Homework 3

#A)

df=read.csv("D:/Old\_Data/math/Data science (toseeh/Files/battery.csv"

#B)

names(df)

dim(df)

summary(df)

str(df)

این مجموعه داده دارای 75 سطر و 6 ویژگی با نام های

Cycle charge discharge depth temp endvolt

است که تمامی این ویژگی ها عددی هستند و داده گمشده ندارند .

Cycle : تعداد دفعات شارژ و تخلیه شارژ است یعنی تعداد دفعاتی که یک دور کامل شارژ

صد درصد تخلیه شود.

Charge : به مقدار جریانی گفته میشود که باعث شارژ باتری در یک ساعت میشود. (آمپر)

Discharge : مقدار جریانی که باعث تخلیه شارژ در یک ساعت میشود.

Depth : میزان تخلیه باتری نسبت به ظرفیت آن. مثلا اگر 75 درصد باتری تخلیه شده باشد عمق آن را 75 درصد گوییم و هر بار باتری را شارژ کنیم باتری حداکثر 75 درصد شارژ میشود.

مثلا اگر تعداد دورهای باتری 1000 باشد، زمانی که 1000 بار چرخه شارژ و دیشارژ به اتمام رسید عمق باتری به 80 درصد میرسه یعنی 20 درصد ان از بین میره و دوباره چرخه 1000 دور انجام میشه تا عمق به 60 درصد برسه. پس 40 درصد باتری از بین رفته و مسلما طول عمر باتری و حداکثر میزان شارژ باتری کم میشود.

Temp : دمای باتری

Endvolt : میزان ولتاژ باتری بعد از شارژ شدن. (ولت)

#c)

داده گمشده نداریم.

#D)

mf=lm(cycles~.,data=df)

رگرسیون با متغیر پاسخ دور و بقیه متغیر ها به عنوان متغیر های مستقل.

mf$coefficients

output:

(Intercept) charge discharge depth temp endvolt

-6665.49451 -460.65473 -13.28588 -31.37477 18.54409 3612.39888

بتا صفر به عنوان عرض از مبدا.intercept

توصیف ضرایب:

اگر میزان عرض از مبدا یک واحد افزایش پیدا کند، تعداد دفعات شارژ و دیشارژ به طور متوسط 6665 دور کاهش می یابد.

اگر میزان جریان لازم برای شارژ باتری در یک ساعت یک آمپر افزایش پیدا کند تعداد دفعات شارژ و دیشارژ به طور متوسط 460.65 دور کم میشود.

اگر میزان جریان لازم برای تخلیه شارژ باتری در یک ساعت یک آمپر افزایش پیدا کند تعداد دفعات شارژ و دیشارژ به طور متوسط 13.28 دور کم میشود.

اگر عمق باتری یک درصد افزایش پیدا کند تعداد دفعات شارژ و دیشارژ به طور متوسط 31.37 دور کم میشود.

اگر دمای باتری یک درجه افزایش پیدا کند تعداد دفعات شارژ و دیشارژ به طور متوسط 18.54 دور زیاد میشود.

اگر ولتاژ پایانی یک ولت افزایش پیدا کند تعداد دفعات شارژ و دیشارژ به طور متوسط 3612.3 دور زیاد میشود.

E)

Summary(mf)

Call:

lm(formula = cycles ~ ., data = df)

Residuals:

Min 1Q Median 3Q Max

-661.1 -199.4 -69.6 60.9 3301.8

Coefficients:

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

(Intercept) -6665.495 13110.540 -0.508 0.612789

charge -460.655 133.683 -3.446 0.000973 \*\*\*

discharge -13.286 43.166 -0.308 0.759174

depth -31.375 45.829 -0.685 0.495886

temp 18.544 5.735 3.234 0.001876 \*\*

endvolt 3612.399 6537.313 0.553 0.582336

---

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

Residual standard error: 483.1 on 69 degrees of freedom

Multiple R-squared: 0.252, Adjusted R-squared: 0.1977

F-statistic: 4.648 on 5 and 69 DF, p-value: 0.001028

از آنجایی که مقدار پی برابر با 0.001 است اگر الفا را 0.05 در نظر بگیریم چون پی از الفا کمتر است فرض اچ صفر یعنی صفر بودن ضرایب ویژگی ها (بتا ها) رد میشود و در نتیجه باید از مدلی استفاده کنیم که تمام ضرایب برای ویژگی ها صفر نباشد.

همچنین از میزان پی برای هر یک از متغیر ها (ویژگی ها) میتوان دریافت که دو ویژگی شارژ و دما در توصیف متغیر پاسخ تاثیر دارند. زیرا فرض اچ صفر برابر است با صفر بودن ضریب آن متغیر و چون مقدار پی برای آن متغیر کمتر از الفا است پس فرض اچ صفر رد میشود یعنی این دو متغیر در توصیف متغیر پاسخ تاثیر زیادی دارد و در مدل باید باشند. (در صورتی که وابستگی بین متغیر ها نباشد)

برای دیگر متغیر ها مقدار پی یا تقریبا با الفا برابر است و یا از الفا بیشتر است به همین دلیل فرض اچ صفر برای انها را میپذیریم. در حقیقت به تنهایی تاثیر زیادی در توصیف متغیر پاسخ ندارند.

اگر متغیر های مستقل را فقط دو متغیر شارژ و دما در نظر بگیریم میزان توصیف متغیر پاسخ به 0.22 میرسد که البته با حالتی که تمام متغیر ها به غیر از دور به عنوان متغیر مستقل بودند تفاوت چندانی ندارد.

library(faraway)

vif(mf)

charge discharge depth temp endvolt

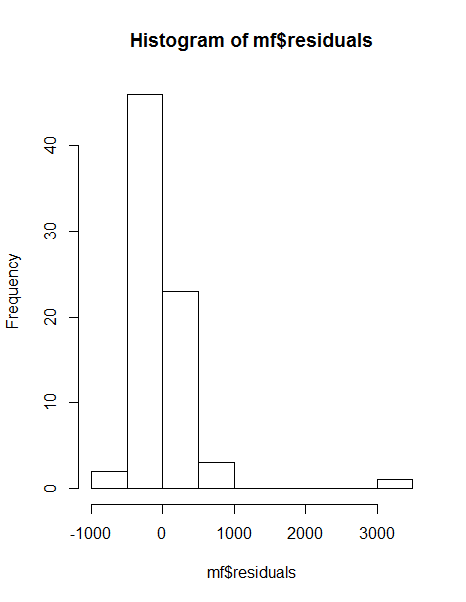
1.103904 1.045912 1.134237 1.055428 1.036999

F)

بررسی نرمال بودن خطا ها و ثابت بودن واریانس ویژگی ها

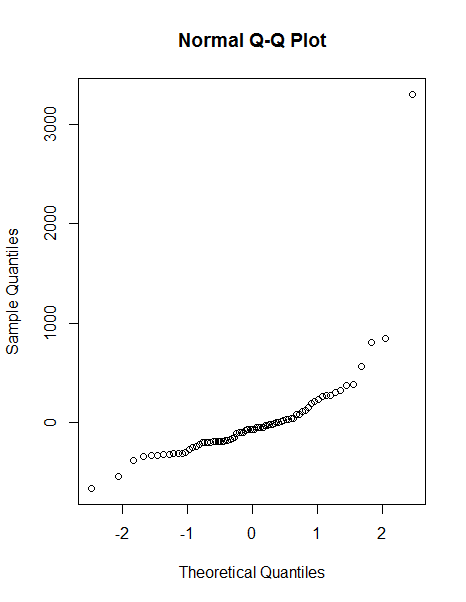
رگرسیون سه فرض اولیه دارد. نرمال بودن توزیع خطا ها ثابت بودن واریانس خطاها و خطی بودن رابطه متغیر پاسخ و متغیر های مستقل

hist(mf$residuals)



به نظر من نرمال نیست اما برای مطمین شدن در مورد نرمال بودن توزیع خطا ها میتوان نمودار چارک ها در نمونه را با مدل نرمال بررسی کرد اگر خط ایجاد شده تقریبا صاف بود یعنی نرمال است. برای اطمینان بیشتر باید ازمون فرض اندرسون را انجام داد.

qqnorm(mf$residuals)



خط واصل نقطه ابتدا و انتهایی را که رسم کنیم نقاط حول این خط نیستند پس نرمال نیست حال برای اطمینان اندرسون را انجام میدهیم.

ad.test(mf$residuals)

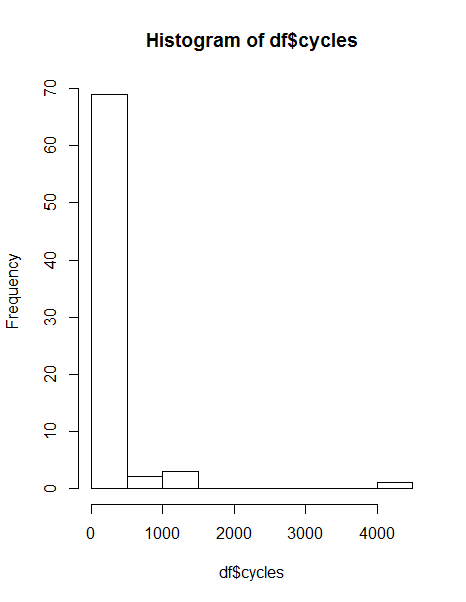
Anderson-Darling normality test

data: mf$residuals

A = 6.4936, p-value = 4.386e-16

فرض اچ صفر نرمال بودن توزیع خطا هاست اما چون مقدار پی از الفا برابر با 0.05 خیلی کمتر است این فرض رد شده و در نتیجه توزیع خطا ها نرمال نیست.

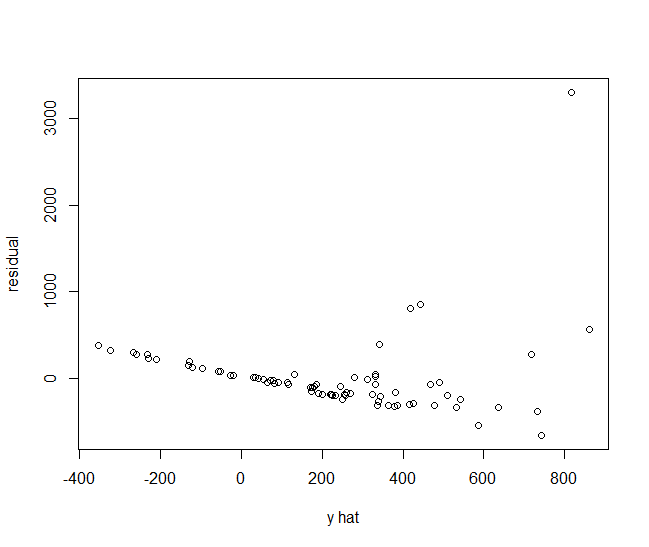
hist(df$cycles)



چون توزیع متغیر پاسخ هم نرمال نیست این امکان وجود دارد که با انتقال متغیر پاسخ توزیع خطا ها را به نرمال نزدیک کرد.

حال ثابت بودن واریانس را بررسی میکنیم.

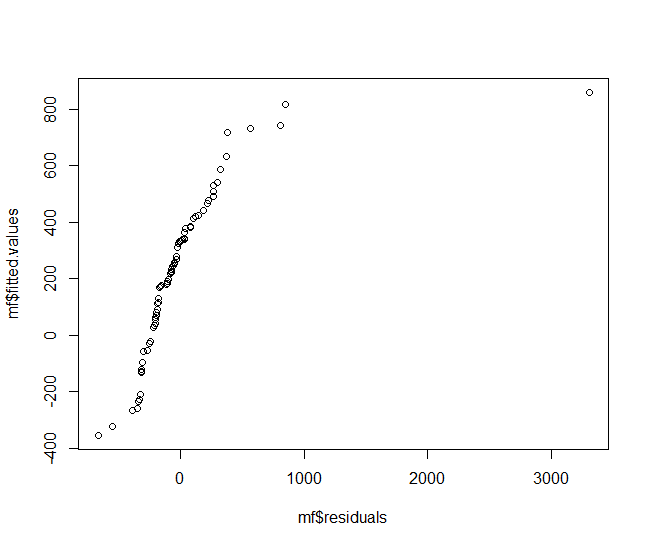
plot(mf$fitted.values,mf$residuals,xlab="y hat",ylab="residual")



الگوی خاصی در این شکل نمیبینم به همین دلیل راه های دیگر را برای تشخیص انجام میدهیم.

qqplot(mf$residuals,mf$fitted.values)

quantile ها در توزیع خطاها و براورد متغیر پاسخ را با هم مقایسه میکند.



به نظر میرسد تغییر در براورد متغیر پاسخ در واریانس خطاها تاثیر دارد اما ازمون فرض انجام میدهیم تا به قطعیت برسیم.

میخواهیم ببینیم ایا براورد متغیر پاسخ روی خطا ها تاثیری دارد یا نه. پس متغیر پاسخ در رگرسیون را خطا ها و متغیر مستقل را براورد cycle در نظر میگیریم. فرض اچ صفر عدم وجود رابطه است.

farz=lm(abs(mf$residuals)~mf$fitted.values)

Coefficients:

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

(Intercept) 110.7561 54.9800 2.014 0.047644 \*

mf$fitted.values 0.5900 0.1575 3.747 0.000356 \*\*\*

---

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

Residual standard error: 366.7 on 73 degrees of freedom

Multiple R-squared: 0.1613, Adjusted R-squared: 0.1498

F-statistic: 14.04 on 1 and 73 DF, p-value: 0.0003556

چون مقدار پی از الفا کمتر است پس فرض اچ صفر رد میشود پس واریانس ها تصادفی نیستند.

چون توزیع خطاها نرمال نیست و توزیع متغیر پاسخ هم نرمال نیست، لگاریتم توزیع متغیر پاسخ را در نظر میگیریم که در پایان به آن میپردازیم. (بعدا میبینیم که هم توزیع براورد پاسخ و هم توزیع خطاها به نرمال نزدیک میشود و واریانس خطاها نیز ثابت میشود.)

G)

برای بررسی داده های پرت ابتدا باید مدل را فیت کنیم. در اینجا مدلی که روی cycle فیت شود مناسب نیست اما روند را روی این مدل انجام میدهیم.

هدف یافتن نقاط پرت در فضای مستقل و پاسخ است و بعد از بررسی میکنیم که ایا وجود این نقاط روی رگرسیون تاثیر میگذارند یا نه. در صورتی که تاثیر گذار بودند و باعث تغییر زیاد ابرصفحه میشدند حذف میکنیم.

h=influence(mf)$hat

length(h)

ابتدا به دنبال سطرهایی میگردیم که دارای لورج بیشتر از 3p/n هستند که p تعداد متغیر های مستقل در مدل فیت شده است و n سایز نمونه است. در اینجا P=5 و n=75.

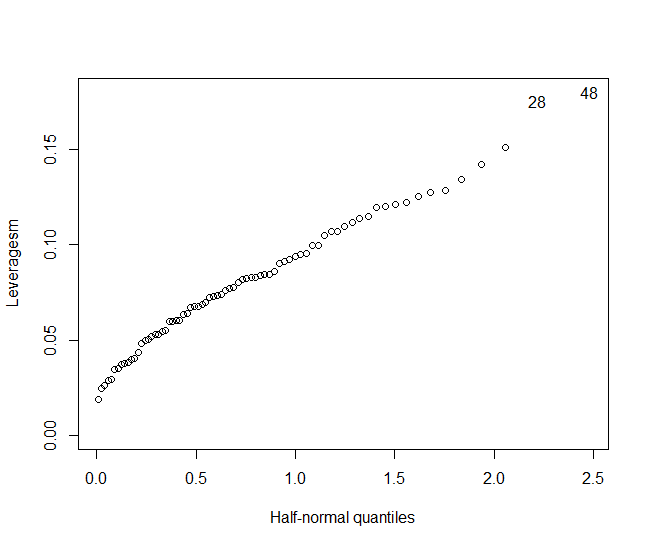
which(h>3\*5/75)

output: integer(0)

پس هیچ سطری با این ویژگی نداریم. حال روش های دیگر را امتحان میکنیم.

راه دیگر بررسی توزیع داده ها است. اگر داده ای پارامتر های توزیع ان با پارامتر های توزیع بقیه داده ها متفاوت باشد آن را پرت گوییم (در صورتی که تعداد ان ها کمتر از 15 درصد کل نمونه باشد). (توزیع داده ها یکسان است اما پارامتر ها ممکن است متفاوت باشد)

halfnorm(influence(mf)$hat, nlab = 2, ylab="Leverages")



دو سطر 28 و 48 ممکن است داده های پرت در فضای ایکس باشند. اما باید بررسی های دیگر انجام شود.

یکی از دیگر تعاریف داده های پرت: اگر مقدار آن در ویژگی ای از 3 برابر (-3 برابر) انحراف معیار آن ویژگی بیشتر (کمتر) بود.

پس دو سطر 28 و 48 را فراخوانی میکنیم.

df[28,]

cycles charge discharge depth temp endvolt

28 32.64789 1.998284 4.058757 8.088917 25.35331 1.980972

df[48,]

cycles charge discharge depth temp endvolt

48 26.55681 1.089765 5.29151 3.922258 15.20134 2.020632

colMeans(df)

cycles charge discharge depth temp endvolt

222.705994 1.006531 3.311309 6.431928 20.134210 1.999859

sqrt(diag(var(df)))

|  |
| --- |
| cycles charge discharge depth temp endvolt  539.32972832 0.44134942 1.33044420 1.30498781 10.06026526 0.00874751 |
|  |
| میزان تخلیه شارژ در سطر 28 از سه برابر انحراف معیار این ویژگی بیشتر است پس این ویژگی احتمالا داده پرت است.  همینطور میزان عمق در سطر 48 از سه برابر انحراف معیار این ویژگی بیشتر است.  حال با روش دقیق تر بررسی میکنیم ایا این نقاط در فضای ایکس پرت هستند یا نه تا مطمین شویم. |

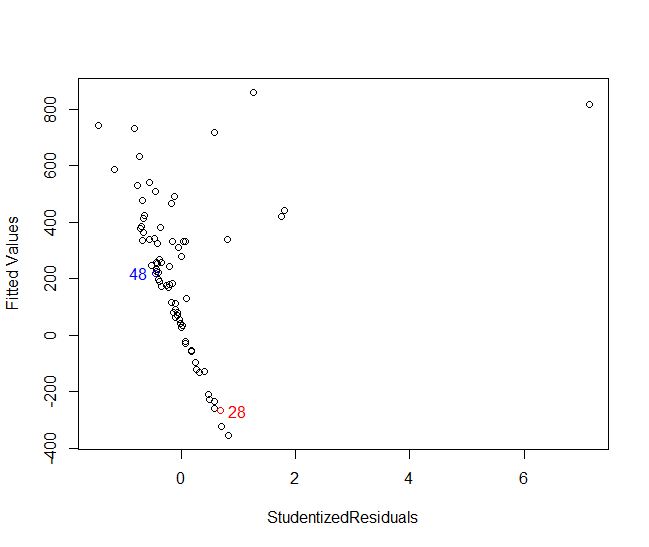
rstandard\_mf=rstandard(mf)

points(rstandard\_mf[28],mf$fitted.values[28],col="red")

text(rstandard\_mf[28]+0.3,mf$fitted.values[28],labels =28,col="red",cex=1)

points(rstandard\_mf[48],mf$fitted.values[48],col="blue")

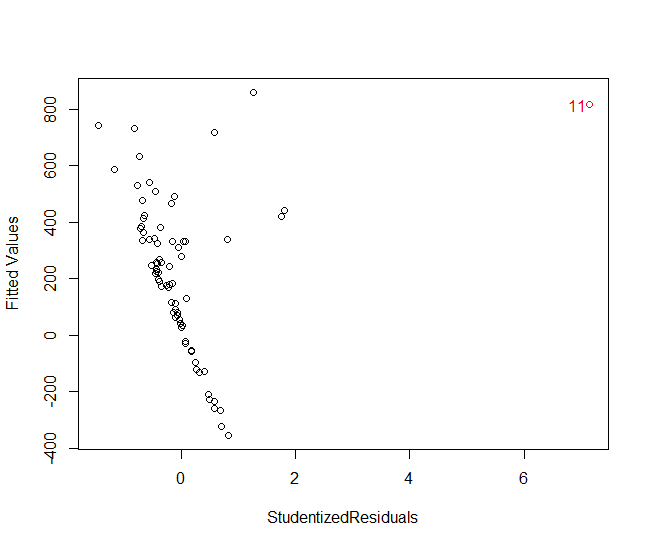
text(rstandard\_mf[48]-0.3,mf$fitted.values[48],labels =48,col="blue",cex=1)



در این روش میبینیم دو داده 28 و 48 به عنوان داده پرت در فضای ایکس نیتند اما اده پرت دیگری وجود دارد که مقدار rstandard برای آن بیشتر از 3 است. برای یافتن این داده:

which(rstandard\_mf>3)

خروجی آن داده 11 است که روی نمودار نشان میدهیم.



df[11,]

cycles charge discharge depth temp endvolt

11 4117.931 0.3383868 3.546622 6.063071 34.08629 2.004976

در صورتی که نمودار نقطه ای را برای این ویژگی ها رسم کنیم متوجه میشویم که داده 11 در مقدار ویژگی دور از بقیه نقاط دور است پس این داده در فضای پاسخ پرت است.

حال از ازمون فرض استفاده میکنیم تا پرت بودن هر یک از این داده را بررسی کنیم.

tmf=rstudent(mf)

با در نظر گرفتن الفا برابر با 0.05 و سایز نمونه 75، داریم

75/0.05=0.0006

T=function(x){

2\*(1-pt(abs(tmf[x]),df=75-5-1))

}

timf=apply(as.matrix(seq(1,75)), 1, T)

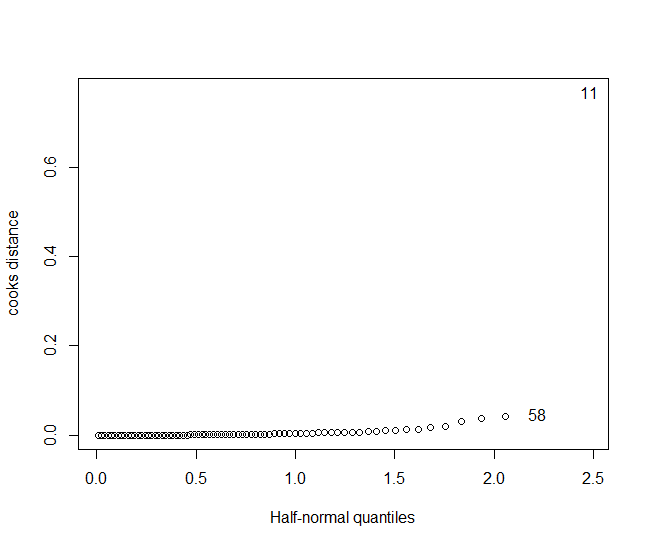
which(timf<0.0006)

خروجی داده 11 است. پس تنها داده ای که مقدار پی برای آن از 0.0006 کمتر است داده 11 است در نتیجه فرض اچ صفر یعنی پرت نبودن این داده رد میشود. پس داده 11 تنها داده پرت است و این داده در فضای پاسخ پرت است. درگر نقاط که پی انها از 0.0006 بیشتر است در فضای پاسخ پرت نیستند اما ممکن است در فضای ایکس پرت باشند. که البته دیدیم پرت نبودند.

حال باید بررسی کنیم ایا این نقاط پرت را حذف کنیم یا نه. یعنی ایا با حذف آن سطر و فیت کردن رگرسیون تفاوت پاسخ و براورد آن برای آن داده زیاد است یا نه.

cook=cooks.distance(mf)

halfnorm(cook,nlab=2,ylab="cooks distance")



داده 11 از بقیه خیلی دور است و باید حذف شود.

J)

vif(mf)

charge discharge depth temp endvolt

1.103904 1.045912 1.134237 1.055428 1.036999

خیر وابستگی وجود ندارد.

حال تمام مراحل را برای لگاریتم متغیر پاسخ انجام میدهیم. اگر هیستوگرام متغیر پاسخ را در نظر بگیریم نرمال نیست اما لگاریتم آن نرمال است.

mf\_logl=lm(log(cycles)~.,data=df)

summary(mf\_logl)

Coefficients:

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

(Intercept) -37.438324 23.776785 -1.575 0.1199

charge -1.211854 0.242443 -4.999 4.2e-06 \*\*\*

discharge -0.161780 0.078285 -2.067 0.0425 \*

depth 0.002927 0.083114 0.035 0.9720

temp 0.116640 0.010400 11.215 < 2e-16 \*\*\*

endvolt 20.510114 11.855827 1.730 0.0881 .

---

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

Residual standard error: 0.8761 on 69 degrees of freedom

Multiple R-squared: 0.7017, Adjusted R-squared: 0.6801

F-statistic: 32.46 on 5 and 69 DF, p-value: < 2.2e-16

میزان پی از الفا برابر با 0.05 کمتر است پس فرض اچ صفر یعنی صفر بودن بتا برای تمام متغیر ها رد میشود.

از طرفی بنابر مقدار پی برای هر یک از متغیر ها میبینیم که برای دو متغیر شارژ و دما مقدار پی از الفا کمتر است پس فرض اچ صفر یعنی صفر بودن ًریب ان متغیر در رگرسیون رد میشود.

برای متغیر های دیگر فرض اچ صفر را میپذیریم و در مدل رگرسیون ضرایب ان ها را صفر در نظر میگیریم.

توصیف ضرایب:

اگر میزان عرض از مبدا یک واحد افزایش پیدا کند، تعداد دفعات شارژ و دیشارژ به طور متوسط 34 دور کاهش می یابد.

اگرمیزان دیگر متغیر ها را ثبت در نظر بگیریم و میزان جریان لازم برای شارژ باتری در یک ساعت یک آمپر افزایش پیدا کند تعداد دفعات شارژ و دیشارژ به طور متوسط 1 دور کم میشود.

اگر میزان جریان لازم برای تخلیه شارژ باتری در یک ساعت یک آمپر افزایش پیدا کند تعداد دفعات شارژ و دیشارژ به طور متوسط 0.16 دور کم میشود.

اگر عمق باتری یک درصد افزایش پیدا کند تعداد دفعات شارژ و دیشارژ به طور متوسط تغییر زیادی نمیکند.

اگر دمای باتری یک درجه افزایش پیدا کند تعداد دفعات شارژ و دیشارژ به طور متوسط تغییر زیادی نمیکند.

اگر ولتاژ پایانی یک ولت افزایش پیدا کند تعداد دفعات شارژ و دیشارژ به طور متوسط 20 دور زیاد میشود.

چون قبلا واریانس ها ثابت نبود پس مدل فیت شده باید شامل تمام ویژگی ها باشد. در حقیقت نمیتوان به مقادیر پی اطمینان نمود و ویژگی ای را از مدل خارج کرد.

mf\_logl=lm(log(cycles)~.,data=df)

vif(mf\_logl)

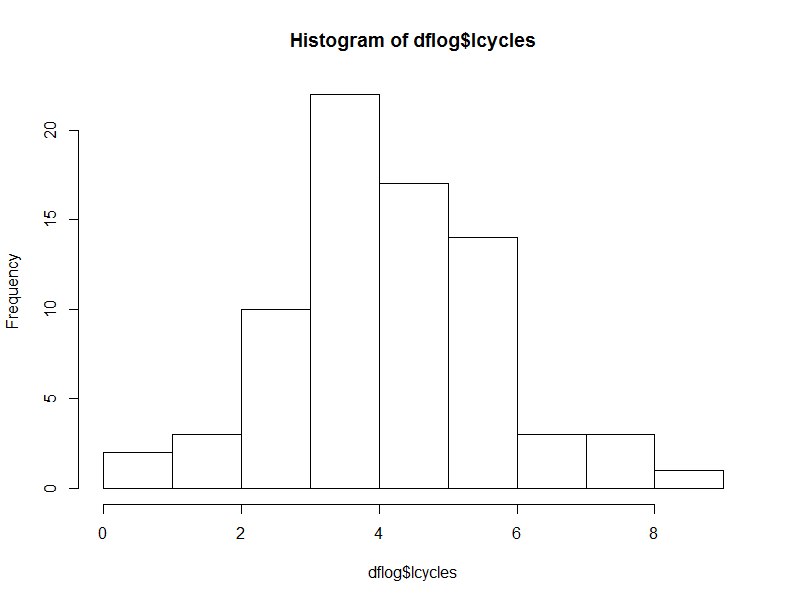
charge discharge depth temp endvolt

1.103904 1.045912 1.134237 1.055428 1.036999

متغیر های مستقل وابستگی ناچیزی با هم دارند.

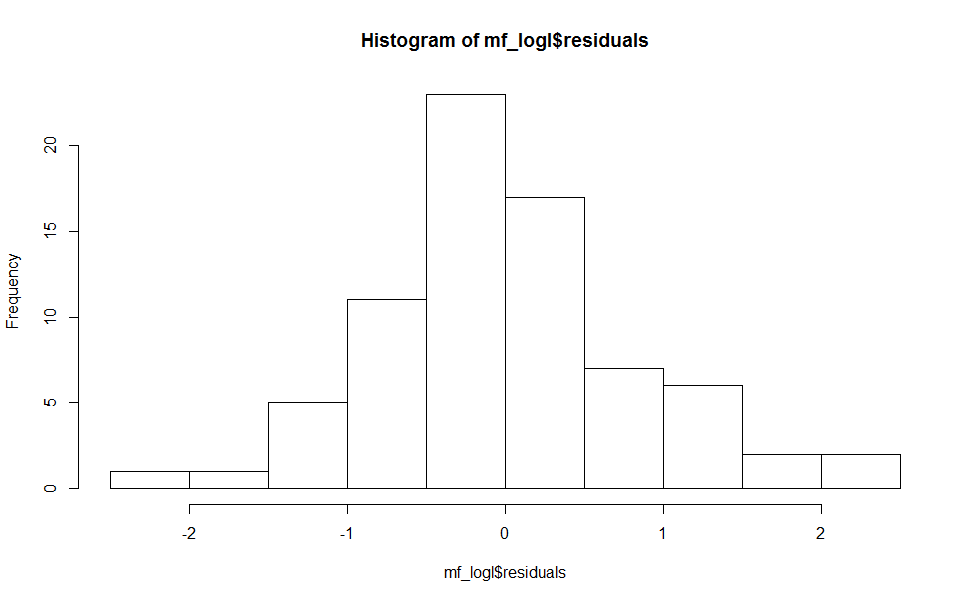
توصیف بتا ها به این معنا است که همه متغیر های مستقل به جز یکی (مانند m) را ثابت در نظر بگیریم. حال ببینیم با تغییر یک واحد در متغیر مستقل m، متوسط متغیر وابسته چقدر تغییر میکند. اگر بین متغیر های مستقل وابستگی وجود داشته باشد این توصیف هم دیگر درست نیست. چون اگر متغیر m را تغییر دهیم تا تغییر در متوسط متغیر پاسخ را بسنجیم، متغیر وابسته به ان نیز خود به خود تغییر میکند.

hist(log(df$cycles))



لگاریتم متغیر پاسخ نرمال شده. حال باید نرمال بودن و ثابت بودن واریانس خطا ها را بررسی کنیم.

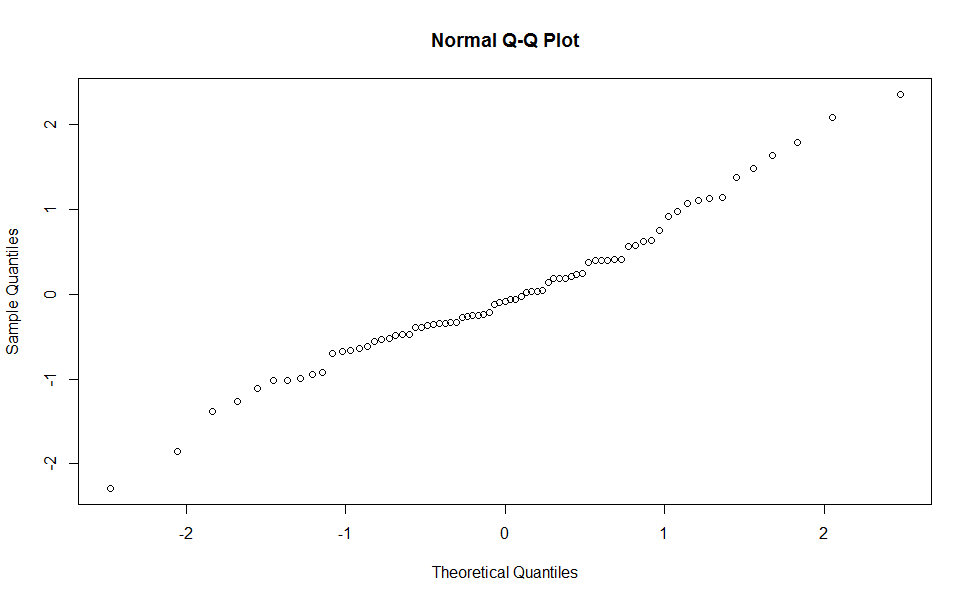
hist(mf\_logl$residuals)



نرمال است اما دو مرحله دیگر برای مطمین شدن را نیز انجام میدهیم.

library(nortest)

qqnorm(mf\_logl$residuals)



و در پایان تست اندرسون با فرض اچ صفر (نرمال بودن).

ad.test(mf\_logl$residuals)

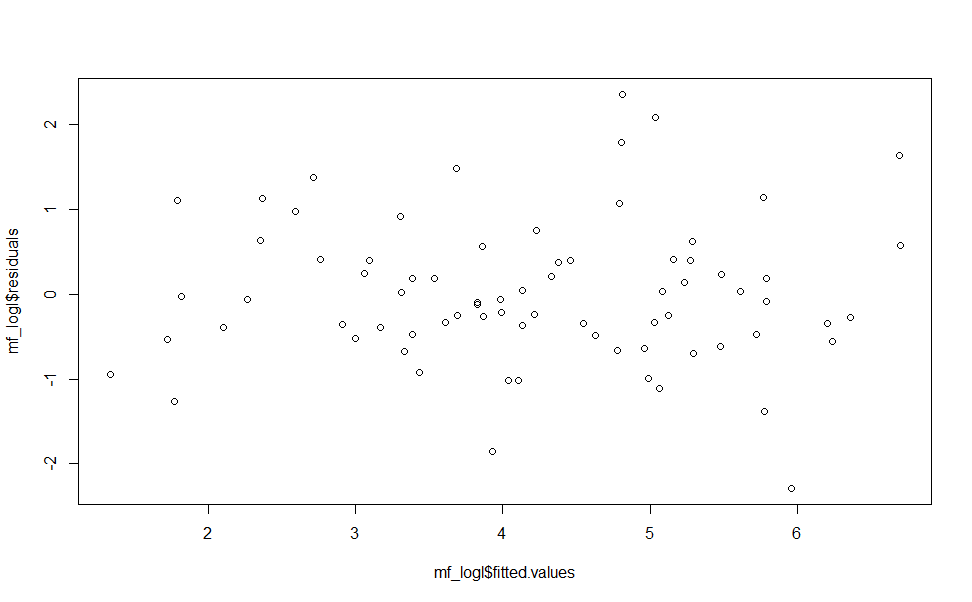
Anderson-Darling normality test

data: mf\_logl$residuals

A = 0.68843, p-value = 0.06936

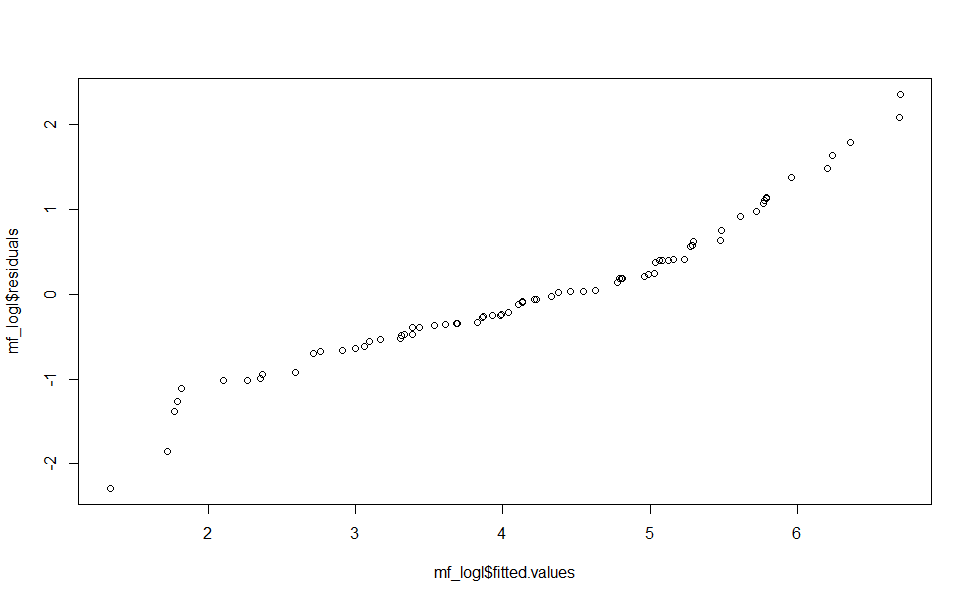
چون مقدار پی از الفا 0.05 بیشتر است فرض اچ صفر را میپذیریم.

plot(mf\_logl$fitted.values,mf\_logl$residuals)



پترن خاصی دیده نمیشود و به نظر میرسد خطاها تصادفی هستند.

qqplot(mf\_logl$fitted.values,mf\_logl$residuals)



حال ازمون فرض با فرض اچ صفر (عدم وجود رابطه بین خطاها و براورد متغیر پاسخ ) را بررسی میکنیم.

farz=lm(mf\_logl$residuals~mf\_logl$fitted.values)

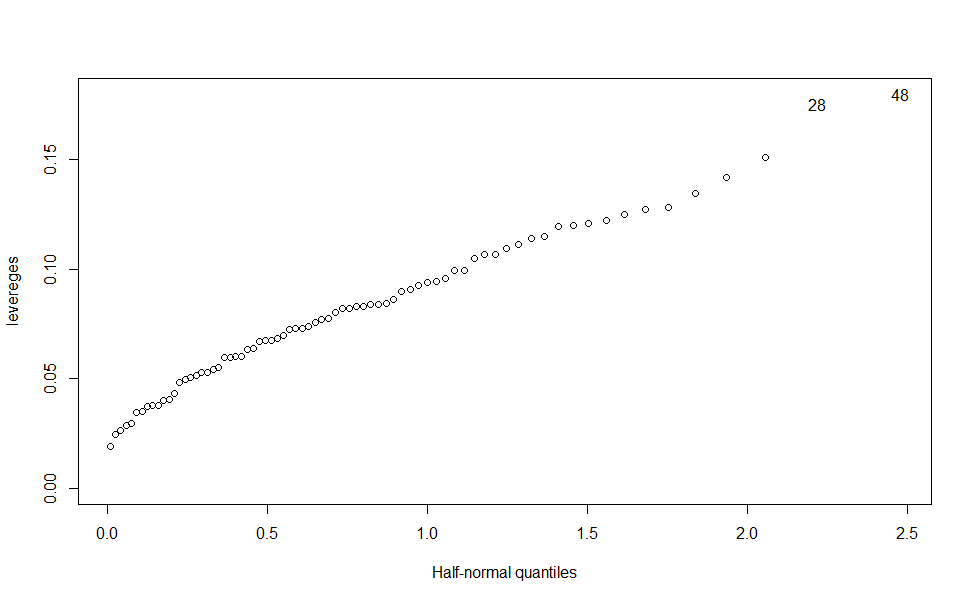
summary(farz)

مقدار پی برابر با یک است و از الفا بیشتر است پس فرض اچ صفر را میپذیریم.

بررسی نقاط پرت:

hlog=influence(mf\_logl)$hat

halfnorm(hlog,nlab=2,ylab="levereges")



باید بررسی کنیم ایا دو داده 28 و 48 در فضای ایکس پرت هستند یا نه.

which(hlog>3\*5/75)

output: named integer(0)

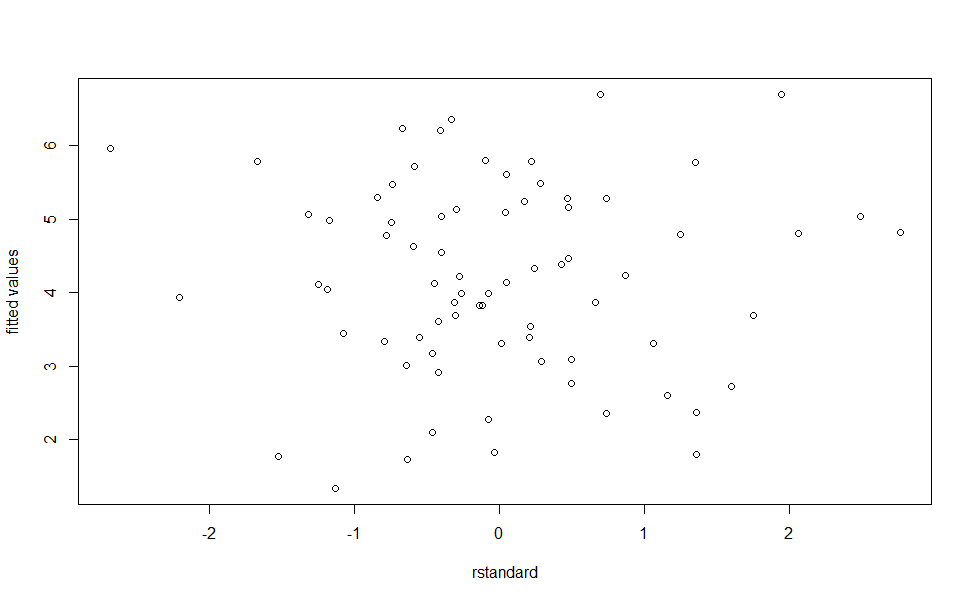
تعداد متغیر های مستقل در مدل فیت شده برابر با 5 و سایز نمونه برابر با 75 است.

راه دیگر بررسی واریانس هر یک از ویژگی ها با مقدار ان ویژگی در داده های خروجی بالا است) )

rstandard\_log=rstandard(mf\_logl)

plot(rstandard\_log,mf\_logl$fitted.values,xlab="rstandard",ylab="fitted val

ues")



which(abs(rstandard\_log)>3)

output: named integer(0)

پس نمیتوان به داده های پرتی که در روش های قبل بیان شد اطمینان نمود.

حال برای یافتن داده های پرت در فضای مدل اماره تی را بررسی میکنیم.

tlog=rstudent(mf\_logl)

#alpha=0.05/75=0.0006

T=function(x){

2\*(1-pt(abs(tlog[x]), df=75-5-1))

}

tilog=apply(as.matrix(seq(1,75)), 1, T)

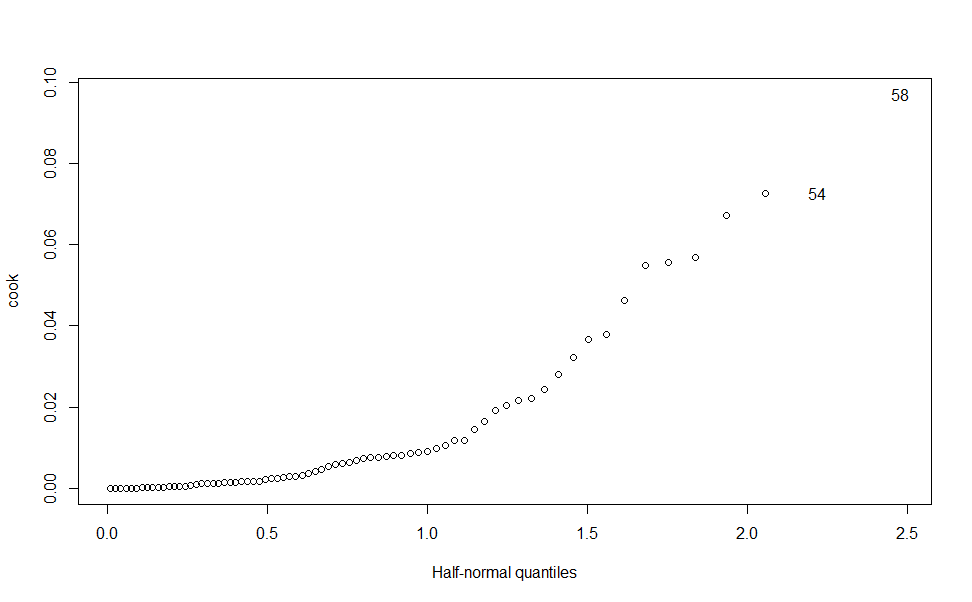
which(tilog<0.0006)

output: integer(0)

پس داده پرت نداریم.

cooklog=cooks.distance(mf\_logl)

halfnorm(cooklog,nlab=2,ylab="cook")



داده 58 نسبتا از بقیه دوتر است اما به داده های نزدیک 54 نزدیک است و چون تعداد نمونه کمه، این نقاط خودشون یک دسته هستند

چون در مراحل قبل نقطه پرتی نداشتیم داده ای را حذف نمیکنیم.

df58=df[-58,]

mf\_logl58=lm(log(cycles)~charge+temp,data=df58)

summary(mf\_logl)

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

23.776785

0.242443

0.078285

0.083114

0.010400

11.855827

summary(mf\_logl58)

23.552384

0.233056

0.075633

0.080003

0.009997

11.754311

تغییر زیادی در واریانس ها بعد از حذف این داده ایجاد نشده. در کل داده پرتی نداریم و نیاز به حذف نیست.