 1.2. Residuals by the predicted value for IMDB score.**



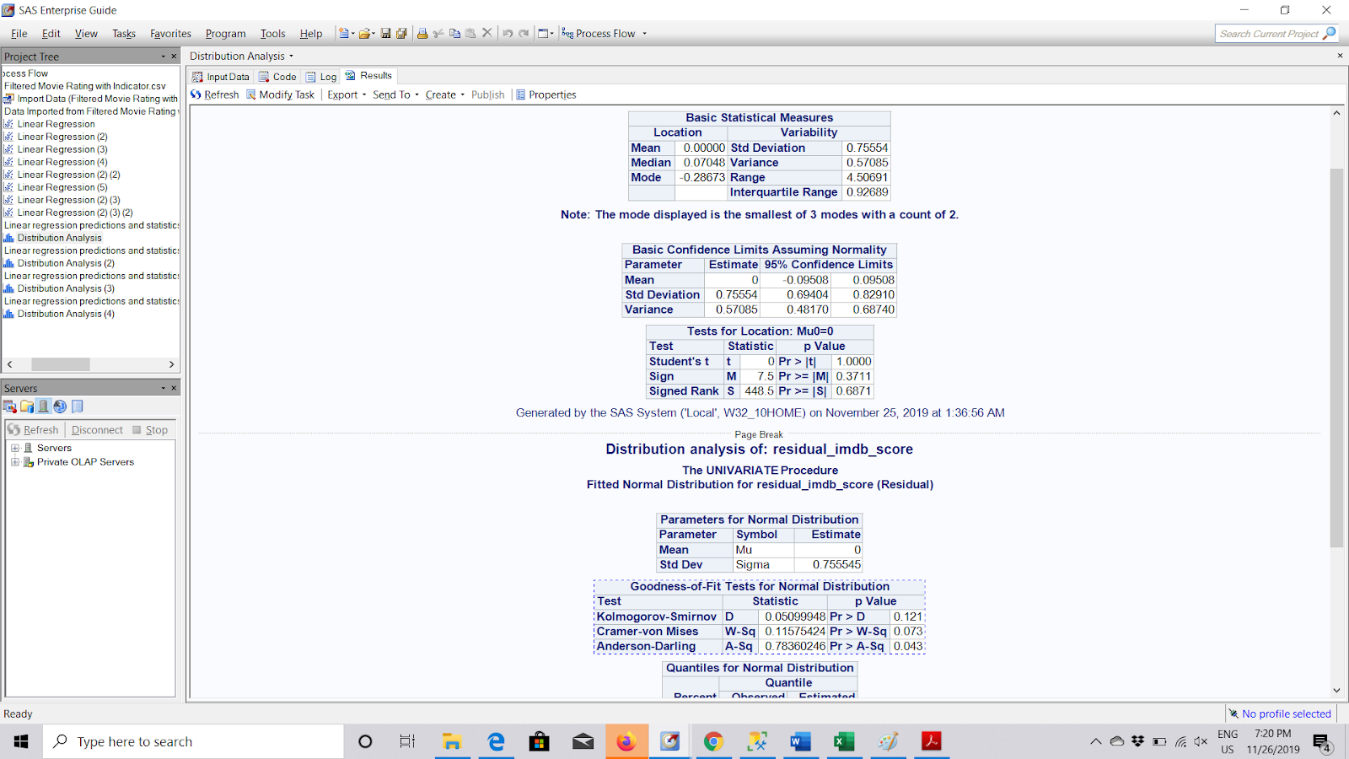
**Figure 1.3. Q-Q Plot of Residuals for IMDM score.**



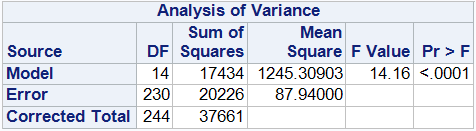
**Figure 1.4. Goodness-of-Fit Test for Normal Distribution.**

****

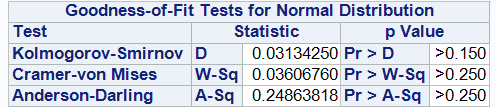
**Figure 1.5. Result of First Linear Regression Model**

****

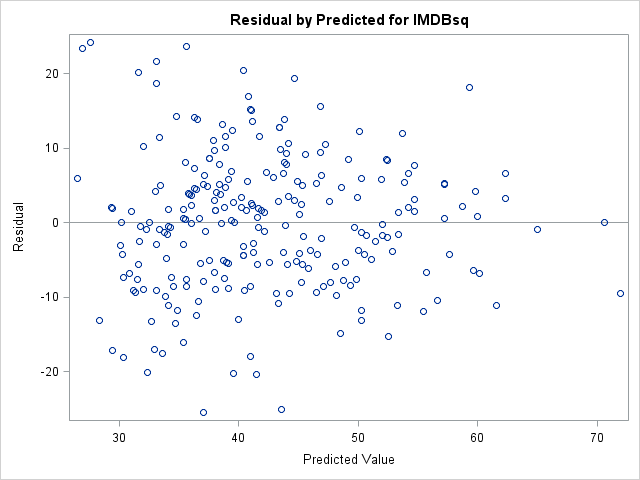
**Figure 1.6. Goodness of Fit test Results for normality of Residuals for First Linear Regression Model**



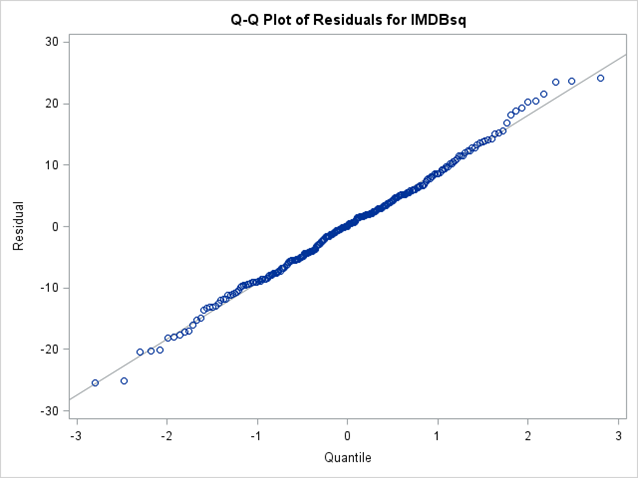
**Figure 2.1. Analysis of Variance**



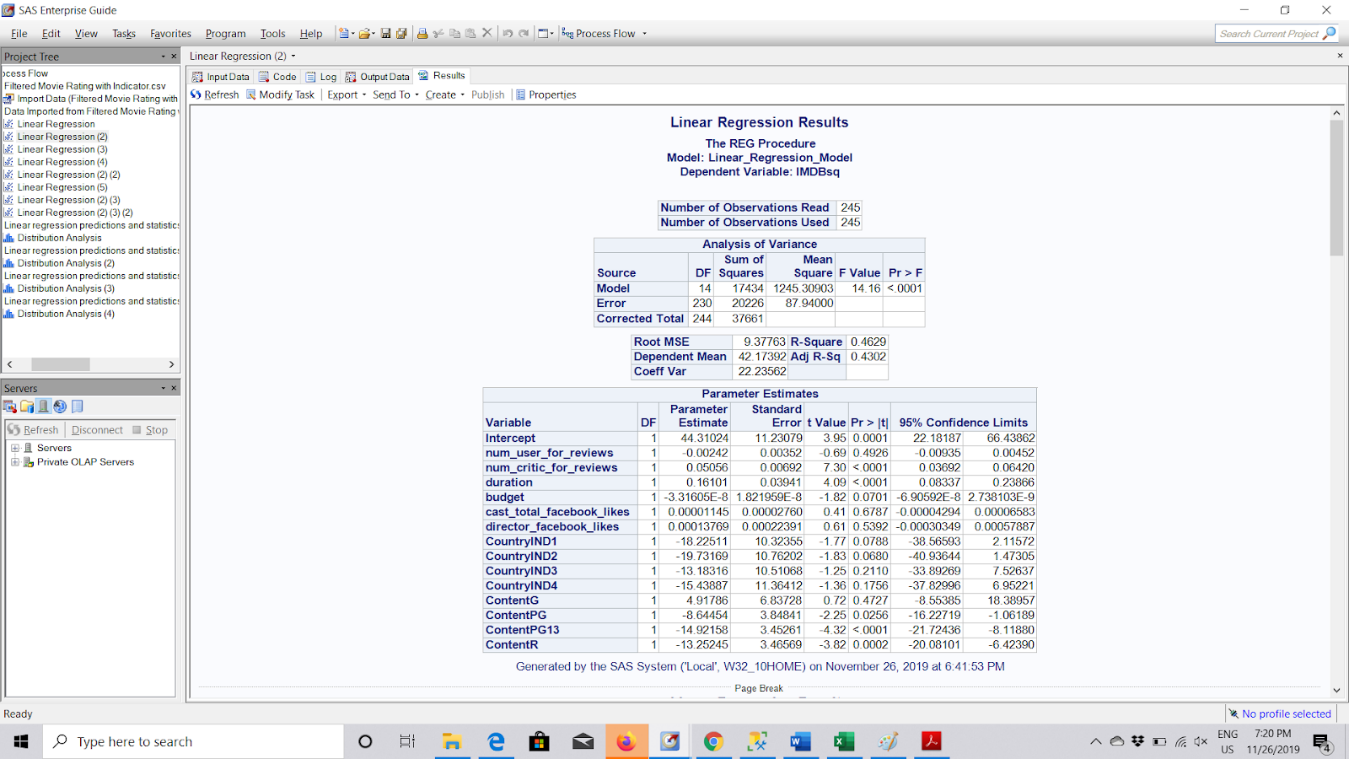
**Figure 2.2. Goodness-of-Fit Test for Normal Distribution.**



**Figure 2.2. Residuals by the predicted value for  IMDB score.**

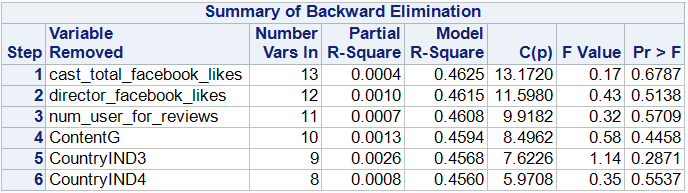


**Figure 2.3. Q-Q Plot of Residuals for IMDM score.**

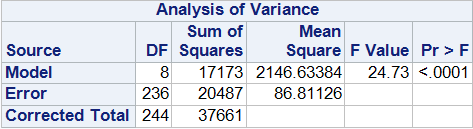
****

**Figure 2.4. Result of Transformed Linear Regression Model with square of Imdb\_score as response variable.**

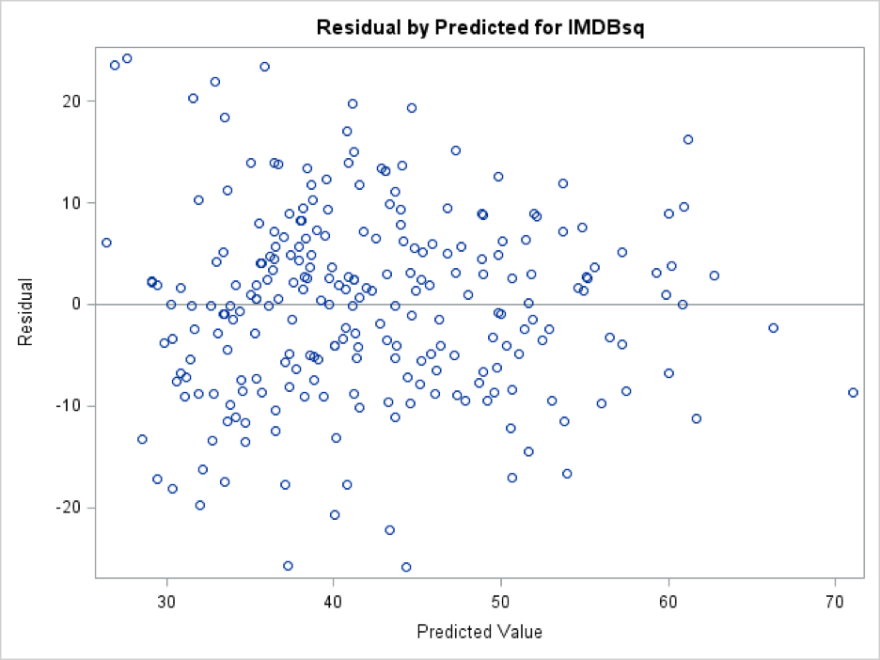
**Figure 2.5. Goodness of Fit test Results for normality of Residuals for Transformed Linear Regression Model**



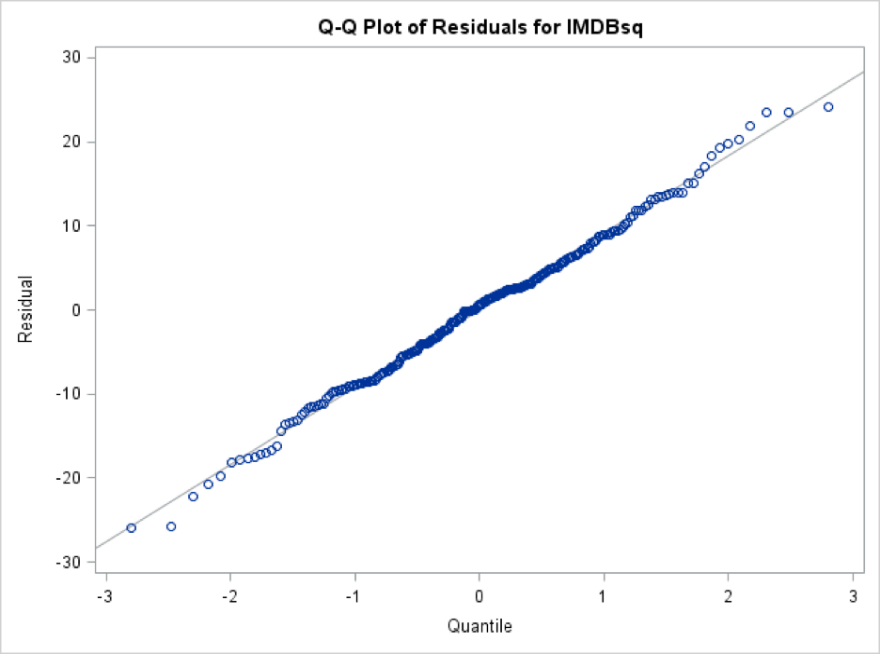
**Figure 3.1. Backwards Elimination Summary**



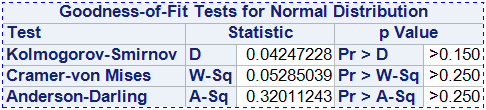
**Figure 3.2. Analysis of Variance**



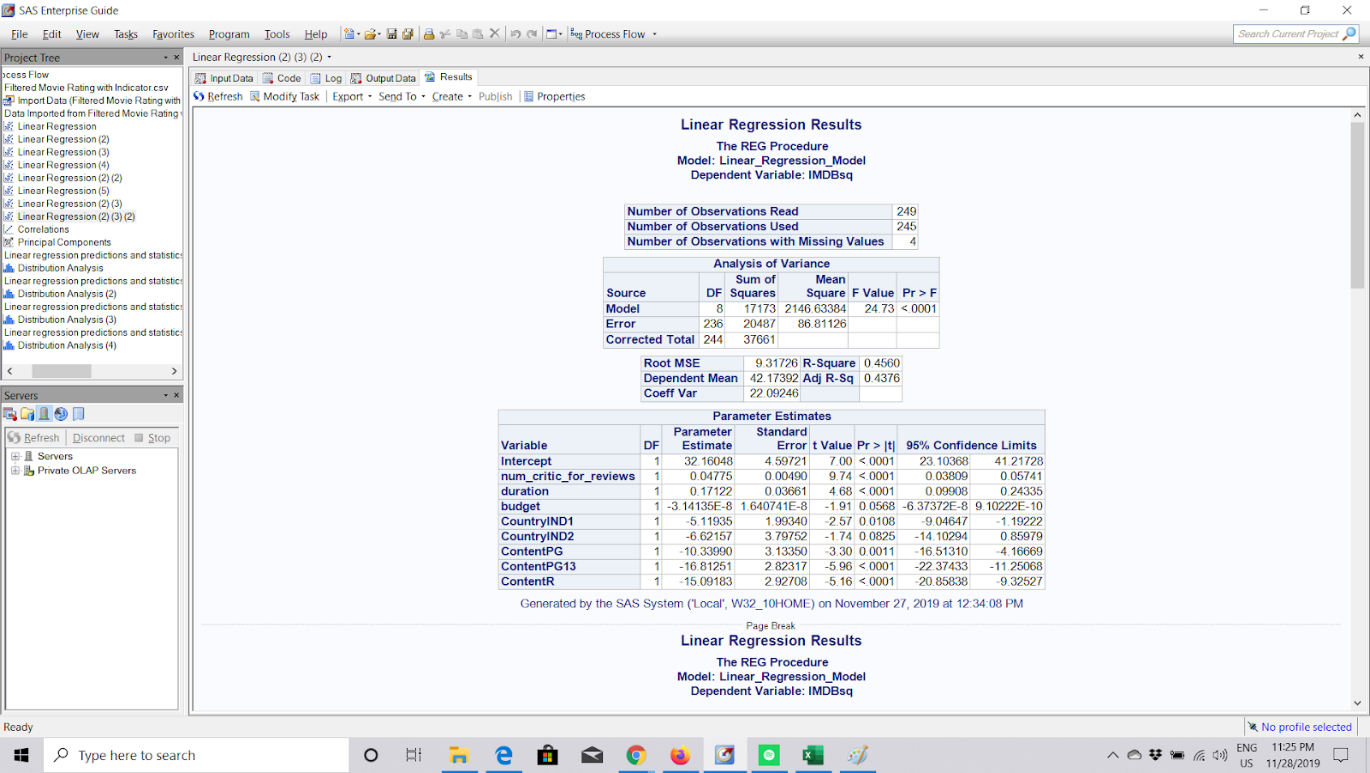
**Figure 3.3. Residuals by predicted values for IMDB score**



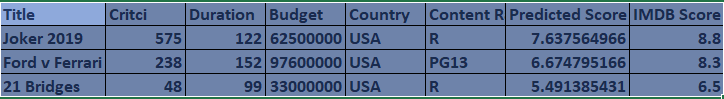
**Figure 3.4. Q-Q Plot of residuals for the IMDB score**



**Figure 3.5. Goodness-of-Fit Tests for Normal Distribution**

****

**Figure 3.6. Result of Reduced Transformed Linear Regression Model with square of Imdb\_score as response variable.**

****

**Figure 3.7. Some of predicted IMBD\_score using reduced transformed Linear Regression Model.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **Degrees of Freedom** | **Pr > ChiSq** |
| **Likelihood Ratio** | 56.124 | 14 | <0.0001 |
| **Score** | 46.489 | 14 | <0.0001 |
| **Wald** | 31.099 | 14 | 0.0054 |

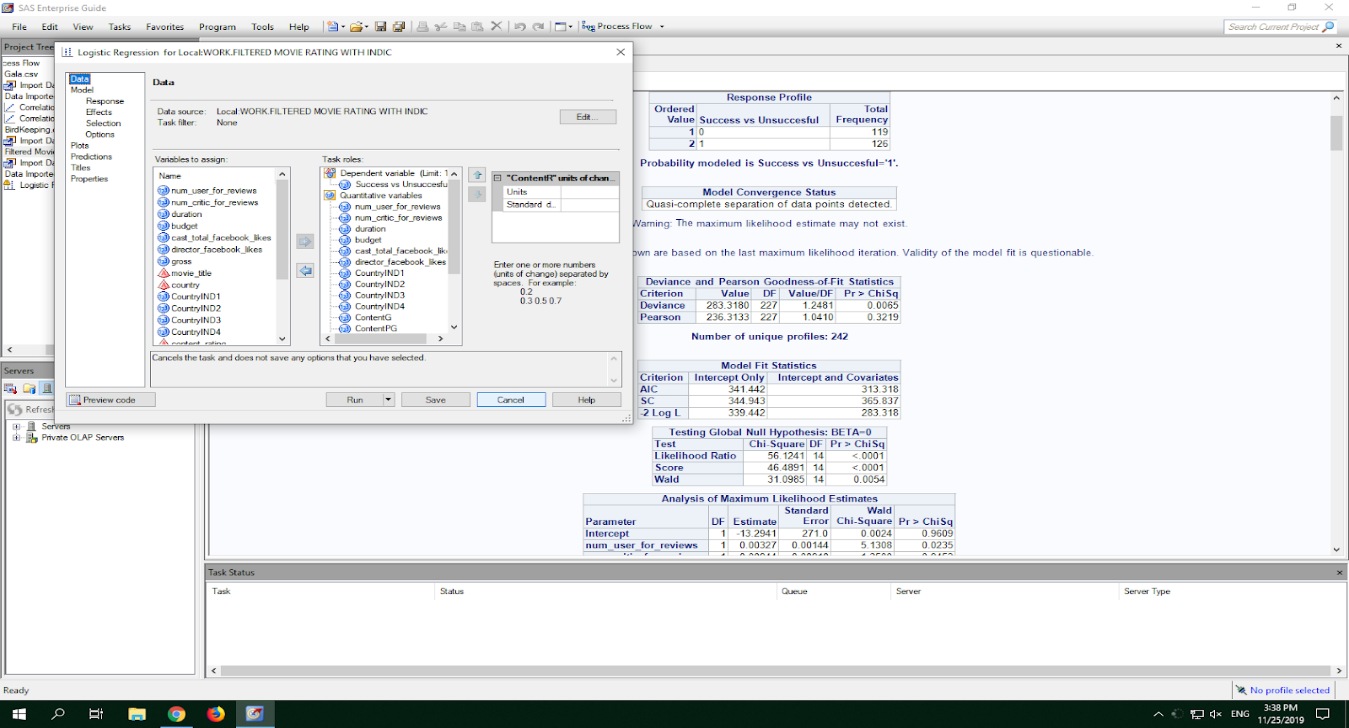
**Table 4.1. Testing Global Null Hypothesis: BETA=0**

|  |  |  |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| AIC | 341.442 | 313.318 |
| SC | 344.943 | 365.837 |
| -2 Log L | 339.442 | 283.318 |

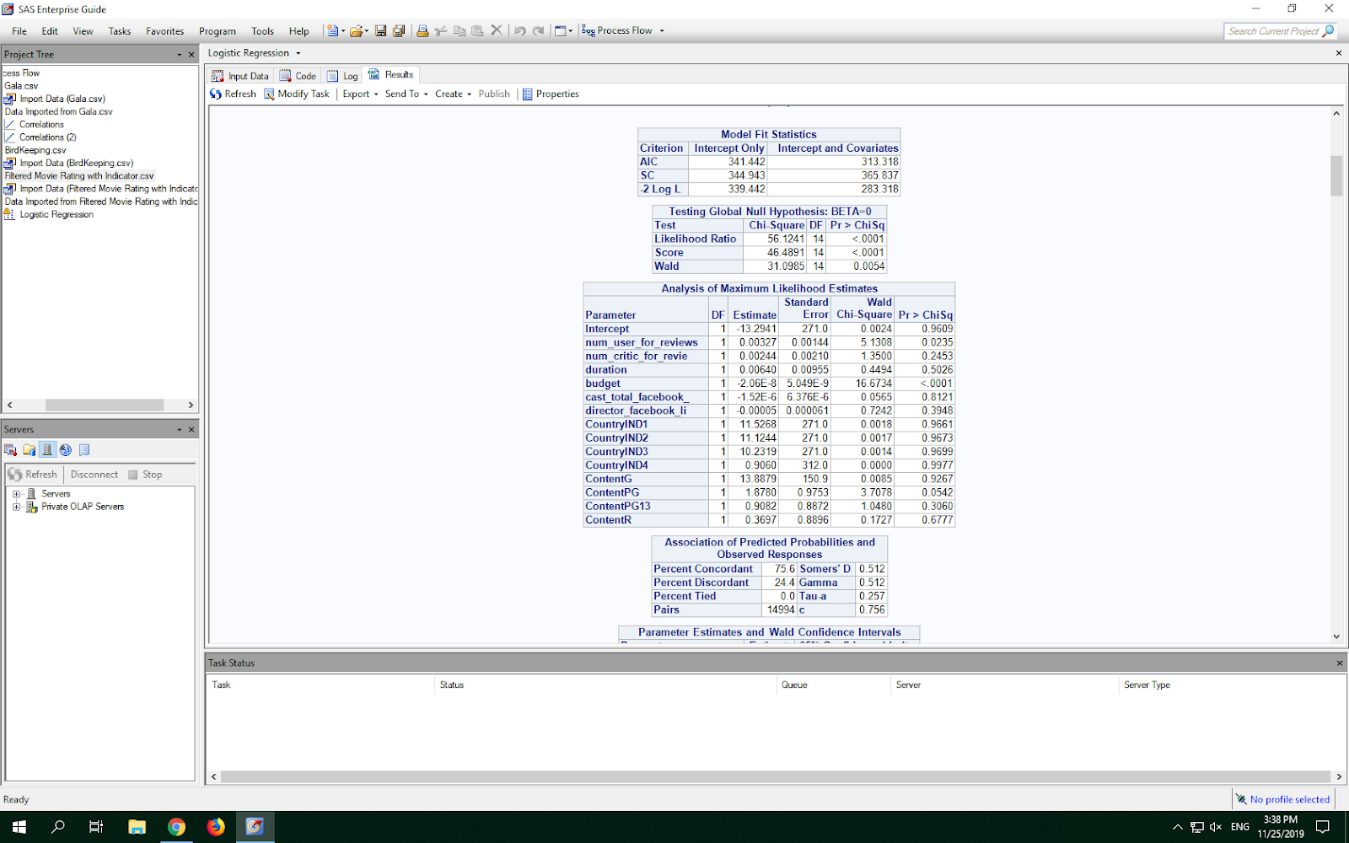
**Table 4.2. Model Fit Statistics**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criterion** | **Value** | **DF** | **Value/DF** | **Pr > ChiSq** |
| Deviance | 283.3180 | 227 | 1.2481 | 0.0065 |
| Pearson | 236.3133 | 227 | 1.041 | 0..3219 |

**Table 4.3. Goodness-of-Fit Statistics**

****

**Figure 4.1. Computing the variables into SAS for Logistic Regression**

****

**Figure 4.2. Analysis of Maximum Likelihood Estimates for Model #4.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **Degrees of Freedom** | **Pr > ChiSq** |
| **Likelihood Ratio** | 36.1376 | 3 | <0.0001 |
| **Score** | 29.0533 | 3 | <0.0001 |
| **Wald** | 25.0028 | 3 | <0.0001 |

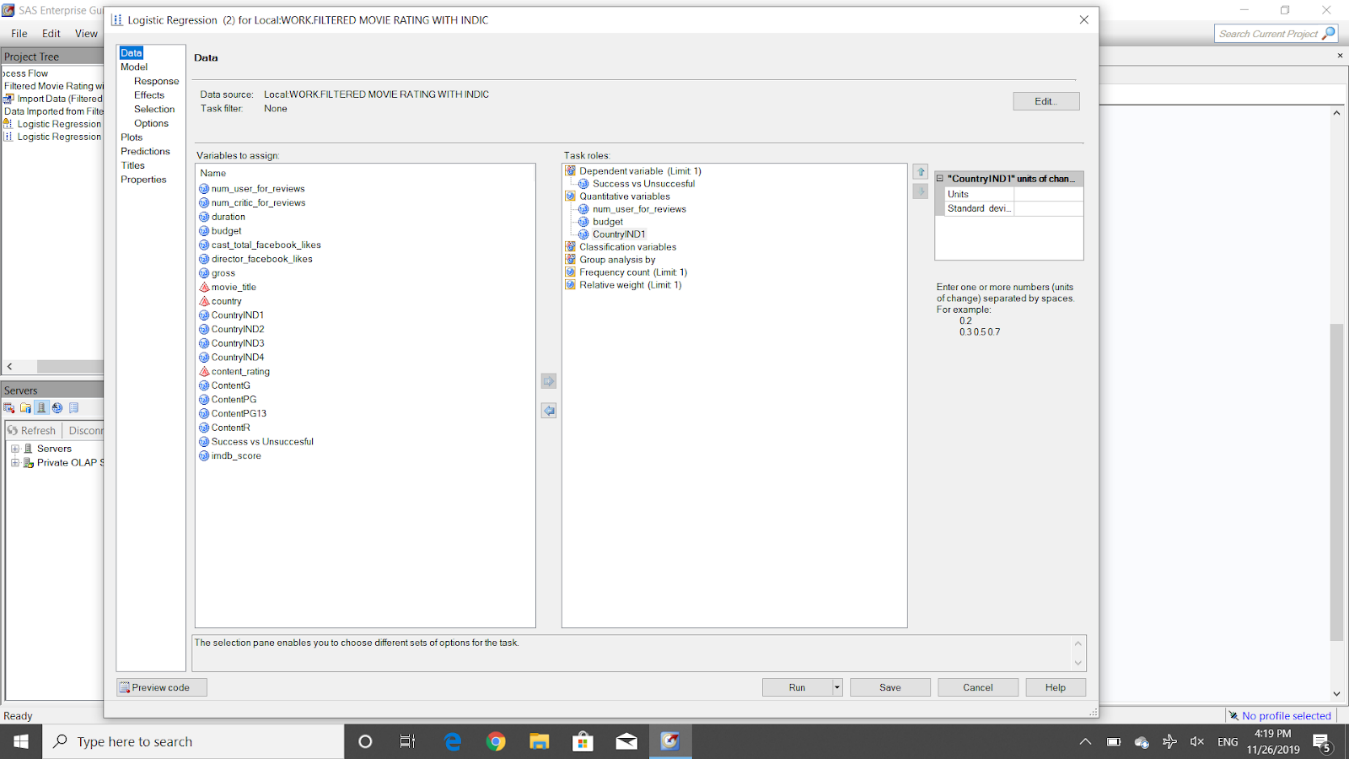
**Table 5.1. Testing Global Null Hypothesis: BETA=0**

|  |  |  |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| AIC | 341.442 | 311.304 |
| SC | 344.943 | 325.309 |
| -2 Log L | 339.442 | 303.304 |

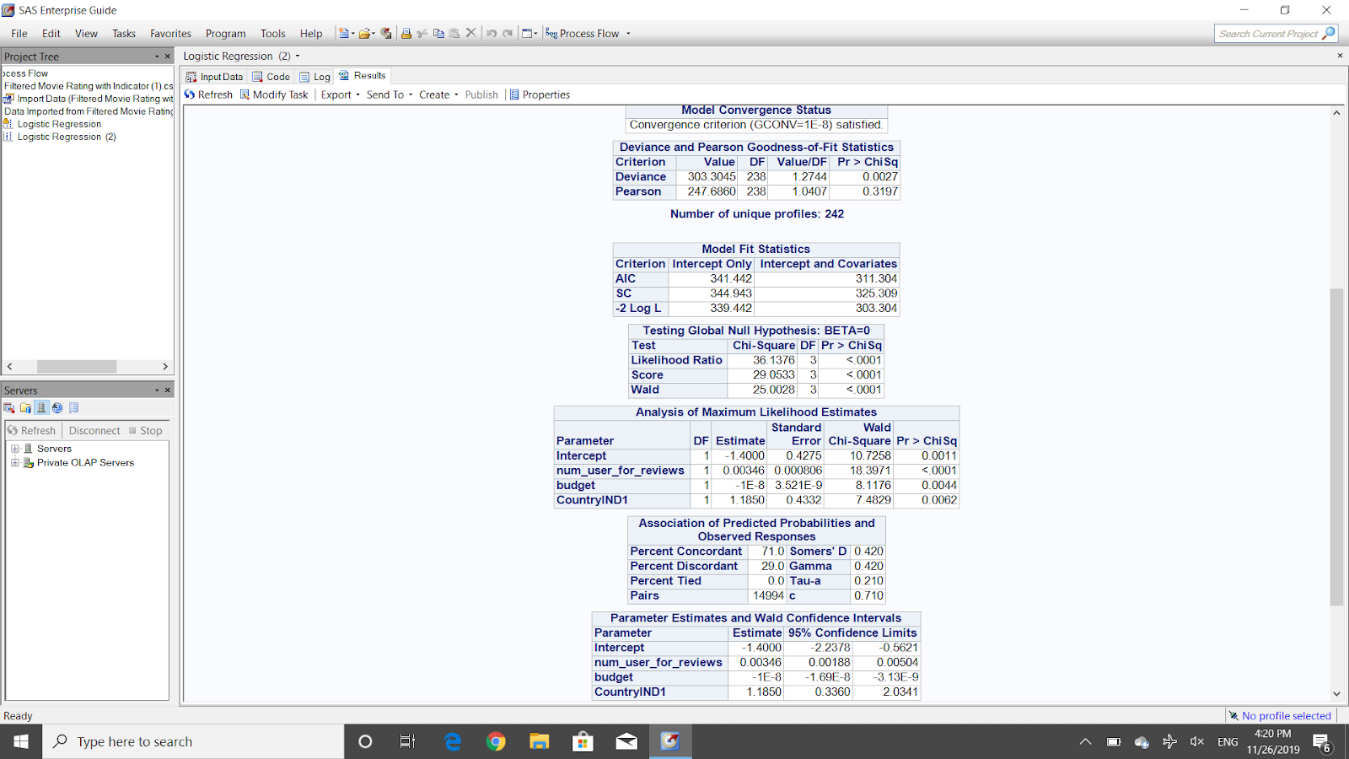
**Table 5.2. Model Fit Statistics**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criterion** | **Value** | **DF** | **Value/DF** | **Pr > ChiSq** |
| Deviance | 303.3045 | 238 | 1.2744 | 0.0027 |
| Pearson | 247.6860 | 238 | 1.0407 | 0.3197 |

**Table 5.3. Goodness-of-Fit Statistics**

****

**Figure 5.4. Computing the variables into SAS for Reduced Logistic Regression**

****

**Figure 5.5. Analysis of Maximum Likelihood Estimates for Model #5, and relative data.**