**Advanced Data Analysis HW6**

**Yingxin Chen**

**Yc2910**

1. Fit a multiple linear regression model:

The multiple linear regression model is :

> summary(fit)

Call:

lm(formula = glu ~ npreg + bp + skin + bmi + age)

Residuals:

Min 1Q Median 3Q Max

-80.398 -16.558 -2.406 18.508 87.546

Coefficients:

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

(Intercept) 48.83674 16.19545 3.015 0.002909 \*\*

npreg -0.49202 0.77323 -0.636 0.525318

bp 0.34774 0.20179 1.723 0.086425 .

skin 0.07238 0.24281 0.298 0.765959

bmi 0.68362 0.45451 1.504 0.134183

age 0.86912 0.25333 3.431 0.000735 \*\*\*

---

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

Residual standard error: 29.34 on 194 degrees of freedom

Multiple R-squared: 0.163, Adjusted R-squared: 0.1414

F-statistic: 7.556 on 5 and 194 DF, p-value: 1.664e-06

1. Validation
2. Linearity/functional form, including the need for any interaction terms

We first have a look at scatter plot of the data



**We can see, only age and bp show some trend with glu.**

**The Multiple R-squared: 0.163, Adjusted R-squared: 0.1414 also demonstrate that the multiple linear model is very poor to fit data.**

We can also see it from the plot of residuals vs x and y:



**We can see the residuals are so significant in all variables, especially it have extreme strong positive relationship with y, thus the data cannot fit this fumction form.**

We next take a look at the needs for interaction forms.

> summary(fit2)

Call:

lm(formula = glu ~ npreg + bp + skin + bmi + age + npreg \* bp +

npreg \* skin + npreg \* bmi + npreg \* age + bp \* skin + bp \*

bmi + bp \* age + skin \* bmi + skin \* age + bmi \* age)

Residuals:

Min 1Q Median 3Q Max

-81.229 -17.800 -2.989 17.870 83.412

Coefficients:

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

(Intercept) -45.928649 84.496284 -0.544 0.5874

npreg -8.968891 7.527588 -1.191 0.2350

bp 0.073327 1.281906 0.057 0.9544

skin -1.042493 2.068301 -0.504 0.6148

bmi 5.128752 3.334424 1.538 0.1257

age 4.869970 2.368012 2.057 0.0411 \*

npreg:bp 0.069404 0.080618 0.861 0.3904

npreg:skin 0.054525 0.085594 0.637 0.5249

npreg:bmi 0.115094 0.186581 0.617 0.5381

npreg:age -0.053040 0.068524 -0.774 0.4399

bp:skin 0.006464 0.025531 0.253 0.8004

bp:bmi 0.005558 0.046474 0.120 0.9049

bp:age -0.008049 0.023931 -0.336 0.7370

skin:bmi -0.027181 0.036612 -0.742 0.4588

skin:age 0.034460 0.015267 2.257 0.0252 \*

bmi:age -0.133438 0.058664 -2.275 0.0241 \*

---

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

Residual standard error: 29.17 on 184 degrees of freedom

Multiple R-squared: 0.2152, Adjusted R-squared: 0.1512

F-statistic: 3.364 on 15 and 184 DF, p-value: 5.088e-05

From the second model which consider all the interaction factors, we can see age, age\*skin and age\*bmi are significant independent variables. We can check with Anova test.

Analysis of Variance Table

Response: glu

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

npreg 1 5803 5802.9 6.8176 0.0097705 \*\*

bp 1 10923 10922.7 12.8326 0.0004360 \*\*\*

skin 1 4312 4311.6 5.0655 0.0255901 \*

bmi 1 1357 1357.5 1.5949 0.2082312

age 1 10134 10134.0 11.9060 0.0006939 \*\*\*

npreg:bp 1 576 576.1 0.6768 0.4117406

npreg:skin 1 893 893.5 1.0497 0.3069173

npreg:bmi 1 202 201.6 0.2368 0.6270775

npreg:age 1 876 875.5 1.0286 0.3118097

bp:skin 1 46 45.6 0.0535 0.8172681

bp:bmi 1 138 137.8 0.1619 0.6879040

bp:age 1 414 413.8 0.4862 0.4865064

skin:bmi 1 828 827.8 0.9726 0.3253352

skin:age 1 2041 2041.4 2.3984 0.1231769

bmi:age 1 4404 4403.7 5.1738 0.0240845 \*

Residuals 184 156614 851.2

---

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

**We can see only bmi and age have an interaction effect (F= 5.17, P<0.05)**

1. Normality

We can check Q-Q plot and plot of residuals and fitted value.



**Neither plot indicates a significant normality assumption.**

> ks.test(fit$resi, "pnorm")

One-sample Kolmogorov-Smirnov test

data: fit$resi

D = 0.4946, p-value < 2.2e-16

alternative hypothesis: two-sided

> shapiro.test(fit$resi)

Shapiro-Wilk normality test

data: fit$resi

W = 0.9856, p-value = 0.03907

**The Kolmogorov-Smirnov test and shapiro.test tests the NULL hypothesis that the samples came from a Normal distribution. This means p-value <= 0.05, then you would reject the NULL hypothesis that the samples came from a Normal distribution.**

1. Homoscedasticity

> gqtest(fit)

Goldfeld-Quandt test

data: fit

GQ = 1.4776, df1 = 94, df2 = 94, p-value = 0.02995

> bptest(fit)

studentized Breusch-Pagan test

data: fit

BP = 4.0722, df = 5, p-value = 0.5391

**The null hypothesis for the GQ and BP test is homoskedasticity. The larger the F-statistic, the more evidence you’ll have against the homoskedasticity assumption and the more likely you have heteroskedasticity. From GQ, the residuals are heteroskedasticity but from BP, it is homoscedasticity.**

**We plot then:**



**We can see heteoskedasticity.**

1. Uncorrelated error

> dwtest(fit)

Durbin-Watson test

data: fit

DW = 2.1802, p-value = 0.9021

alternative hypothesis: true autocorrelation is greater than 0

**p > 0.05 provides no evidence of correlation.**

1. Check for outliers and influential points.



> h <- hatvalues(fit)

> a <- 2\*(5+1)/200

> a

[1] 0.06

> which(h >= a)

9 11 13 14 36 80 111 116 157 187 188 190

**So, there are several outliers with respect to x.**

> tlist<-studres(fit)

> which(abs(tlist)>=qt(1-0.1/(2\*200),200-5-1))

named integer(0)

**So there is no outlier with respect to y.**

1. Remedial measures
2. For lack of functional form, we can try transformations and non-linear model.
3. For lack of normality, we can try transformation and robust regression methods.
4. For heteoskedasiticity, we can try transformation or build variance structure into model.
5. For outliers with respect to x, we can delete those outliers.

**Codes:**

glu <- Pima.tr$glu

npreg <- Pima.tr$npreg

bp <- Pima.tr$bp

skin <- Pima.tr$skin

bmi <- Pima.tr$bmi

age <-Pima.tr$age

fit <- lm(glu ~ npreg + bp + skin + bmi + age)

summary(fit)

pairs(~glu+npreg + bp + skin + bmi + age, main= 'scatter plot of data')

anova(fit)

par(mfrow=c(2,3))

plot(fit$resi~age, main = 'resi vs age')

plot(fit$resi~bp, main = 'resi vs bp')

plot(fit$resi~npreg, main = 'resi vs npreg')

plot(fit$resi~skin, main = 'resi vs skin')

plot(fit$resi~bmi, main = 'resi vs bmi')

plot(fit$resi~glu, main = 'resi vs glu')

fit2 <- lm(glu ~ npreg + bp + skin + bmi + age + npreg\*bp + npreg\*skin + npreg\*bmi + npreg\*age + bp\*skin + bp\*bmi + bp\*age + skin\*bmi + skin\*age + bmi\*age)

summary(fit2)

anova(fit2)

par(mfrow=c(1,2))

qqnorm(fit$resi)

plot(fit$fitted, fit$resi , xlab="Fitted",ylab='Residuals')

ks.test(fit$resi, "pnorm")

shapiro.test(fit$resi)

gqtest(fit)

bptest(fit)

dwtest(fit)

h <- hatvalues(fit)

a <- 2\*(5+1)/200

which(h >= a)

tlist<-studres(fit)

which(abs(tlist)>=qt(1-0.1/(2\*200),200-5-1))