## 

Do you know body fat?

04.26.2017

Multiple linear regression model project

Group3

# 

# Problem1

# Since we have BMI(Body Mass Index) = weight/height^2, we discard these two variables Height (Ht) and Weight (Wt). Our response variable is % Body fat (\_Bfat), predictors are Red cell count (RCC), White cell count (WCC), Hematocrit (Hc), Hemoglobin (Hg), Plasma ferritin concentration (Ferr), BMI (Body mass index), Sum of skinfolds (SSF) and Lean body mass (LBM).

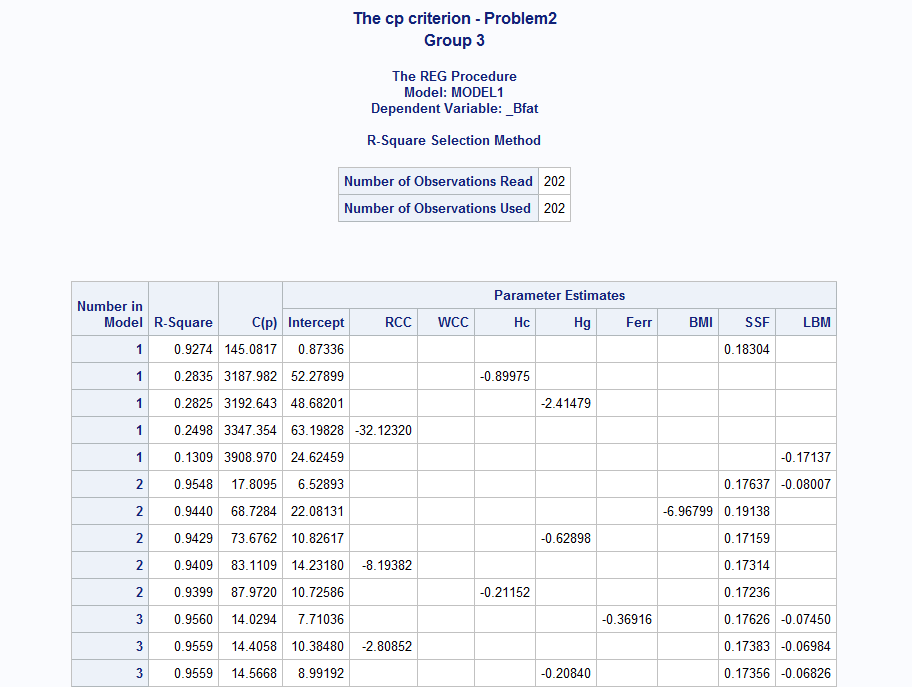
## Below are the plots of each of the predictors

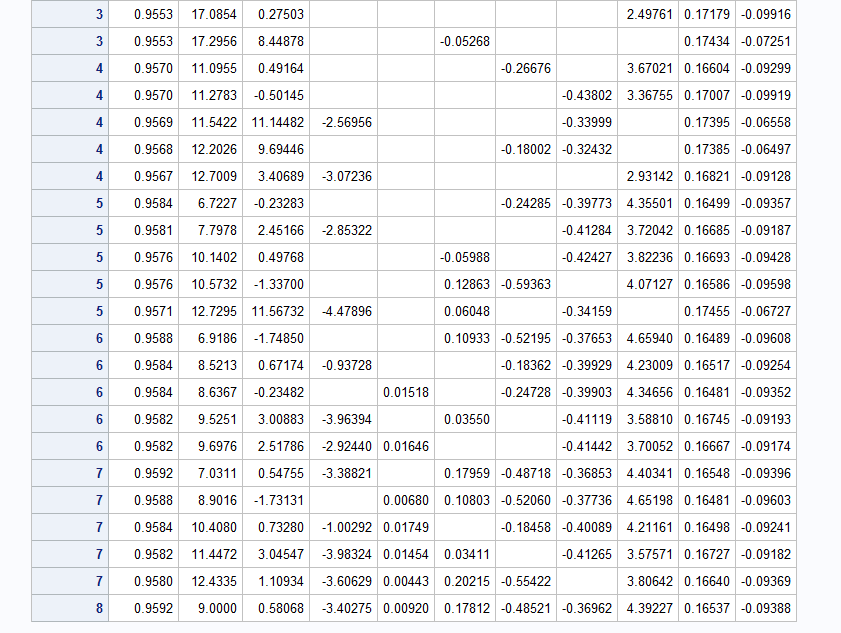
|  |  |
| --- | --- |
| 1.1.PNG | 1.2.PNG |
| 1.3.PNG | 1.4.PNG |
| 1.5.PNG | 1.6.PNG |
| 1.8.PNG | 1.7.PNG |

As we can see in the output, some plots have approximately linear relationship and others do not appear to have a specific relationship with the response variable. Thus, no transformation was needed.

# Problem2

## The Cp criterion was used to select the best subset of variables for the data. All 10 predictors are shown in the table below.





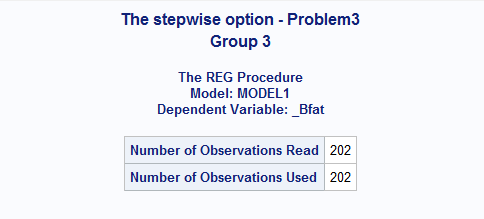
We can see that the R-value of all these subsets are all very close to 1. The best subset of the data is the one with 6 predictors and an R- square value of 0.9588 and a C(p) value of 6.9186. Since it has the smallest C(p) value and is simple compare with other feasible subset.

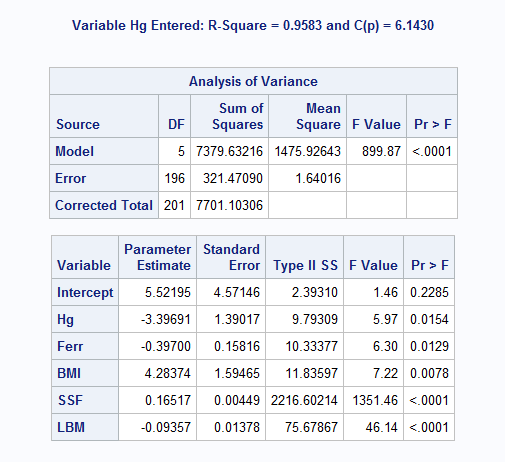
We have our linear regression model: bodyfat = -1.74850+ 0.10933\*Hc -0.52195\* Hg -0.37653\* Ferr + 4.65940 \* BMI + 0.16489\* SSF -0.09608\* LBM

## 

# Problem 3

The stepwise option was used to report the best subset of variables for the data. All 5 predictors are shown in the table below.  This subset has the value of C(p) = 6.1430





We have our linear regression model :

bodyfat = 5.52195 -3.39691\* Hg -0.39700\* Ferr + 4.28374\* BMI + 0.16517\* SSF -0.09357\* LBM

# Problem 4

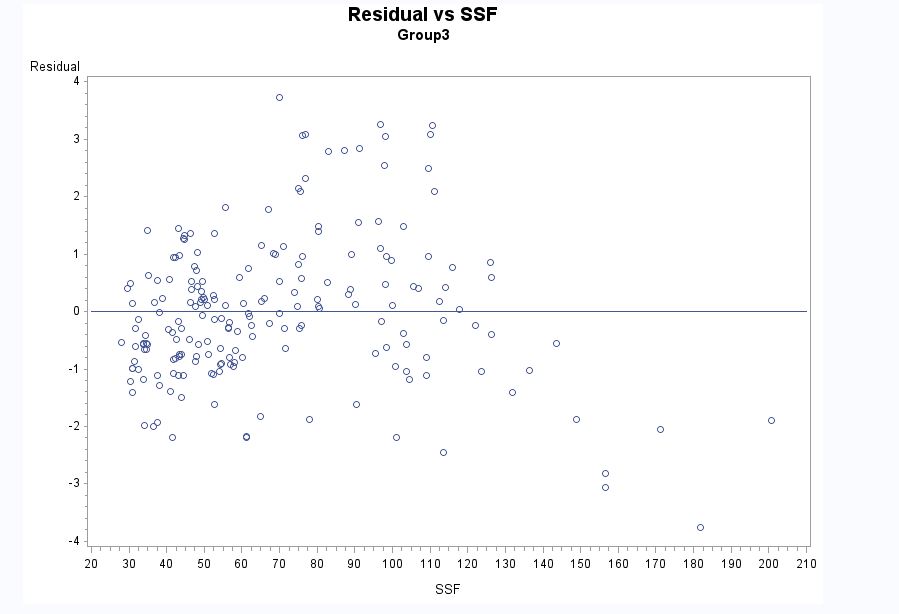
The final model we used is the model from problem 3 and the assumptions were checked the below.

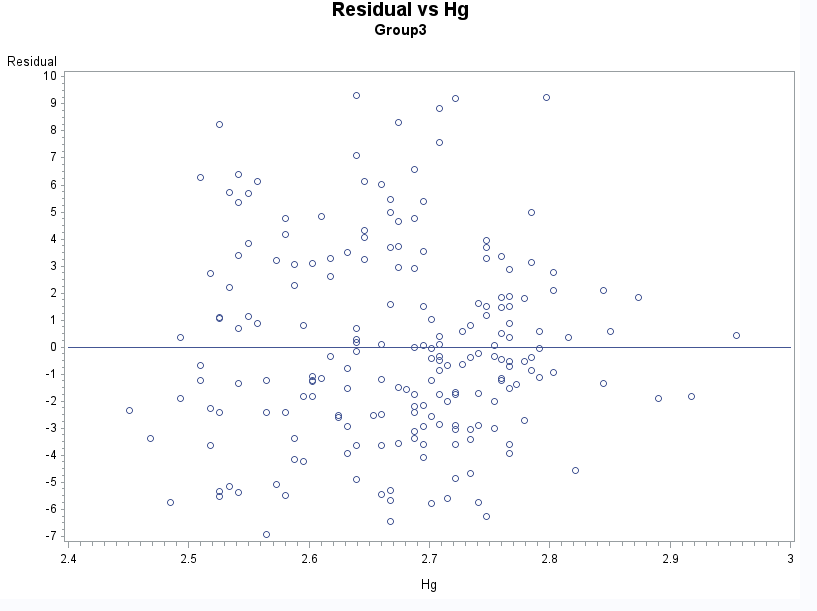
bodyfat = 5.52195 -3.39691\* Hg -0.39700\* Ferr + 4.28374\* BMI + 0.16517\* SSF -0.09357\* LBM

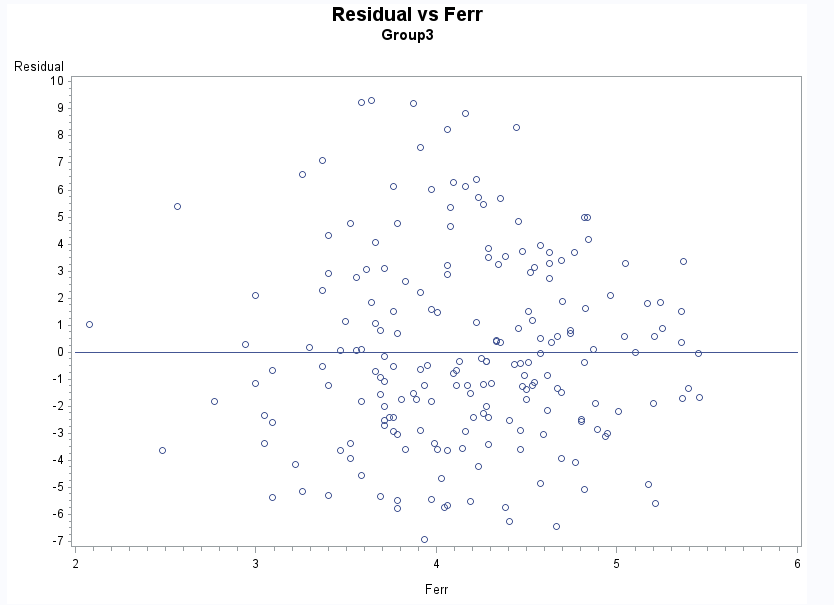
## Linear relationship

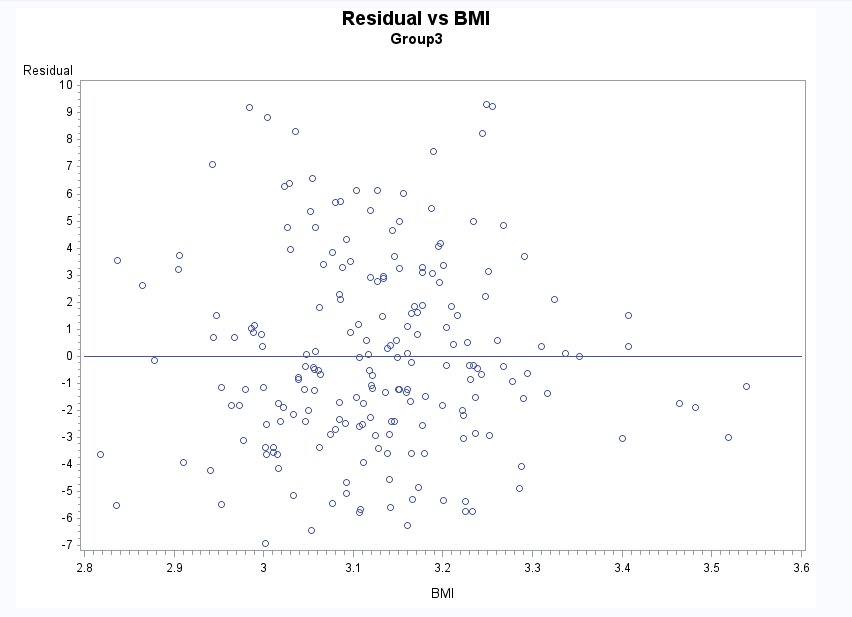
From problem 1, we could know response variables vs each predictors have linear relationship. Therefore, the linear relationship is satisfied.

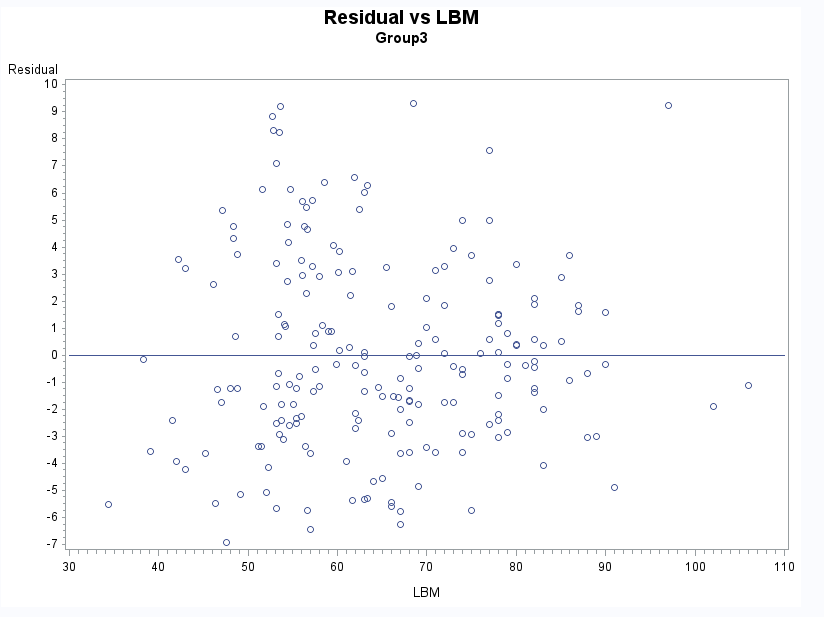
## Constant variance





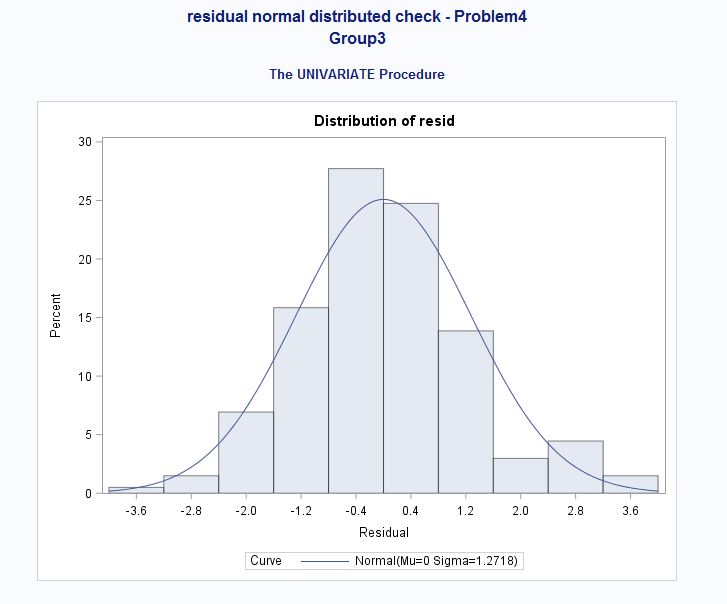


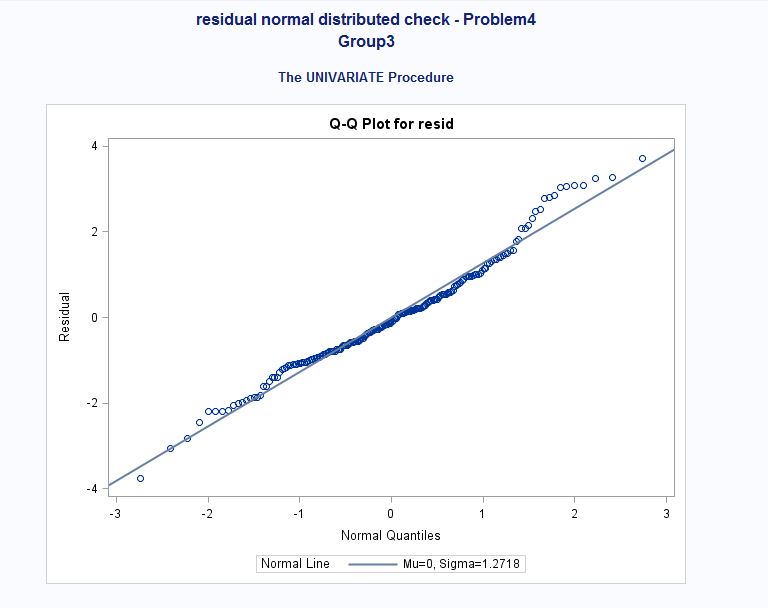




From the residual vs each predictors’ plots, we can see that there is no obvious pattern and there is a constant variance. Therefore, we think the data obey the constant variance assumption.

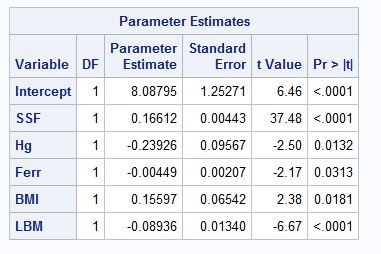
## Residual normally distributed





According to the histogram, the residual is roughly normally distributed because the curve is bell shaped. Also, based on the QQ plot, the dots are in a line that appears to be slightly right-tailed, but approximately normal.

## Independence



According to the SAS output, the predictors are independent. Therefore, the independence assumption is satisfied.

# Problem 5

Examining other diagnostics of the same “best” model:

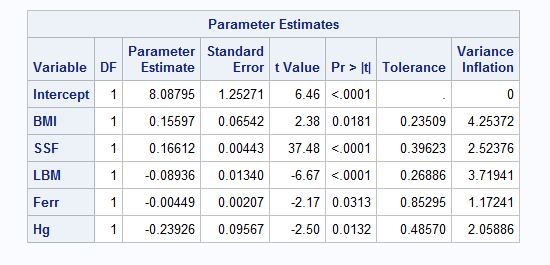
bodyfat = 5.52195 -3.39691\* Hg -0.39700\* Ferr + 4.28374\* BMI + 0.16517\* SSF -0.09357\* LBM

The partial residual plots for the five predictors (RCC, Hg, Ferr, BMI, LBM) are shown in the figures below. There are no outliers.

|  |  |
| --- | --- |
| SSFpartial.JPG |  |
|  |  |
|  |  |

**Summary of Tests**

|  |  |  |
| --- | --- | --- |
| Diagnostic Test | Significant Values | Conclusion |
| Partial Residual Plots | All | All predictors are of value |
| Student Residual | All values lie within the interval except for 1 | 1 outlier |
| Cook’s D | the 1 outlier is impactful | Need to take further investigate for that outlier |
| Tolerance/VIF | All variance inflation values are less than 10 | No excessive multicollinearity |



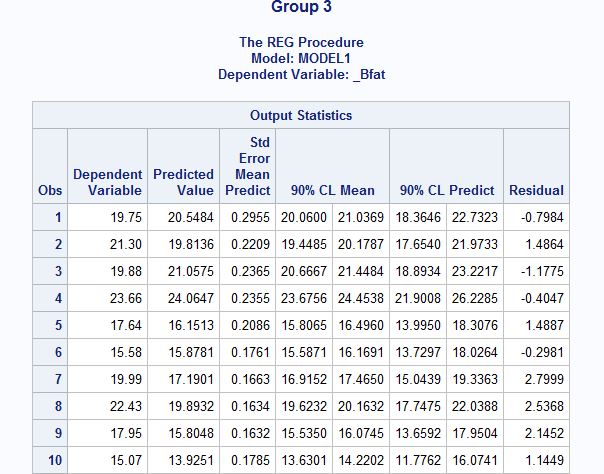
# Problem 6

1. Equation of the regression model

bodyfat = 5.52195 -3.39691\* Hg -0.39700\* Ferr + 4.28374\* BMI + 0.16517\* SSF -0.09357\* LBM

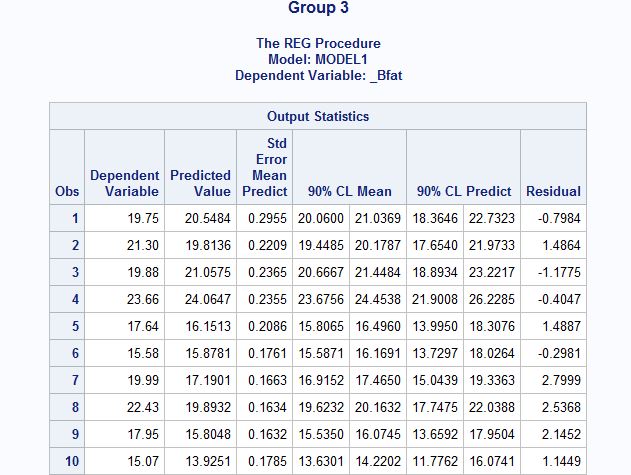
1. 90% confidence interval for the mean of the response variable

First ten observations shown below.

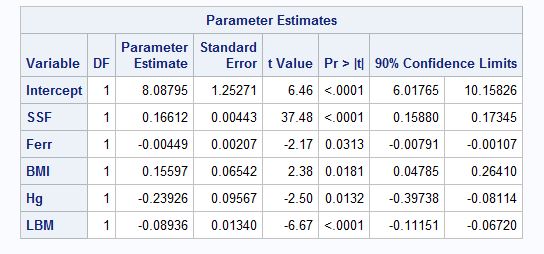


1. 90% prediction interval for the mean of the response variable

First ten observations shown below.



1. 90% confidence intervals for the regression coefficients.



**---------------------------------------------------SAS code:-------------------------------------------------------**

\*Response variables: %bodyfat(\_Bfat)

\*Predictor:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Ht |  | Height in cm |
|  | Wt |  | Weight in kg |
|  | LBM |  | Lean body mass |
|  | RCC |  | Red cell count |
|  | WCC |  | White cell count |
|  | Hc |  | Hematocrit |
|  | Hg |  | Hemoglobin |
|  | Ferr |  | Plasma ferritin concentration |
|  | BMI |  | Body mass index = weight/height^2 |
|  | SSF |  | Sum of skin folds |

----------------------------------------------------------------\*1----------------------------------------------------------

----------------------------------------------\*plot response variablesl vs predictor--------------------------------

data newbodyfat;

set work.bodyfat;

proc sort data= newbodyfat;

by RCC;

title1 '%Bodyfat vs RCC - Question 1 ';

title2 'Group3';

symbol1 v=circle i=sm80;

proc gplot data=newbodyfat;

plot \_Bfat \* RCC;

run;

proc sort data= newbodyfat;

by WCC;

title1 '%Bodyfat vs WCC - Question 1 ';

title2 'Group3';

symbol1 v=circle i=sm80;

proc gplot data=newbodyfat;

plot \_Bfat \* WCC;

run;

proc sort data= newbodyfat;

by Hc;

title1 '%Bodyfat vs Hc - Question 1 ';

title2 'Group3';

symbol1 v=circle i=sm80;

proc gplot data=newbodyfat;

plot \_Bfat \* Hc;

run;

proc sort data= newbodyfat;

by Hg;

title1 '%Bodyfat vs Hg - Question 1 ';

title2 'Group3';

symbol1 v=circle i=sm80;

proc gplot data=newbodyfat;

plot \_Bfat \* Hg;

run;

proc sort data= newbodyfat;

by Ferr;

title1 '%Bodyfat vs Ferr - Question 1 ';

title2 'Group3';

symbol1 v=circle i=sm80;

proc gplot data=newbodyfat;

plot \_Bfat \* Ferr;

run;

proc sort data= newbodyfat;

by BMI;

title1 '%Bodyfat vs BMI - Question 1 ';

title2 'Group3';

symbol1 v=circle i=sm80;

proc gplot data=newbodyfat;

plot \_Bfat \* BMI;

run;

proc sort data= newbodyfat;

by SSF;

title1 '%Bodyfat vs SSF - Question 1 ';

title2 'Group3';

symbol1 v=circle i=sm80;

proc gplot data=newbodyfat;

plot \_Bfat \* SSF;

run;

proc sort data= newbodyfat;

by LBM;

title1 '%Bodyfat vs LBM - Question 1 ';

title2 'Group3';

symbol1 v=circle i=sm80;

proc gplot data=newbodyfat;

plot \_Bfat \* LBM;

run;

----------------------------------------------------------------\*2----------------------------------------------------------

-----------------------------------------------------------\*Cp criterion-------------------------------------------------

title1 'The cp criterion - Problem2';

title2 'Group 3';

proc reg data= newbodyfat;

model \_Bfat = RCC WCC Hc Hg Ferr BMI SSF LBM /selection= rsquare cp b best=5;

run;

----------------------------------------------------------------\*3----------------------------------------------------------

.

-------------------------------------------------------------\*Stepwise--------------------------------------------------

title1 'The stepwise option - Problem3';

title2 'Group 3';

proc reg data=newbodyfat;

model \_Bfat = RCC WCC Hc Hg Ferr BMI SSF LBM  / selection=stepwise;

run;

----------------------------------------------------------------\*4----------------------------------------------------------

proc reg data=newbodyfat;

model \_Bfat = SSF Hg Ferr BMI LBM ;

output out =diag r=resid p = pred;

Run;

----------------------------------------------------------\*residual plot--------------------------------------------------

title1 'Residual vs SSF';

title2 'Group3';

proc sort data=diag;

by SSF;

symbol1 v=circle i=rl;

proc gplot data=diag;

plot resid\*SSF;

Run;

title1 'Residual vs Hg';

title2 'Group3';

proc sort data=diag;

by Hg;

symbol1 v=circle i=rl;

proc gplot data=diag;

plot resid\*Hg;

Run;

title1 'Residual vs Ferr';

title2 'Group3';

proc sort data=diag;

by Ferr;

symbol1 v=circle i=rl;

proc gplot data=diag;

plot resid\*Ferr;

Run;

title1 'Residual vs BMI';

title2 'Group3';

proc sort data=diag;

by BMI;

symbol1 v=circle i=rl;

proc gplot data=diag;

plot resid\*BMI;

Run;

title1 'Residual vs LBM';

title2 'Group3';

proc sort data=diag;

by LMB;

symbol1 v=circle i=rl;

proc gplot data=diag;

plot resid\*LBM;

Run;

-----------------------------------------------\*QQ plot and histogram-----------------------------------------------

title1 'residual normal distributed check - Problem4';

title2 'Group3';

proc univariate data=diag plot normal;

var resid;

qqplot resid /normal(L=1 mu=est sigma=est);

histogram resid /normal(L=1 mu=est sigma=est);

Run;

----------------------------------------------------------------\*5----------------------------------------------------------

------------------------------------------------------\*partial residual plots--------------------------------------------

title1 'partial residual plot - question 5';

title2 'SSF partial residual plot';

title3 'group 3';

proc reg data=newbodyfat;

model \_Bfat SSF = Hg Ferr BMI LBM ;

output out = partialSSF r=residBfat residSSF;

symbol1 v=circle i=rl;

axis1 label=('SSF');

axis2 label=(angle=90 'Body fat');

proc gplot data=partialSSF;

plot residBfat\*residSSF/ haxis = axis1 vaxis=axis2 vref=0;

Run;

title1 'partial residual plot - question 5';

title2 'Hg partial residual plot';

title3 'group 3';

proc reg data=newbodyfat;

model \_Bfat Hg = Rcc Ferr BMI LBM ;

output out = partialHg r=residBfat residHg;

symbol1 v=circle i=rl;

axis1 label=('Hg');

axis2 label=(angle=90 'Body fat');

proc gplot data=partialHg;

plot residBfat\*residHg/ haxis = axis1 vaxis=axis2 vref=0;

Run;

title1 'partial residual plot - question 5';

title2 'Ferr partial residual plot';

title3 'group 3';

proc reg data=newbodyfat;

model \_Bfat Ferr = RCC Hg BMI LBM ;

output out = partialFerr r=residBfat residFerr;

symbol1 v=circle i=rl;

axis1 label=('Ferr');

axis2 label=(angle=90 'Body fat');

proc gplot data=partialFerr;

plot residBfat\*residFerr/ haxis = axis1 vaxis=axis2 vref=0;

Run;

title1 'partial residual plot - question 5';

title2 'BMI partial residual plot';

title3 'group 3';

proc reg data=newbodyfat;

model \_Bfat BMI = RCC Hg Ferr LBM ;

output out = partialBMI r=residBfat residBMI;

symbol1 v=circle i=rl;

axis1 label=('BMI');

axis2 label=(angle=90 'Body fat');

proc gplot data=partialBMI;

plot residBfat\*residBMI/ haxis = axis1 vaxis=axis2 vref=0;

Run;

title1 'partial residual plot - question 5';

title2 'LBM partial residual plot';

title3 'group 3';

proc reg data=newbodyfat;

model \_Bfat LBM = RCC Hg Ferr BMI ;

output out = partialLBM r=residBfat residLBM;

symbol1 v=circle i=rl;

axis1 label=('LBM');

axis2 label=(angle=90 'Body fat');

proc gplot data=partialLBM;

plot residBfat\*residLBM/ haxis = axis1 vaxis=axis2 vref=0;

Run;

----------------------------------------------------------\*r influence----------------------------------------------

proc reg data=newbodyfat;

model \_Bfat = SSF Hg Ferr BMI LBM / r influence;

output out = diag1 r = resid1;

run;

---------------------------------------------------------\*Tolerance/VIF-----------------------------------------

proc reg data=newbodyfat;

model \_Bfat = BMI SSF LBM Ferr Hg / tol vif;

run;

----------------------------------------------------------------\*6----------------------------------------------------------

title1 "90% confidence interval and prediction interval - Problem 6";

title2 "Group 3";

proc reg data=newbodyfat alpha=0.1;

model \_Bfat= SSF Ferr BMI Hg LBM/ clb cli clm;

run;