AMS 315 Project 2

Zian Shang

Instructor: Benjamin Hechtman

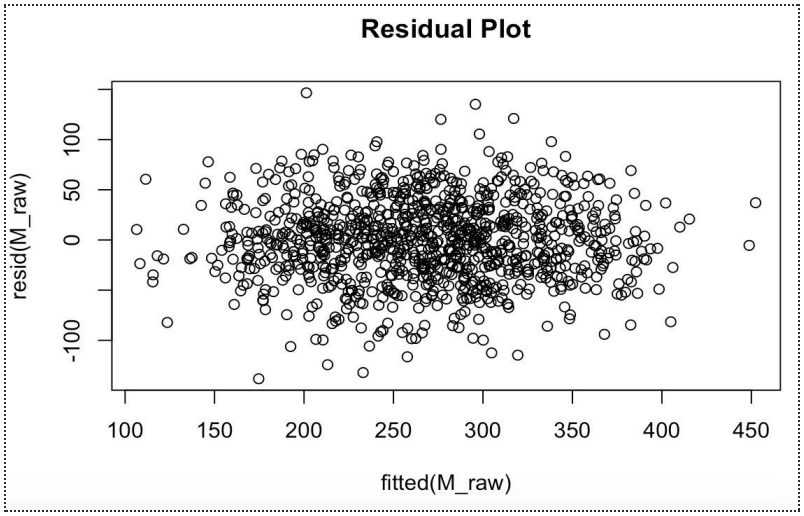
11/22/2022

Introduction

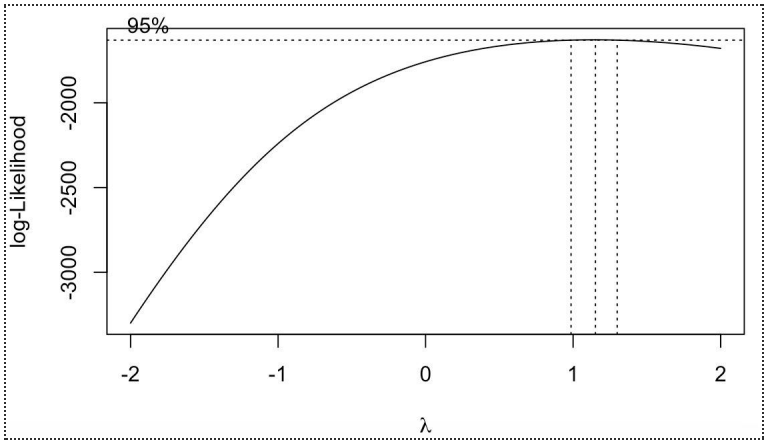
In the study of Caspi et al., they tested why stressful experiences lead to depression in some people but not in others. They found that a functional polymorphism in the promoter region of the 5-HT T moderates the influence of stressful life events on depression. Individuals with a short allele of the 5-HT T promoter polymorphism exhibited more depressive symptoms in relation to stressful life events than individuals with a long allele. Their paper used multiple regression techniques as the methodology for its findings. In this project, we are given a synthetic dataset with one independent variable Y, 4 environmental variables E, and 20 genetic variables G. Our goal is to estimate the function that the TA used to generate our data, just like Caspi et al. did their study to find their special gene-environment interaction.

Methodology

For this project, I decided to use R. First, I imported the file using the "read.csv()" function and stored the imported values in the variable "DataSet". After that, I fitted a model with just the environmental variables to explain the outcome using the "lm()" function and name it "M_E" (apdx 2). The summary of M_E gave me an initial R^2 value for comparison with further transformation values. Then, I assessed the contribution of the genetic variables by fitting all 24 independent variables into a linear regression model named "M_raw". I also created a residual plot for this raw model that is shown below:



To seek a transformation of the dependent variable that has homoscedastic residuals, I imported the library "MASS" to use its "boxcox(M_raw)" function. This gave me a graph that looks like this:



After careful measuring, I decide that the value of λ in my graph is a little bit smaller than 1+1/7, so I chose 1.14 as the transformation exponent of Y. I applied this transformation to Y by creating another linear regression model using the function "lm()" and assigned it to the variable "M_trans". I then plotted the new residual plot for "M_trans" to see if the data was displayed as a flat ellipse (apdx 3). I also used the "summary()$adj.r.square" function to generate the adjusted R^2 values for both M_raw and M_trans to compare with M_E.

Next, I loaded the R code leaps and used the function "regsubset()" to perform stepwise regression. With the function "kable()", I was able to produce some proposed models shown below:

Table: Model Summary

model	adjR2	BIC
(Intercept)+E3:E4	0.471725338145586	-627.869495115599
(Intercept)+E4+E3:E4	0.500272179251872	-677.73520481024
(Intercept)+E3+E4+E3:E4	0.518002367844189	-708.09584360806
(Intercept)+E3+E4+E3:E4+G14:G18	0.526174906328692	-719.357951031702
(Intercept)+E3+E4+E3:E4+G2:G12+G14:G18	0.533574463092544	-729.254537677062

I carefully calculated the differences between the adjusted R^2 and Bayesian Information Criterion(BIC) of these five models. According to this table, I decided to use the 5th model. To check if any main effect variables are missing, I copied the previous transformed linear regression model as "M_main" and drew another table with the command "kable(temp$coefficients[abs(temp$coefficients[,4]) <= 0.001,], caption='Sig Coefficients') ". It gave me a table of significant coefficients, shown below:

Table: Sig Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
E3	25.15261	1.385054	18.16002	0.000000
E4	37.26521	1.334828	27.91761	0.000000
G12	30.23456	8.531653	3.54381	0.000413

It showed that the variables E3, E4, and G12 are the main effect, so these three variables are likely to be included in the final model. I also produced a second-stage model to examine whether there is a significant pairwise interaction. I produced the model with function “lm()” and assigned it to “M_2stage”. Then, I drew the table using the following command “kable(temp$coefficients[abs(temp$coefficients[,3]) >= 4,])
”:

	Estimate	Std. Error	t value	Pr(>|t|)
E3	21.71312	5.021898	4.323688	1.69e-05
E4	33.25177	4.926518	6.749549	0.00e+00

The p-values of both second-stage variables are all significantly smaller than 1, which means there are no strong pairwise interactions between them. Thus, there does exist a plausible final model. Therefore, I created my final model with E3, E4, and G12 with the “lm()” function and summarized it to get the real function(apdx 4). At last, I drew the ANOVA table for my final model (apdx 5).

Results

The initial adjusted R^2 using only the environmental variables is 0.5308. The adjusted R^2 of the raw model using both the environmental and genetic variables is 0.5373. Based on the graph of log-likelihood, I chose the transformation of my model to be Y^1.14, and the R^2 with this transformation is 0.5387. We can see that there is an apparent increase in the R^2 value after transformation. Then, the table of significant coefficients showed that the three variables, E3, E4, and G12, are the main effects that are significant to the model, so they are likely to be in the true model. The second stage table proved that there is no significant interaction between variables. Therefore, based on the summary, in the end, my final model is Y^1.14 = -45.162 + 24.916E3 + 37.497E4 + 31.903G12 and the adjusted R^2 is 0.5371.

Conclusion and discussion

In conclusion, a final model for the given dataset is found using R. There is no pairwise interaction between the genetic variables since they do not appear in the second stage table. There are some limitations in this approach to fitting a model to this dataset. The first is that both the initial and new residual plots can not show the shape of a “flat ellipse” well. My initial plot consisting of only the experimental variables also looks like an ellipse. There is no way to tell which plot is more “flat-elliptical”. The second problem is that the graph of the log-likelihood does not mark the point where the line reaches the 95% CI itself, so I have to guess it to find a proper transformation. Also, when I tested which variables in model 5 should be included in the final model, the variable G14 has one star beside it. This means that there is a really small possibility that G14 affects the final model. G14 may be in the final model if using another method for analysis.

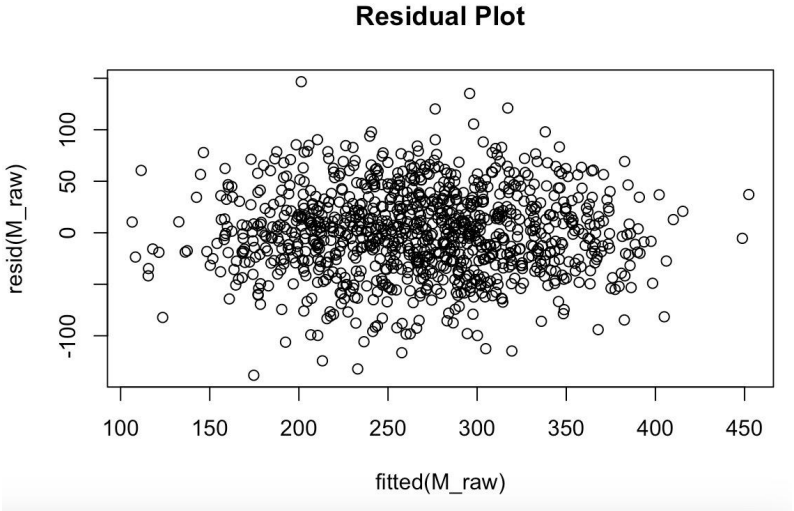
Appendix

1. source code

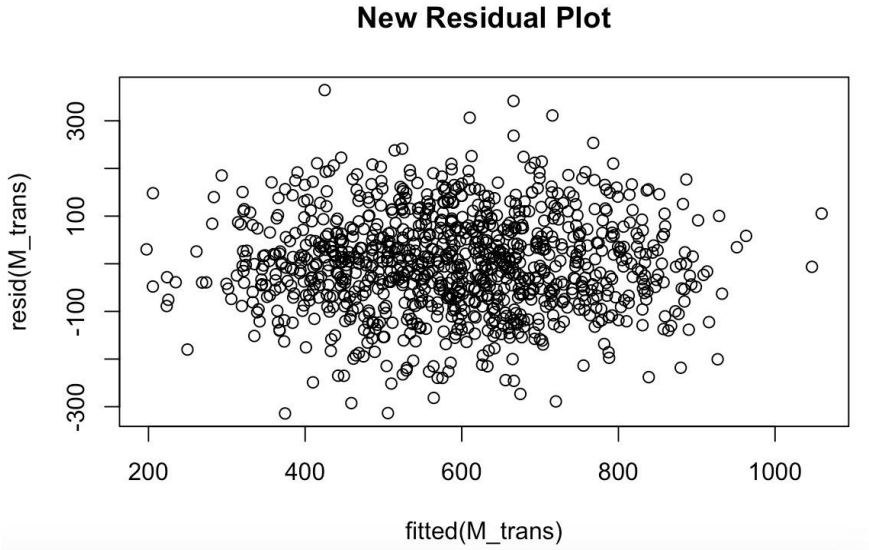
```
> DataSet <- read.csv('/Users/zianshang/Downloads/342000_project2.csv', header = TRUE)
> View(DataSet)
> # Fit a model with only the environmental variables
> M_E <- lm(Y ~ E1+E2+E3+E4, data=DataSet)
> View(M_E)
> summary(M_E)
> summary(M_E)$adj.r.squared
[1] 0.5308293
> M_raw <- lm(Y ~
(E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+G17+G18+G19+G20)^2,
data=DataSet)
> View(M_raw)
> # create and examine the residual plot
> plot(resid(M_raw) ~ fitted(M_raw), main='Residual Plot')
> # use the Box-Cox transformation to seek a transformation of the dependent variable that has apparently homoscedastic
residuals
> library(MASS)
> boxcox(M_raw)
> # apply approximate transformation 1.14 to response variable y
> M_trans <- lm( I(Y^1.14) ~
(E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+G17+G18+G19+G20)^2,
data=DataSet)
> View(M_trans)
> summary(M_raw)$adj.r.square
[1] 0.5373412
> summary(M_trans)$adj.r.square
[1] 0.5387447
> # new residual plot
> plot(resid(M_trans) ~ fitted(M_trans), main='New Residual Plot')
> # Stepwise Regression
> install.packages("leaps")
> library(leaps)
> M <- regsubsets( model.matrix(M_trans)[,-1], I((DataSet$Y)^.5), nbest = 1 , nvmax=5, method = 'forward', intercept = TRUE )
> View(M)
> temp <- summary(M)
> # produces a proposed model
> install.packages("knitr")
> library(knitr)
> Var <- colnames(model.matrix(M_trans))
> M_select <- apply(temp$which, 1, function(x) paste0(Var[x], collapse='+'))
> kable(data.frame(cbind( model = M_select, adjR2 = temp$adjr2, BIC = temp$bic)), caption='Model Summary')
> # make sure any main effects that are significant are in the model
> M_main <- lm( I(Y^1.14) ~
E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+G17+G18+G19+G20,
data=DataSet)
> temp <- summary(M_main)
> kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.001, ], caption='Sig Coefficients')
> # use the 5th function since E3 E4 E12 are all significant
> M_2stage <- lm( I(Y^1.14) ~ (E3+E4+G12)^2, data=DataSet)
> temp <- summary(M_2stage)
> kable(temp$coefficients[ abs(temp$coefficients[,3]) >= 4, ])
```

```
> # The lines of code below create and then provide the summary values for the chosen final model
> M_final <- lm(I(Y^1.14) ~ E3 + E4 + G12, data = DataSet)
> summary(M_final)
>kable(anova(M_final),caption = "ANOVA table")
```

2. environmental variables residual plot



3. new residual plot



4. final model summary

```
Call:
lm(formula = I(Y^1.14) ~ E3 + E4 + G12, data = DataSet)

Residuals:
    Min      1Q  Median      3Q     Max
-346.97  -78.42   -1.35   79.42  395.78

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -45.162     19.498  -2.316 0.020745 *
E3            24.916      1.376  18.110  < 2e-16 ***
E4            37.497      1.326  28.283  < 2e-16 ***
G12           31.903      8.476   3.764 0.000177 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 121.9 on 1000 degrees of freedom
Multiple R-squared:  0.5385,    Adjusted R-squared:  0.5371
F-statistic: 388.9 on 3 and 1000 DF,  p-value: < 2.2e-16
```

5. ANOVA table for final model

Table: ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
E3	1	5394660.5	5394660.54	363.09411	0.0000000
E4	1	11729414.3	11729414.34	789.46234	0.0000000
G12	1	210466.3	210466.30	14.16569	0.0001771
Residuals	1000	14857471.7	14857.47	NA	NA