DATA1646 Report

Introduction

Recent studies have shown that most common genetic risks produce modest effect at least when considered individually, which indicates that complex diseases are the result of the joint efforts of many genetic and environmental contributors, and possibly their interactions. The objective in this report is to find the correct model that describes the given data generated synthetically as well as the association between the outcome variable and independent variables. Box-Cox transformation will be used to find the non-linear transformation of the dependent variable if needed.

Methodology

The original dataset contains 1000 obeservations with one dependent variable and twenty-six independent variables (including 6 environmental variables and 20 genetic variables). A detailed chart that summarized the statistics of all variables is included in appendix Figure 1. I used mice() function in "mice " package which implements an iterative Markov Chain Monte Carlo type of algorithm to impute missing values by creating multiple imputations. The method for numerical variables is " Predictive Mean Matching "(pmm) and the method for categorical variables is " Logistic Regression "(logreg). I also turn G variables into indicator variables using as.factor() function in R.

I employ two steps to solve the problem. The fisrt step is trying to use stepwise methods to eliminate variables with negligible effects. The second step is to apply all-possible-regressions (or best subset regression) to my selected set of candidate variables. Then I compare each candidate models formerly selected by each metric(adjusted R squared, BIC). Finally I perform model adequacy checking and model diagnostics by checking residual plots and influential points.

I first build the regression model on all variables to investigate the main effects using lm() function. The p-value is < 2.2e-16. This means that, at least, one of the predictor variables is significantly related to the outcome variable. The adjusted R squared for this model is 0.3626 meanning that there is small asscociation between the independent variables and the dependent variable. The residual plot is a horizontal line without distinct patterns is an indication for a linear relationship.The Scale-Location plot residuals are spread equally along a horrizontal line. The studentized Breusch-Pagan test with p-value 0.3967 indicates we cannot reject the null hypothesis constant variance of error terms. Thus there is no need for transformations of response variable. There are some possible outliers (#786,#479,#130) that may affect our regression model. Moreover, if we look at the p-values of the estimated coefficients, we see that only some predictors are statistically significant related to y. Significant independent variables (E3+G2+G9+ G19) are picked out for further analysis. This means we need to perform some variable selection. One reason for variable selection is that the variance of the prediction y will increase and the precision of the parameter estimates will decrease as the number of regressors increases.

We have 26 E and G variables and 325(26 choose 2) interactions terms. It is not feasible to examine all 2^(26+325) regression models. We first consider variable selection without adding interaction terms. We will use regsubsets function in leaps package to perform stepwise regressions to check whether we could exclude any regressor with insignificant effects. Regsubsets can be used to identify different best models with different size. The function provides 3 methods to perform all possible regressions: (1) forward selection , (2) backward elimination , and (3) stepwise regression, which is just a combination of the previous two. Note that backward elimination requires the number of observations larger than the number of variables, so that the full model can be fit. During the entire modeling process I choose to apply all three procedures in order to learn more things that might be overlooked by using only one selection method from the data. The metric I used to evaluate each model is the adjusted R_2 and the Bayesian Information Criterion (BIC). After the selection, both the three procedures agree on two set of variables , one for biggest adjusted R_2 and one for smallest BIC. Now I have two candidate models without interaction terms, i.e model.m1(E3+E5+G2+G3+G4+G6+G8+G9+G11+G14+G18+G19+G20) with largest adjusted R_2 and model.m2(E3+G2+G9+G19) with smallest BIC.

After I have picked out my candidate variables without interactions, next task is to assess the contribution of interactions among the variables. I added the second order interactions by using the second power and variables from model.m1 in the model request. Doing the hypothesis test gives me significant candidates including interacion terms. After I compared the hypothesis test result with the result I got in model.m1 and model.m2, I picked two sets of important variables for doing my second order interactions variable selections, namely E3+E5+G2+G3+G8+G9+G18+G11+G19+G20 and E3+G2+G9+G19. The approach is similar to previous method using regsubsets except now I have to do three types of variable selections for each set of candidate variables. Comparing the best models recommended by each criterion gives me 2 candidate models: model.m3 (Y~E3+E3:G2+E3:G9+G9:G19) and model.m4 (Y~E3+G2+E3:G9+G2:G18+G9:G19).

Since there is possible third order or higher order interactions among all variables, I continue to use the variables with highest correlation with response variables (E3+G2+G9+ G19) to do my third order interactions analysis. These individual regressors appear both in the t tests and variable selection results. After I incorprated third order interactions in my analysis, I get a candidate model model.m5: Y~E3+G2+G9+G9:G19.

The stepwise regression has indicated 5 final candidate models(model.m1, model.m2, model.m3, model.m4 and model.m5). The big picture is E3+G2+G9+G19 and E3:G9 interaction term seem to be are significantly associated with the response variables and hence should be used to perform all-possible-regressions. I use regsubsets() function in r , set method = "exhasustive" and use third order interactions. This produces best subsets ranked by various criteria. After performing a comprehensive analysis of all my candidate models using lm(),anova() and BIC, adjusted R squared values. The models I choosed are model.m3 and model.m6 (Y~ E3+E3:G2+E3:G9+G9:G19+E3:G2:G9+G2:G9:G19). However, further analysis shows that model.m6 has a serious multicolineary problem and dose not have satisfactory performance in other statistics such as BIC. Since I have the third-order interactions have already shown serious multicolineary problem, there is no need to consider fourth-order interaction. My best model is set to be model.m3.

The final step is applying specific analysis through the model. So far we have assumed that there is a linear relationship between the predictors and y. If the relationship is quite far from linear, then it may yield an unstable model. I will check the linear regression assumptions by starting with the residual plots. The residuals can be contained in a horizontal band and no obvious pattern in plots(Fig.2). This suggests that we can assume linear relationship between the predictors and the outcome variables.We can also see from the plots some potential outliers(#786,#63,#479). Then I will check the Normal Q-Q plot(Fig.2): There is small deviance from the line. However it does not affect the model greatly, so we can assume normality.The third scale-location plot(Fig.2) shows a horizontal line with equally spread point, suggesting no heteroscadacity problemm, which implys no need for a transformation on regressors or higher order terms. The studentized Breusch-Pagan test with p value 0.4242 also indicates constant variance in error terms. We use Cook's distance to check the influence of a value. This metric considers both leverage and residual size. We consider leave these points in our model for further analysis. The plot(Fig.2) above highlights the top 3 most extreme points (#63, #943 and #97) including outliers and leverage points. Next I will perform model diagnostics to the individual influential points separately. My cutoff values for diagnostics are as follows: DIFFITS > 2*square root of 5/1000 and Cook's distance >1 are considered influential points. Hatvalues > 8/1000 are considered high leverage points. Absolute r student values> 2 are considered outliers. I use vif() function in car package to examine multicolinearty. The result shows no seriously problematic VIF(variable inflation factor) values. I also applied lack of fit test, partial F test and analysis of variance to my final model model.m3. The result is summarized in the following section.

Results

The variables that I chose were E3,E3:G2,E3:G9,G9:G19. My final model is:
Y = 6.5763664+ 0.0966821E_3 +0.0029542 E_3G_9 + 0.0031117E_3G_2-0.1209036 G_9G_{19}. Our analysis found associations with genetic variables and the environmental. In specific, the E_3G_2 , E_3G_9 and G_9G_{19} interaction. I use anova() to do a partial F test for a reduced subset model(model.m3) and a full model contains all interactions and variables(full.model2) . The first test is partial F test, P value is 0.85 hence I accept the null hypothesis that the extra coefficents in the full model is zero. The second test is lack of fit test, I use pureErrorAnova() to test my model.m4. The P value is 0.0136<0.001. Hence I accept the null hypothesis that the model provides an adequate model in the significance level 0.001. Our analysis of variance table anova() shows that each coefficient in my model has P-value less than 2.2e-16. The adjusted R-squared is 0.3549 The residual standard error is s 0.1747 on 994degrees of freedom and the F-statistic is110.9 of 5 and 994 degrees of freedom with p-value: 2.2e-16. My candidate models are summarized in Table2. To further estimate the accuracy for our model such as the confidence intervals, we can see they all exclude 0(Table 2). Model adequacy checking are in (Fig2) and some of the influential points are displayed in Table 3.

Conclusion and Discussion

Our model includes variables that has significant main effect on the outcome variable and the 95% confidence interval excludes 0 for each term. The model has p-value less than 0 in t-test. Our model is significant and there are no obvious problems with assumptions or other indications of model inadequacy. However our model has a low value for R squared meaning the proportion of variability explained by the model is low. This is because a lot of variability occurs in the measurements of the y. Here the variability in the response probably ouccurs because the response is a subjective measurement of psychology problem. Our result is affected by the variance of our test subjects. Another limitation occured in our regression analysis is that since we have included interaction terms in our model, there might be some mild multicolinearty problems. While strong multicollinearity might produce poor estimates of the individual model parameters, this does not necessarily mean the model performs badly in prediction. If our data are restricted to be near the regions of the X space, then precise predictions can often be made. Since we also discard model.m6 because of its high multicolinearty, we could have the possibility of omitting one or more third-order interactiton terms. Another limitation is for influential points, since we cannot know if there is an error in measurement , so we have no justification for their removals. Alternatively, we could use robust estimation techniques to deal with influential points.

Reference

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

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

Risch, N., Herrell, R., Lehner, T., Liang, K.-Y., Eaves, L., Hoh, J., … Merikangas, K. R. (2009, June 17). Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938776/

(PDF) Introduction to Linear Regression Analysis, 5th Ed ... books.google.com/books/about/Introduction_to_Linear_Regression_Analys.html?id=0yR4KUL4VDkC.

Appendix

Figure 1

```
       Y                E1              E2              E3              E4              E5              E6       
 Min.   :10.84   Min.   :47.29   Min.   :47.31   Min.   :47.07   Min.   :45.54   Min.   :47.24   Min.   :46.25  
 1st Qu.:11.44   1st Qu.:49.25   1st Qu.:49.25   1st Qu.:49.32   1st Qu.:49.30   1st Qu.:49.33   1st Qu.:49.28  
 Median :11.58   Median :49.95   Median :49.99   Median :50.02   Median :49.98   Median :50.01   Median :50.01  
 Mean   :11.58   Mean   :49.96   Mean   :49.97   Mean   :50.01   Mean   :49.96   Mean   :50.00   Mean   :50.01  
 3rd Qu.:11.74   3rd Qu.:50.63   3rd Qu.:50.62   3rd Qu.:50.70   3rd Qu.:50.64   3rd Qu.:50.63   3rd Qu.:50.71  
 Max.   :12.19   Max.   :53.22   Max.   :53.43   Max.   :53.04   Max.   :53.30   Max.   :52.76   Max.   :53.17  
 NA's   :105     NA's   :28      NA's   :16      NA's   :25      NA's   :25      NA's   :24      NA's   :16     
    G1              G2              G3              G4              G5              G6              G7              G8              G9              G10       
 0   :206   0   :192   0   :207   0   :193   0   :214   0   :203   0   :207   0   :198   0   :206   0   :187  
 1   :791   1   :797   1   :790   1   :799   1   :778   1   :790   1   :785   1   :793   1   :790   1   :809  
 NA's:  3   NA's: 11   NA's:  3   NA's:  8   NA's:  8   NA's:  7   NA's:  8   NA's:  9   NA's:  4   NA's:  4  
   G11             G12             G13             G14             G15             G16             G17             G18             G19             G20       
 0   :177   0   :177   0   :201   0   :199   0   :212   0   :197   0   :195   0   :207   0   :205   0   :174  
 1   :820   1   :816   1   :795   1   :798   1   :784   1   :796   1   :799   1   :784   1   :785   1   :816  
 NA's:  3   NA's:  7   NA's:  4   NA's:  3   NA's:  4   NA's:  7   NA's:  6   NA's:  9   NA's: 10   NA's: 10  
```

Figure 2

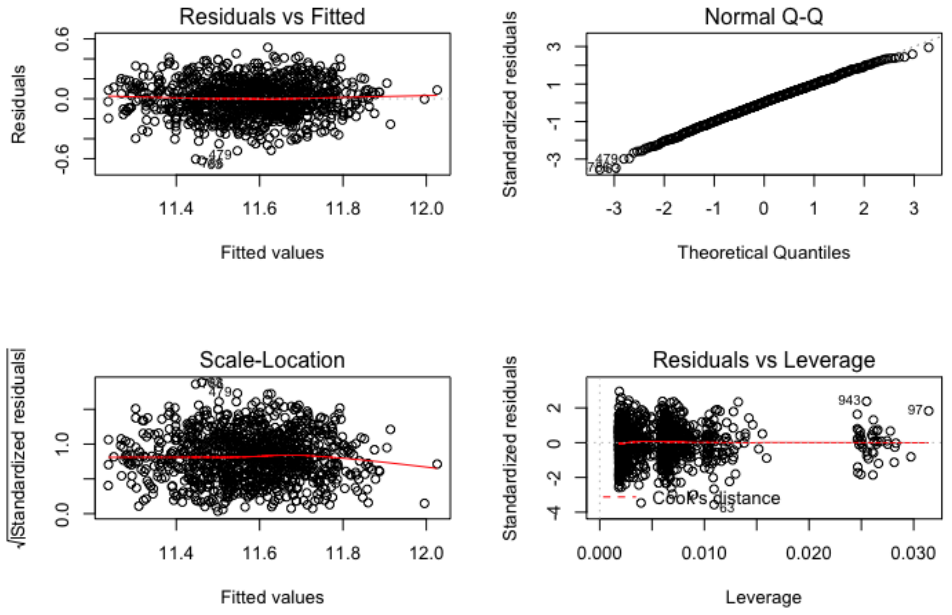

Table 1

model.m1 's variables : E3+E5+G2+G3+G4+G6+G8+G9+G11+G14+G18+G19+G20
model.m2 's variables : E3+G2 +G9+G19
model.m3 's variables : E3+E3:G9+G9:G19+E3:G2
model.m4 's variables : E3+G2+E3:G9+G9:G19+G2:G18
model.m5 's variables : E3+G2+G9+E3:G9+G9:G19
model.m6 's variables : E3+E3:G2+E3:G9+G9:G19+E3:G2:G9+G2:G9:G19
full.model 2 : all Es , Gs and any second order interaction terms among them

Table 2

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	6.5763664	0.2798909	23.496	< 2e-16	6.027121	7.125615
E3	0.0966821	0.0056100	17.234	< 2e-16	0.085673	0.107690
E3:G9	0.0029542	0.0006041	4.890	1.17e-06	0.001768	0.004139
G9:G19	-0.1209036	0.0152151	-7.946	5.19e-15	-0.150761	-0.091046
G2:E3	0.0031117	0.0002808	11.082	< 2e-16	0.002560	0.003662

Table 3(influential points)

Id	cd	lev	r	dffit
4	0.007101851	0.012976818	-1.8023187	-0.2066579	-0.0413489688
30	0.004061152	0.006010368	2.0105006	0.1563377	0.0085344267
36	0.003380180	0.029735686	-0.8133503	-0.1423874	0.0102246418
54	0.014266593	0.024964019	-1.8306354	-0.2929198	0.0021601456
63	0.023394751	0.010930621	-3.5850882	-0.3768850	-0.0318754770
87	0.004544200	0.005981293	2.1324441	0.1654161	0.0083202551
93	0.012604112	0.024670654	-1.7308228	-0.2752753	-0.0121726297

97	97	0.018088947	0.031434823	1.8308569	0.3298340	-0.0295534885
99	99	0.006185047	0.028144048	-1.1321823	-0.1926676	-0.0198946305

Technical Appendix

```
E <- read.csv("E_1646.csv",header = TRUE)
Y <- read.csv("Y_1646.csv",header = TRUE)
G <- read.csv("G_1646.csv",header = TRUE)
data <- data.frame(cbind(Y, E, G))
str(data)
library(leaps)
library('dplyr')
library('caret')
library(mice)
library(broom)
library(knitr)
data <- data %>%
  mutate(
    G1= as.factor(G1),
    G2 = as.factor(G2),
    G3 = as.factor(G3),
    G4 = as.factor(G4),
    G5 = as.factor(G5),
    G6 = as.factor(G6),
    G7 = as.factor(G7),
    G8 = as.factor(G8),
    G9 = as.factor(G9),
    G10 = as.factor(G10),
    G11= as.factor(G11),
```

```
  G12 = as.factor(G12),
  G13 = as.factor(G13),
  G14 = as.factor(G14),
  G15 = as.factor(G15),
  G16 = as.factor(G16),
  G17 = as.factor(G17),
  G18 = as.factor(G18),
  G19 = as.factor(G19),
  G20 = as.factor(G20),
 )
names(data) <- c('Y', paste0('E', 1:6), paste0('G', 1:20))
summary(data)
set.seed(123)
imdata <- mice(data, print=FALSE)
meth <- imdata$meth
for(i in 1:20){meth[paste0("G",i)] <- "logreg" }
imData <- mice(data, maxit = 3, print=FALSE,method=meth)
complete <- complete(imData)
md.pattern(complete)
# To test if there is any relationship between the outcome and the predictors, we start with fitting a multiple linear regression model using all the predictors
full.model <- lm(Y ~., data = complete)
main <- summary (full.model)
main

full.model2 <- lm(Y~(.)^2,data=complete)
main2 <- summary (full.model2)
main2
```

```
full.model3 <- lm(Y~(.)^3,data=complete)
main3 <- summary (full.model3)
main3
library(knitr)
kable(main$coefficients[ abs(main$coefficients[,4]) <= 0.001, ], caption='Sig Coefficients')

kable(main2$coefficients[ abs(main2$coefficients[,4]) <= 0.01, ], caption='Sig Coefficients')
plot(full.model)
install.packages('lmtest')
lmtest::bptest(full.model1)
fit1.back <- regsubsets(Y~., data = complete, nvmax = 26,nbest = 1,method = "backward", really.big = T)
fit1.step<- regsubsets(Y~., data = complete, nvmax = 26,nbest = 1,method = "seqrep", really.big = T)
fit1.forward<- regsubsets(Y~., data = complete, nvmax = 26,nbest = 1,method = "forward", really.big = T)
fit1.back.sum <- summary(fit1.back)
fit1.back.sum
fit1.step.sum <- summary(fit1.step.reg)
fit1.step.sum
fit1.forward.sum<-summary(fit1.forward)
fit1.forward.sum
which.min(fit1.back.sum$bic)
which.max(fit1.back.sum$adjr2)
which.min(fit1.step.sum$bic)
which.max(fit1.step.sum$adjr2)
which.min(fit1.forward.sum$bic)
which.max(fit1.forward.sum$adjr2)

model.m1 <- lm(Y~E3+E5+G2+G3+G4+G6+G8+G9+G11+G14+G18+G19+G20,data= complete)
model.m2 <- lm(Y~E3+G2+G9+G19,data=complete)
```

```
summary(model.m1)

summary(model.m2)

glance(model.m1)

glance(model.m2)

anova(model.m1,model.m2)

car::vif(model.m1)

car::vif(model.m2)

set.seed(123)

inter <- lm(Y~(E3+E5+G2+G3+G4+G6+G8+G9+G11+G14+G18+G19+G20)^2, data = complete)

temp <- summary(inter)

kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.1, ], caption='Sig Coefficients')

num <- 2^(55)

fit2.seq <- regsubsets(Y~(E3+E5+G2+G3+G8+G9+G18+G11+G19+G20)^2, data = complete,
nvmax=num,nbest = 1,method = "seqrep", really.big = TRUE)

fit2.forward <- regsubsets(Y~(E3+E5+G2+G3+G8+G9+G18+G11+G19+G20)^2, data =
complete,nvmax=num, nbest = 1,method = "forward", really.big = TRUE)

fit2.back <- regsubsets(Y~(E3+E5+G2+G3+G8+G9+G18+G11+G19+G20)^2, data = complete,
nvmax=num, nbest = 1,method = "backward", really.big = TRUE)

fit2.seq.sum <- summary(fit2.seq)

fit2.seq.sum

fit2.forward.sum <- summary(fit2.forward)

fit2.forward.sum

fit2.back.sum <- summary(fit2.back)

fit2.back.sum

which.min(fit2.seq.sum$bic)

which.max(fit2.seq.sum$adjr2)

which.min(fit2.forward.sum$bic)

which.max(fit2.forward.sum$adjr2)

which.min(fit2.back.sum$bic)

which.max(fit2.back.sum$adjr2)
```

```
fit3.forward <- regsubsets(Y~(E3+G2+G9+G19)^2, data = complete, nvmax = 1024,nbest = 1,method = "forward", really.big = T)

fit3.seq <- regsubsets(Y~(E3+G2+G9+G19)^2, data = complete, nvmax = 1024,nbest = 1,method = "seqrep", really.big = T)

fit3.back <- regsubsets(Y~(E3+G2+G9+G19)^2, data = complete, nvmax = 1024,nbest = 1,method = "backward", really.big = T)


fit3.forward.sum <- summary(fit3.forward)

fit3.back.sum <- summary(fit3.back)

fit3.seq.sum <- summary(fit3.seq)

fit3.forward.sum

fit3.back.sum

fit3.seq.sum

which.min(fit3.forward.sum$bic)

which.max(fit3.forward.sum$adjr2)

which.min(fit3.seq.sum$bic)

which.max(fit3.seq.sum$adjr2)

which.min(fit3.back.sum$bic)

which.max(fit3.back.sum$adjr2)

model.m3 <- lm(Y~E3+E3:G2+E3:G9+G9:G19,data=complete)

model.m4 <- lm(Y~E3+G2+E3:G9+G2:G18+G9:G19,data = complete)

summary(model.m3)

summary(model.m4)

glance(model.m3)

glance(model.m4)

car::vif(model.m3)

car::vif(model.m4)

anova(model.m3,model.m4)

fit4.forward <- regsubsets(Y~(E3+G2+G9+G19)^3, data = complete, nvmax = 1000,nbest = 1,method = "forward", really.big = T)
```

```
fit4.seq <- regsubsets(Y~(E3+G2+G9+G19)^3, data = complete, nvmax = 1000,nbest = 1,method =
"seqrep", really.big = T)

fit4.back <- regsubsets(Y~(E3+G2+G9+G19)^3, data = complete, nvmax = 1000,nbest = 1,method =
"backward", really.big = T)

fit4.forward.sum <- summary(fit4.forward)

fit4.back.sum <- summary(fit4.back)

fit4.seq.sum <- summary(fit4.seq)

fit4.forward.sum

fit4.back.sum

fit4.seq.sum

which.min(fit4.forward.sum$bic)

which.max(fit4.forward.sum$adjr2)

which.min(fit4.seq.sum$bic)

which.max(fit4.seq.sum$adjr2)

which.min(fit4.back.sum$bic)

which.max(fit4.back.sum$adjr2)

temp1 <- lm(Y~E3+G2+G9+G9:G19,data = complete)

temp2 <- lm(Y~E3+G2+G9+G9:G19+E3:G9,data = complete)

summary(temp1)

summary(temp2)

glance(temp1)

glance(temp2)

car::vif(temp1)

car::vif(temp2)

model.m5 <- temp1

modl.m1 <-lm(Y~E3+E5+G2+G3+G4+G6+G8+G9+G11+G14+G18+G19+G20,data= complete)

model.m2<-lm(Y~E3   +G2                  +G9+G19,data=complete)

model.m3<-lm(Y~E3                  +E3:G9+G9:G19+E3:G2,data=complete)

model.m4<-lm(Y~E3   +G2            +E3:G9+G9:G19+G2:G18,data = complete)
```

```
model.m5<-lm(Y~E3   +G2           +G9+E3:G9+G9:G19,data = complete)
library(leaps)
subset <- regsubsets(Y~(E3+G2+G9+G19)^3,nbest=5,data = complete,really.big = T,method='exhaustive')
which.min(summary(subset)$bic)
which.min(summary(subset)$cp)
which.max(summary(subset)$adjr2)
coef(subset,16)
coef(subset,26)
model.m6 <- lm(Y~  E3+E3:G2+E3:G9+G9:G19+E3:G2:G9+G2:G9:G19 ,data =complete)
x <- glance(model.m4)%>%
  dplyr::select(adj.r.squared,AIC,BIC,p.value)
kable(x)
y <- glance(model.m3)%>%
  dplyr::select(adj.r.squared,AIC,BIC,p.value)
kable(y)
z <- glance(model.m6)%>%
  dplyr::select(adj.r.squared,AIC,BIC,p.value)
kable(z)
car::vif(model.m6)
pureErrorAnova(model.m6)

library(broom)
model.diag.metrics <- augment(model.m3)
head(model.diag.metrics)
plot(model.m3, 1)
plot(model.m3, 2)
plot(model.m3, 3)
plot(model.m3, 5)
```

```
plot(model.m3, 4)

par(mfrow=c(2,2))

par(mfrow=c(1,1))

plot(model.m3)

Diagnostics <- data.frame(Id = 1:1000, cd = cooks.distance(model.m3), lev = hatvalues(model.m3), r =
rstudent(model.m3), dffit = dffits(model.m3),dfbeta = dfbeta(model.m3))

print(subset(Diagnostics,abs(r)>=2 ))

print(subset(Diagnostics, lev > 10/1000))

library(dplyr)

print(subset(Diagnostics, cd > 1 |abs(dffit)> 2*sqrt(5/1000) ))

model.diag.metrics %>%

  top_n(5, wt = .cooksd)

print(car::vif(model.m3))

summary(model.m3)

confint(model.m3)

library(alr3)

library(lmtest)

anova(model.m3)

anova(model.m3,full.model2)

bptest(model.m3)

lrtest(model.m3,full.model2)

pureErrorAnova(model.m3)
```

End of Report