R code used in BEAT

Linear regression:

> model <- lm(Age ~ Gria\_1, data = data)

> model <- lm(Age ~ Gria\_5 + Cdkn2a\_1 + TET2\_2, data = data)

model <- lm(Age ~ Gria\_5 + TET2\_2, data = data)

> summary(model)

Call:

lm(formula = Age ~ Gria\_5 + TET2\_2, data = data)

Residuals:

Min 1Q Median 3Q Max

-10.6055 -2.8331 -0.3352 2.0009 10.6328

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.6025 4.5781 1.879 0.0684 .

Gria\_5 4.3924 0.4766 9.216 5.26e-11 \*\*\*

TET2\_2 -1.1500 0.4699 -2.447 0.0194 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.826 on 36 degrees of freedom

Multiple R-squared: 0.7905, Adjusted R-squared: 0.7789

F-statistic: 67.92 on 2 and 36 DF, p-value: 6.048e-13

> summary(model)$coefficient

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.602530 4.5781437 1.879043 6.835367e-02

Gria\_5 4.392402 0.4766270 9.215597 5.257740e-11

TET2\_2 -1.149964 0.4698538 -2.447492 1.939429e-02

> confint(model)

2.5 % 97.5 %

(Intercept) -0.6823754 17.8874360

Gria\_5 3.4257581 5.3590467

TET2\_2 -2.1028712 -0.1970559

> sigma(model)/mean(data$Age)

[1] 0.4163734

> AIC(model)

[1] 238.3235

> BIC(model)

[1] 244.9777

Results

All possible combinations of three and two CpG sites for top correlated CpG sites were used to make multiple regression models. The top two models based on AIC and R squared values were chosen for all samples, females only, and males only. Asterisks represent the best model based on AIC score.

Combined male and female samples:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model Name** | **Multiple Regression Combos** | **Adjusted R Squared** | **Residual error** | **P-Value** | **AIC** | **BIC** |
| Model 1\* | Gria\_5 + TET2\_2 | 0.7789 | 4.826 | 6.05E-13 | 238.3235 | 244.9777 |
| Model 2 | Gria\_5 + Cdkn2a\_1 + TET2\_2 | 0.7779 | 4.836 | 3.78E-12 | 239.3985 | 247.7163 |

Females only models:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model Name** | **Multiple Regression Combos** | **Adjusted R Squared** | **Resid Error** | **P-Value** | **AIC** | **BIC** |
| Female 1\* | Gria 5, TET2 4 | 0.8877 | 4.282 | 7.96E-07 | 90.85048 | 93.68268 |
| Female 2 | Gria 5, Cdkn2a 1, TET2 4 | 0.8852 | 4.328 | 4.85E-06 | 91.86672 | 95.40697 |

Males only models:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model Name** | **Multiple Regression Combos** | **Adjusted**  **R Squared** | **Resid Error** | **P-Value** | **AIC** | **BIC** |
| Male 1\* | Gria\_1 + Cdkn2a\_5 + TET2\_2 | 0.9006 | 2.633 | 8.27E-11 | 120.209 | 126.099 |
| Male 2 | Gria\_1 + Cdkn2a\_3 + TET2\_2 | 0.8861 | 2.819 | 3.22E-10 | 123.485 | 129.375 |

LOOCV:

model <- train(Age ~ Gria\_5 + TET2\_2, data= data, method = "lm",

trControl = train.control)

Cross-Validation of all models

Cross-validation was done for all models using the “leave one out cross validation” method (LOOCV).

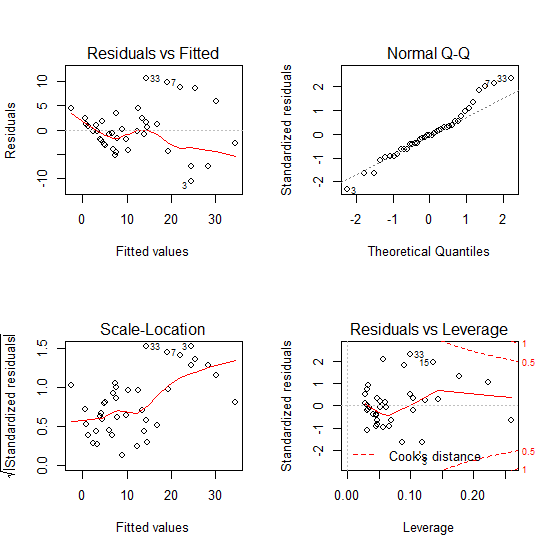
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name** | **sample size** | **RMSE** | **R squared** | **MAE** |
| Model 1 | 38 | 5.14 | 0.74 | 3.87 |
| Model 2 | 38 | 5.44 | 0.72 | 4.12 |
| Female 1 | 14 | 4.49 | 0.87 | 3.44 |
| Female 2 | 14 | 7.31 | 0.68 | 5.83 |
| Male 1 | 23 | 3.11 | 0.86 | 2.37 |
| Male 2 | 23 | 3.31 | 0.84 | 2.53 |

Normality plots:

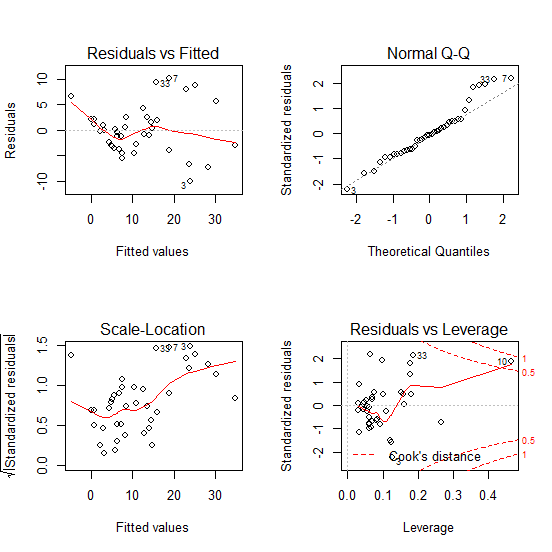
par(mfrow=c(2,2))

plot(beat1)

Model 1:



Model 2:



> shapiro.test(mydata$Gria\_5)

Shapiro-Wilk normality test

data: mydata$Gria\_5

W = 0.90957, p-value = 0.004198

> shapiro.test(mydata$TET2\_2)

Shapiro-Wilk normality test

data: mydata$TET2\_2

W = 0.98015, p-value = 0.7089

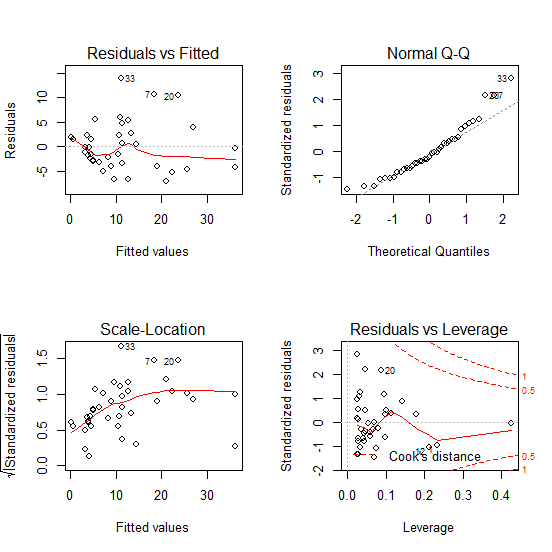
> shapiro.test(mydata$Cdkn2a\_1)

Shapiro-Wilk normality test

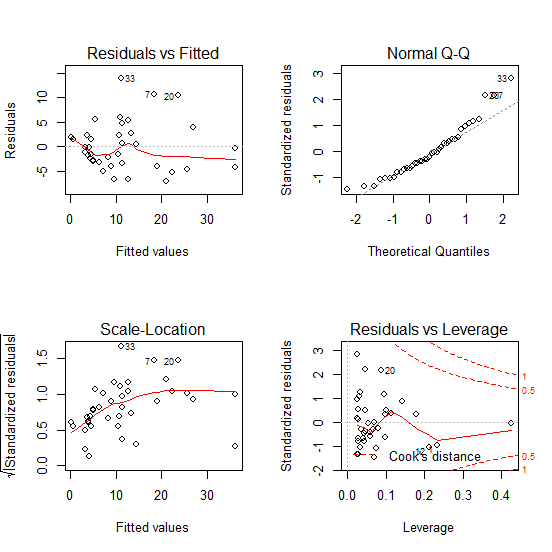
data: mydata$Cdkn2a\_1

W = 0.92642, p-value = 0.01382

Female 1:



Female 2:



> shapiro.test(data$Gria\_5)

Shapiro-Wilk normality test

data: data$Gria\_5

W = 0.89616, p-value = 0.08318

> shapiro.test(data$TET2\_4)

Shapiro-Wilk normality test

data: data$TET2\_4

W = 0.90431, p-value = 0.1108

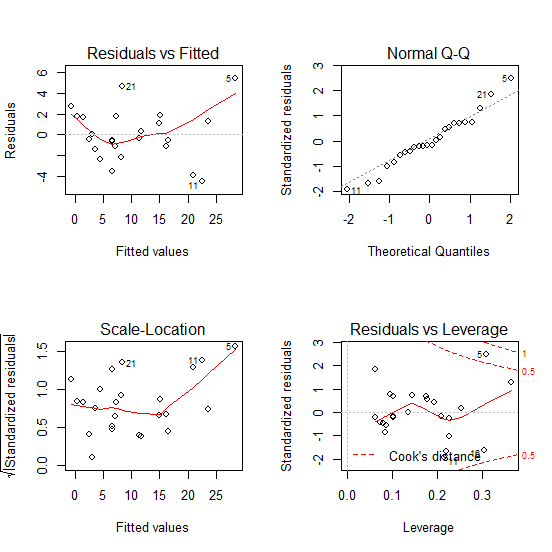
> shapiro.test(data$Cdkn2a\_1)

Shapiro-Wilk normality test

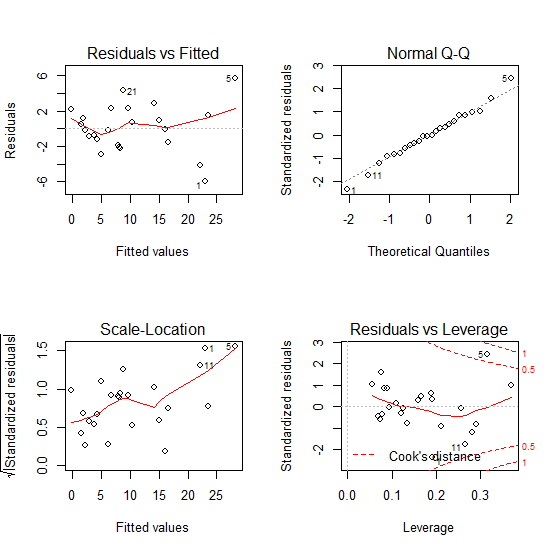
data: data$Cdkn2a\_1

W = 0.84, p-value = 0.01257

Male 1:



Male 2:



> shapiro.test(data$Gria\_1)

Shapiro-Wilk normality test

data: data$Gria\_1

W = 0.95629, p-value = 0.3685

> shapiro.test(data$TET2\_2)

Shapiro-Wilk normality test

data: data$TET2\_2

W = 0.96102, p-value = 0.4593

> shapiro.test(data$Cdkn2a\_3)

Shapiro-Wilk normality test

data: data$Cdkn2a\_3

W = 0.97484, p-value = 0.7854

> shapiro.test(data$Cdkn2a\_5)

Shapiro-Wilk normality test

data: data$Cdkn2a\_5

W = 0.94033, p-value = 0.1659

ANCOVA analysis

data<-read.csv("C:/BEAT\_R\_Files/BEAT\_DATASET\_CV.csv")

head(data)

mod<- lm(Age~Gria\_5 + TET2\_2 + Sex, data=data)

summary(mod)

anova(mod)

> mod6<- lm(Age~ (TET2\_2 + Sex)\* Sex, data=data)

> summary(mod6)

Call:

lm(formula = Age ~ (TET2\_2 + Sex) \* Sex, data = data)

Residuals:

Min 1Q Median 3Q Max

-13.662 -4.293 -1.081 4.543 18.685

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 44.4055 8.9061 4.986 1.67e-05 \*\*\*

TET2\_2 -3.4881 0.9848 -3.542 0.00115 \*\*

SexM -3.5103 12.5709 -0.279 0.78170

TET2\_2:SexM -0.5619 1.5157 -0.371 0.71307

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.058 on 35 degrees of freedom

Multiple R-squared: 0.4321, Adjusted R-squared: 0.3834

F-statistic: 8.876 on 3 and 35 DF, p-value: 0.0001644

> anova(mod6)

Analysis of Variance Table

Response: Age

Df Sum Sq Mean Sq F value Pr(>F)

TET2\_2 1 1185.40 1185.40 18.2566 0.0001408 \*\*\*

Sex 1 534.55 534.55 8.2327 0.0069312 \*\*

TET2\_2:Sex 1 8.92 8.92 0.1374 0.7130727

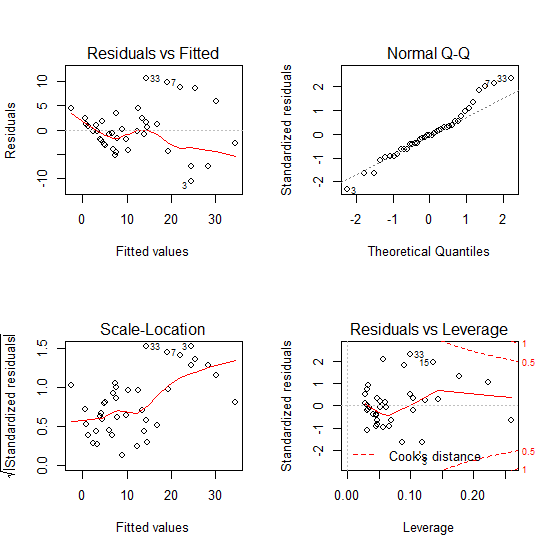
Residuals 35 2272.56 64.93

---

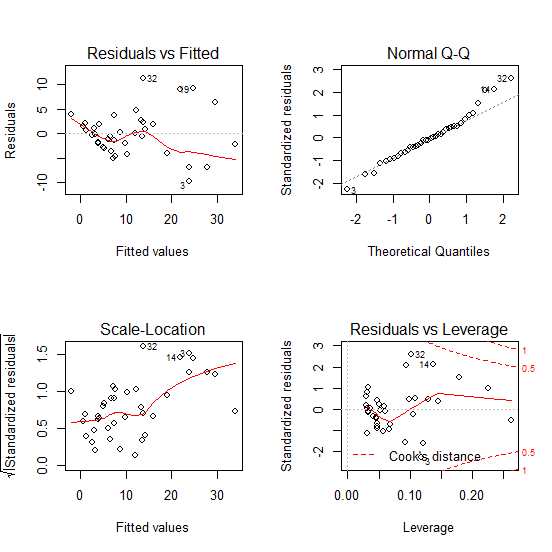
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Outlier Investigation

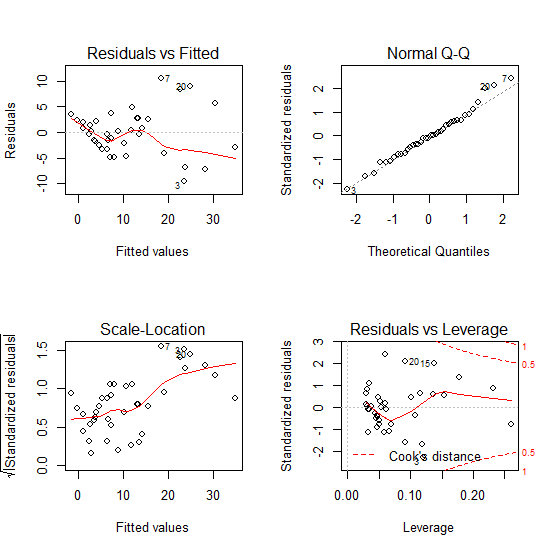
BEAT ALL



BEAT NO 7



BEAT NO 33



BEAT NO 7 or 33

