Multiple Regression Computing

**HuYuDataInsight LLC**

Zhaowei Cai April 25, 2024

**1.Introduction**

We will use the methods in multiple linear regression to find out the model for the dataset. By reading the data from data1.csv, we can find that there are totally 1001 observations and 39 variables. Y is the outcome variable, E1 to E8 are the eight environment variables, and G1 to G30 are the thirty genetic independent indicator variables.

It is obvious that the model is related to the multiple linear regression, and the function should include the interactions of some variables1. Firstly, we will do an analysis of multiple linear regression of Y and all other 38 variables in R. Our goal is trying to select the important variables. Secondly, we will find optimal lambda for Box-Cox transformation of our linear model to let the dataset normally distributed2. Thirdly, we will consider the interactions of variables and perform stepwise multiple regression to select the significant variables that generate our dataset3. Lastly, we will perform an anova test to get the anova table and the result of our model.

**2.Methods**

We perform the linear model of Y and all other variables in R and name it as “model1”. We show the summary of model1 and select **E1 E2 E5 G3 G13 G24 G28** as the important variables. According to the table of Coefficients, the p-values of these selected variables are all smaller than 0.01. We also perform the pairwise correlation test of all the variables by Pearson method4. The result shows that the correlations between **Y&E1**, **Y&E2**, **Y&E5**, **Y&G3**, **Y&G13**, **Y&G24**, **Y&G28** are 0.40, 0.58, 0.02, 0.23, 0.17, 0.16, 0.19.

To find the optimal lambda for Box-Cox transformation of model1, we set the range of lambda to be -2 to 2 with an increment of 1/10. When maximum of y happens, the value of lambda is 0.9494949, so the optimal lambda can be approximately taken as 1. When lambda is equal to 1, the transformation function is:

Then, we perform the linear regression in R by using at most four-way interactions of selected independent variables **E1 E2 E5 G3 G13 G24 G28**. We name this linear model as “full.model”. Then we perform the stepwise multiple regression of full.model in both directions. The result of stepwise regression shows that these variables are significant: **E1 E2 E5 G3 E1\*E2 E1\*E5 E2\*E5 E2\*G13 G3\*G13 G13\*G24 E1\*E2\*E5 E1\*E2\*G13 E5\*G3\*G13 E5\*G3\*G13\*G24**.

Then, we perform the linear regression by using the significant variables above. We name this linear model as “full.model2”. Then we perform the stepwise multiple regression of full.model2 in both directions. The result shows that these variables are significant: **E1 E2 E5 G3 E1\*E2 E1\*E5 E2\*E5 G3\*G13 G13\*G24 E1\*E2\*E5 E5\*G3\*G13**. However, the p-values of there variables are not all smaller than 0.01, so we drop the variables of which the p-values are larger than 0.01. Then our selected important variables are: **E1 E2 G3 G13\*G24 E5\*G3\*G13**.

Then, we continue to perform the linear regression by using the selected variables above and name it as “full.model3”. We perform the stepwise multiple regression of full.model3 in both directions. The result shows that the significant variables are: **E1 E2 G3 G13\*G24**, The p-values of Intercept and these variables are all smaller than 2\*exp(-16), and thus smaller than 0.01. It meets the requirements of project, so we stop performing the stepwise regression again.

Finally, we test our model by anova method and get the analysis of variance table.

**3.Results**

The result of Box-Cox transformation is shown in Figure3.1, we can find that the optimal value of lambda is approximately 1. We can also find that the data is normally distributed after transformation.

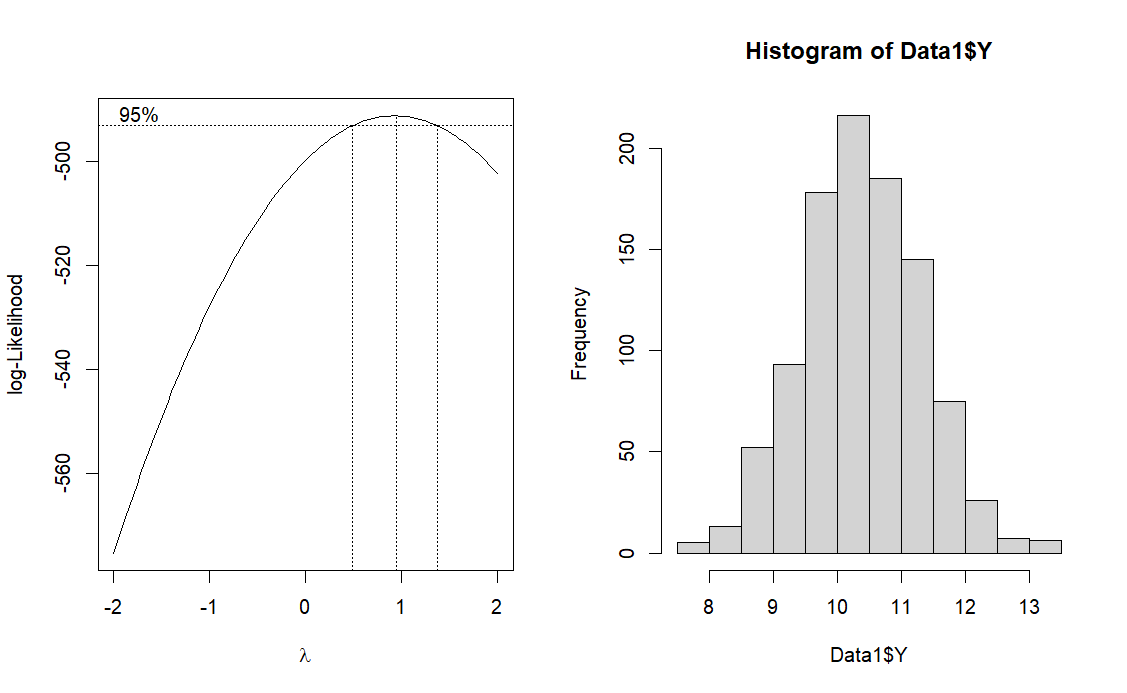


Figure3.1 Plot of Box-Cox Transformation

The result of stepwise multiple regression of full.model3 is shown in Table3.1. We can see that the final significant variables that we select for our model are **E1 E2 G3 G13\*G24**. The p-values of all these variables, including Intercept, are smaller than 0.01. The residual standard error is 0.5482 on 996 degrees of freedom. The value of multiple R-squared is 0.6428, the value of adjusted R-squared is 0.6413.

Table3.1 Result of Stepwise Regression

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Residuals | Min | 1Q | Median | 3Q | Max |
|  | -1.86253 | -0.37659 | 0.01102 | 0.36064 | 2.18949 |
| Coefficients: | Estimate | Std. Error | t value | Pr(>|t|) |  |
| Intercept | 5.26490 | 0.13294 | 39.60 | <2e-16 | \*\*\* |
| E1 | 0.27583 | 0.01204 | 22.90 | <2e-16 | \*\*\* |
| E2 | 0.38041 | 0.01205 | 31.58 | <2e-16 | \*\*\* |
| G3 | 0.44025 | 0.03777 | 11.66 | <2e-16 | \*\*\* |
| G13\*G24 | 0.78348 | 0.05978 | 13.11 | <2e-16 | \*\*\* |

The formula of our model is as follows. The estimated values of are 5.26490, 0.27583, 0.38041, 0.44025, 0.78348.

We also plot the results of stepwise multiple regression, including the Residuals vs Fitted values, Normal Q-Q plot, Scale-Location, Residuals vs Leverage. Please see the details in Figure3.2.

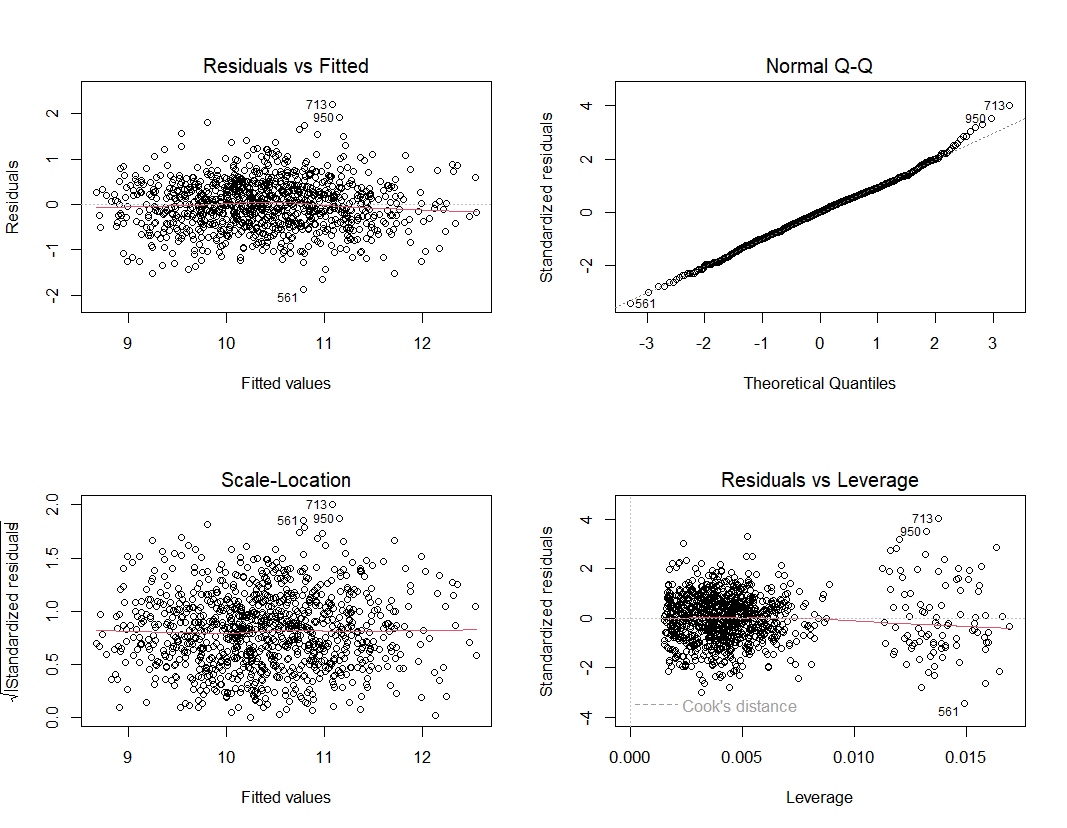


Figure3.2 Plot of Stepwise Multiple Regression

From the Residuals vs Fitted plot and Scale-Location plot, we can conclude that the residuals are heteroskedastic. From Normal Q-Q plot, we find that most of the points are on the red line, so it is normally distributed. From the Residuals vs Leverage plot, there are no outliers.

The result of analysis of variance table is shown in Table3.2.

Table3.2 Analysis of Variance Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| E1 | 1 | 133.483 | 133.483 | 444.20 | <2.2e-16 |
| E2 | 1 | 311.143 | 311.143 | 1035.42 | <2.2e-16 |
| G3 | 1 | 42.288 | 42.288 | 140.73 | <2.2e-16 |
| G13\*G24 | 1 | 51.615 | 51.615 | 171.76 | <2.2e-16 |
| Residuals | 996 | 299.298 | 0.300 |  |  |

We can find that the p-values of selected variables are much smaller than 0.01, which meets the requirements of project.

**4.Discussion and Conclusion**

We use the method of stepwise regression to select the important variables of the dependent outcome variable Y, the selected variables are independent and from a given dataset. However, there are still some limitations of my procedures.

The result of selecting variables may vary from different samples of dataset that we use. Therefore, the fit models are also different. In addition, overfitting is a possible result during stepwise regression5. If we dive into the fundamentals of stepwise regression, we will know that some variables that are not relevant to Y may be selected to improve the fit. Another limitation is that the estimated values of parameters may be biased when we perform stepwise regression6.

**References**

[1] Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. Practical assessment, research, and evaluation, 8(1), 2.

[2] Atkinson, A. C., Riani, M., & Corbellini, A. (2021). The box–cox transformation: Review and extensions.

[3] Johnsson, T. (1992). A procedure for stepwise regression analysis. Statistical Papers, 33(1), 21-29.

[4] Zhou, H., Deng, Z., Xia, Y., & Fu, M. (2016). A new sampling method in particle filter based on Pearson correlation coefficient. Neurocomputing, 216, 208-215.

[5] Coolen, A. C. C., Sheikh, M., Mozeika, A., Aguirre-Lopez, F., & Antenucci, F. (2020). Replica analysis of overfitting in generalized linear regression models. Journal of Physics A: Mathematical and Theoretical, 53(36), 365001.

[6] Gunst, R. F., & Mason, R. L. (1977). Biased estimation in regression: an evaluation using mean squared error. Journal of the American Statistical Association, 72(359), 616-628.