DATA61646 Report

Introduction

The objective is to find the correct model that describes the data in. We need to find the association between the outcome variables and the independent variables. Box-Cox transformation is used to find the non-linear transformation of the dependent variable if needed.

Methodology

We used R to solve the problem. The original data file contains one dependent variable and twenty-four independent variables (including 4 environmental variables and 20 genetic variables). I first analyzed the data using just environmental outcomes. The lm () function in R is used to create a regression model and then we used summary () function to look at the model. Next, we assessed the contribution of the genetic variables after I have controlled for the environmental variables assuming that we have only up to second-order interactions. I obtained the residual plot (see figure 1). We also used Box-Cox transformation to see whether I should apply a transformation of the dependent variables. The graph implies that I do not need any transformation of my model. We obtained our adjusted R squared (0.595651) and residual plot (see figure 2). The residual plot was not patternelss. I next examine whether I could simplify the model using stepwise regression. The statistical package leaps is loaded and we use regsubset () to do the test (see table 1). The result shows that we had a huge increase in adjusted R squared and Bayesian Information Criterion (BIC) in the third model. As a result, I choose the variables in the third model as my key independent variables. Then I want to make sure the significant effect of my variables. I also used the second order interactions by using the second power in the model request. After I choose my model, I perform t-test. The terms that I choose all have P-value less than 0.001(see table 2). In the final fit, I used anova () function to examine my model (see table 3).

Results

The variables that I chose were E1, E3, G17, G19. My final model is: Y=7091.33E1+12307.81G17+12307.81E3*G19 -282322.86. Our analysis found associations with genetic variables after the environmental variables had been controlled. In specific, the E3-G19 interaction. Our analysis of variance table to show that each term in my model all has P-value less than 0.001 and t value bigger than 4. The adjusted R-squared is 0.6077. The multiple R-squared is 0.6089. The residual standard error is 20800 on 996 degrees of freedom and the F-statistic is 516.9 on 3 and 996 DF with p-value: < 2.2e-16.

Conclusion

Our model includes variables that has significant main effect on the outcome variable. The model has p-value less than 0.001 in both t-test and F-test. The model increases the adjusted R squared by a certain level. To better access our model, we could further estimate of the accuracy for our model such as the confidence intervals in addition to P-value.

Reference

Cummings P, Rivara FP. Reporting Statistical Information in Medical Journal Articles. *Arch Pediatr Adolesc Med*. 2003;157(4):321–324. doi:10.1001/archpedi.157.4.321

Avshalom Caspi, et al. Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science*. 2003; 301: 386-389. (2003); doi: 10.1126/science.1083968

Appendix

Figure 1

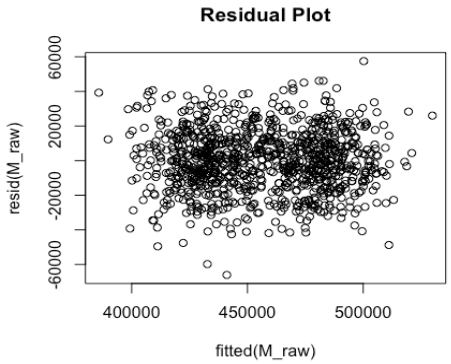

Figure 2

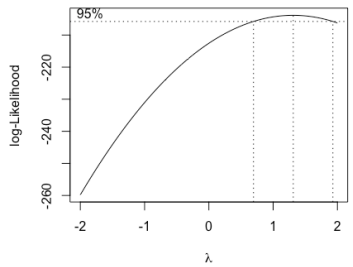

Table 1

model	adjR2	BIC
(Intercept)+E1:G19	0.528098731571488	-738.171484114516
(Intercept)+E1+E1:G19	0.572439797253254	-830.940918231432
(Intercept)+E1+E1:G19+E3:G17	0.606753911646391	-908.696184012364
(Intercept)+E1+E1:G19+E3:G17+G10:G20	0.60854047741993	-907.346424591583
(Intercept)+E1+E1:G19+E3:G17+E3:G19+G10:G20	0.610260166644434	-905.8468967428

Table 2

Summary TABLE

	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	-282322.86	63833.26	-4.423	1.08e-05
E1	7091.33	638.52	11.106	< 2e-16
G17	12307.81	1315.76	9.3541	< 2e-16
E3:G19	476.680	13.17	36.202	< 2e-16

Table 3

ANOVA TABLE

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
E1	1	62681187861	62681187861	144.85260	0
G17	1	41159631950	41159631950	95.11753	0
E3:G19	1	567129694990	56712969499	1310.60391	0
Residuals	996	430993049790	432723946	NA	NA

Technical Appendix

```
install.packages("leaps")

P2 <- read.csv(file="~/Desktop/P2 61646.csv",header=TRUE,sep=",")
M_E <- lm(Y ~ E1+E2+E3+E4, data=P2)
summary(M_E)

M_raw <- lm( Y ~
(E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+G1
 7+G18+G19+G20)^2, data=P2)
summary(M_raw)$adj.r.square
plot(resid(M_raw) ~ fitted(M_raw), main='Residual Plot')
library(MASS)
boxcox(M_raw)

M_trans <- lm( Y ~
 (E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+G1
 7+G18+G19+G20)^2, data=P2)
```

```
summary(M_trans)$adj.r.square
plot(resid(M_trans) ~ fitted(M_trans), main='New Residual Plot')

library(leaps)

M <- regsubsets(  Y ~
 (E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+G1
 7+G18+G19+G20)^2, data=P2,
nbest = 1 , nvmax=5,method = 'forward', intercept = TRUE )
temp <- summary(M)

 Var <- colnames(model.matrix(M_trans))
 M_select <- apply(temp$which, 1,  function(x) paste0(Var[x],
 collapse='+'))
library(knitr)
kable(data.frame(cbind( model = M_select, adjR2 = temp$adjr2, BIC =
 temp$bic)), caption='Model Summary')

M_main <- lm(  Y~ (E1+E3+G17+G19), data=P2 )temp <- summary(M_main)
temp <- summary(M_main)
kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.001, ],
 caption='Sig Coefficients')

M_2nd <- lm(  Y~ (E1+E3+G17+G19)^2, data=P2 )temp  <- summary(M_2nd)
temp  <- summary(M_2nd)
kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.01, ],
 caption='2nd Interaction')

M_2stage <- lm( Y ~ E1+G17+E3:G19, data=P2)
Summary (M_2stage)
kable(anova(M_2stage), caption='ANOVA Table')
```

End of Report