AMS 315 Project 1

Zarif Ahmed

May 2025

Part A

Introduction

The objective of this assignment is to investigate a synthetic dataset made using some function that models the interactions between genetic variables and environmental factors. We wish to reverse engineer the model used to produce the synthetic data using multiple regression techniques to identify significant main effects and interaction terms. The data itself is based on GxE interactions that were studied in the Caspi et al. (2003) paper which identified significant interactions between genetic and environmental variables. The report investigated the interaction between the 5-HTT gene and stressful life events on the risk of depression. The Caspi report found that there is a gene-environment interaction where individuals with short alleles on the 5-HTT gene are more likely to develop depression after facing stressful life events.

Methods

All statistical computations and analyses mentioned were done in R. The code used is given in Figure 1. First we draw scatterplots of Y vs. Ei to find the relation between the dependent variable vs every environmental variables. Figure 1 in the Appendix shows that E1 has a somewhat non-linear relation with Y where as E2, E3, E4 has no impact on Y on their own. This suggests we should treat E1 as a non-linear environmental variable. While E2, E3 and E4 had no impact on Y on their own, they could still impact Y if there is any interactions between these environmental variables and genetic variables. Afterwards we create an initial model with just the environmental variables and use the Box-cox transformation provided by the MASS library to get a transformation variable lambda, that produces the highest log-likelihood. In our case the value we found is .38 so we will transform Y with the power of .38 for all subsequent analysis. We calculated the Pearson correlations between $Y^{.38}$ and all 24 independent variables. The top 10 variables with the strongest absolute correlations were selected for further analysis. We applied Bonferroni inequality, to address the issue of multiple comparisons, by multiplying the raw p value by the number of predictors, 11 in our case. A term was found to be significant if its adjusted p was ¡=.5. We found interactions between our independent variables by performing lasso regression on the top 10 variables plus $E1^2$. We make an intermediate model using top 2 interactions from the lasso model and the top 3 features from the Bonferroni inequality. Afterwards we further reduce the model by taking the t-test of the final independent variables we've chosen and all combinations of interactions that can be found from the.

Results

The final model that we come up with is:

$$Y^{0.38} = \beta_0 + \beta_1 E_1^2 + \beta_2 (G_{11} \cdot G_{15}) + \varepsilon$$

The adjusted r^2 value in this model was .4124. This is higher than our all environmental variable model which had an adjust r^2 value of .3723.

```
Call:
lm(formula = I(Y^0.38) ~ I(E1^2) + G11:G15, data = Dat)

Residuals:
     Min       1Q   Median       3Q      Max
-162.004  -36.139    0.639   37.998  193.564

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.85008    4.22179  27.678  < 2e-16 ***
I(E1^2)       1.18111    0.04049  29.171  < 2e-16 ***
G11:G15      24.22697    4.35909   5.558 3.33e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55.37 on 1260 degrees of freedom
Multiple R-squared:  0.4134,    Adjusted R-squared:  0.4124
F-statistic: 443.9 on 2 and 1260 DF,  p-value: < 2.2e-16
```

Figure 1: Final equation ANOVA table

Conclusions and Discussion

The results of our analysis shows that the addition of genetic variables on top of our environmental variable did improve our fit resulting in a higher adjusted r^2 value. However there are several limitations to our findings. Even though our results improved with the addition of genetic variables our r^2 value was still pretty low. Our methodology involved alot of trial and error in finding the correct variables to fit our model. The data we worked on is also synthetic so our findings may not generalize to real world biological data.

Appendix

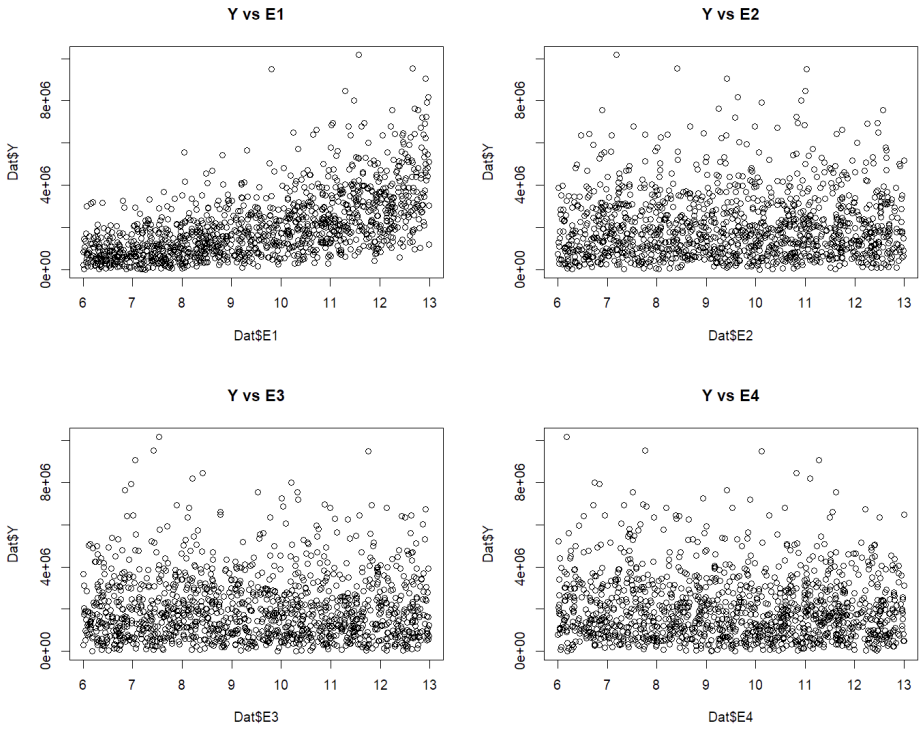

Figure 2: Scatter plot of Y vs Ei

```r
library(MASS)
library(leaps)
library(knitr)
library(glmnet)

setwd("~/AMS 315 Project 2")
Dat <- read.csv('P2_422817.csv', header=TRUE)

# Look for environmental variables with the most impact
plot(Dat$E1, Dat$Y, main = "Y vs E1")
plot(Dat$E2, Dat$Y, main = "Y vs E2")
plot(Dat$E3, Dat$Y, main = "Y vs E3")
plot(Dat$E4, Dat$Y, main = "Y vs E4")

# Find the max lambda value from M_raw that will give an estimated transformation for y
M_raw <- lm(Y ~ E1 + I(E1^2) + E2 + E3 + E4, data = Dat)
plot(resid(M_raw) ~ fitted(M_raw), main='Residual Plot')
summary(M_raw)

bc <- boxcox(M_raw)
lambda <- bc$x[which.max(bc$y)] # Optimal lambda found to be .383838

# See the new transformed model
M_trans <- lm(I(Y^.38) ~ E1 + I(E1^2) + E2 + E3 + E4, data = Dat)
plot(resid(M_trans) ~ fitted(M_trans), main='Residual Plot')
summary(M_trans)

# Find correlation between Y and each IV
all_vars <- c("E1","E2","E3","E4", "G1", "G2", "G3", "G4", "G5", "G6", "G7", "G8", "G9", "G10",
              "G11", "G12", "G13", "G14", "G15", "G16", "G17", "G18", "G19", "G20")
correlations <- cor((Dat$Y)^.38, Dat[,all_vars])
cor_vector <- as.vector(correlations)
names(cor_vector) <- colnames(correlations)
sorted_cors <- cor_vector[order(abs(cor_vector), decreasing = TRUE)]
top_10 <- names(sorted_cors)[1:10]
top_10

# Run model with the top 10 most correlated variables plus E1^2
M_main <- lm(I(Y^.38) ~ E1 + I(E1^2) + E2 + E4 + G7 + G11 + G12 + G13 + G14 + G15 + G19, data=Dat)
temp <- summary(M_main)
kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.1, ], caption='Sig Coefficients')
```

Figure 3: First half of our code used to find our model

```r
# View bonferroni results for the most significant terms
bonferroni_results <- data.frame(
  Variable = rownames(temp$coefficients),
  Coefficient = temp$coefficients[, 1],
  Original_p = temp$coefficients[, 4],
  Bonferroni_p = temp$coefficients[, 4] * 11,
  Significant = temp$coefficients[, 4] * 11 <= .5
)

# Sort by adjusted p-value
bonferroni_results <- bonferroni_results[order(bonferroni_results$Bonferroni_p), ]
kable(bonferroni_results)

# Perform Lasso technique on the interactions
X <- model.matrix(~ (E1 + I(E1^2) + G11 + G13 + G12 + G14 + G15 + E4 + G7 + E2 + G19)^2, data = Dat)[, -1]
y <- (Dat$Y)^0.38
cv_fit <- cv.glmnet(X, y, alpha = 1)
coefs <- coef(cv_fit, s = "lambda.min")
sig_vars <- rownames(coefs)[which(coefs != 0)]
sig_interactions <- grep(":", sig_vars, value = TRUE)
sig_interactions
coefs_df <- as.matrix(coefs)
interactions <- coefs_df[grep(":", rownames(coefs_df)), , drop = FALSE]
interactions <- interactions[interaction_coefs != 0, , drop = FALSE]
sorted_interactions <- sort(abs(interactions[,1]), decreasing = TRUE)
sorted_interactions

# Model with most significant variables and interactions
M_almost <- lm(I(Y^0.38) ~ I(E1^2)+ G15 + G13 +
                 G11:G15 + G11:G14, data = Dat)
summary(M_almost)
semi_final_summary <- summary(M_almost)

M_2stage <- lm( I(Y^.38) ~ (I(E1^2)+G1+G11+G15+G13)^2, data=Dat)
temp <- summary(M_2stage)
kable(temp$coefficients[ abs(temp$coefficients[,3]) >= 4, ])

M_final <- lm(I(Y^0.38) ~ I(E1^2) +
                G11:G15, data = Dat)
summary(M_final)
```

Figure 4: Second half of our code