Introduction

In this project, the data is from www.sci.usq.edu.au. The data give the body fat, triceps skinfold thickness, thigh circumference and mid arm circumference for twenty healthy females aged from 20 to 34. The body fat persons was obtained by a cumbersome and expensive procedure requiring the immersion of the person in water. It would therefore be helpful if a regression model with some or all of these predictor variables could provide reliable predictions of the amount of body fat, since the measurements needed for the predictor variables are easy to obtain. Then variables are as follow：

Y：Body Fat

X_1：Triceps skinfold thickness

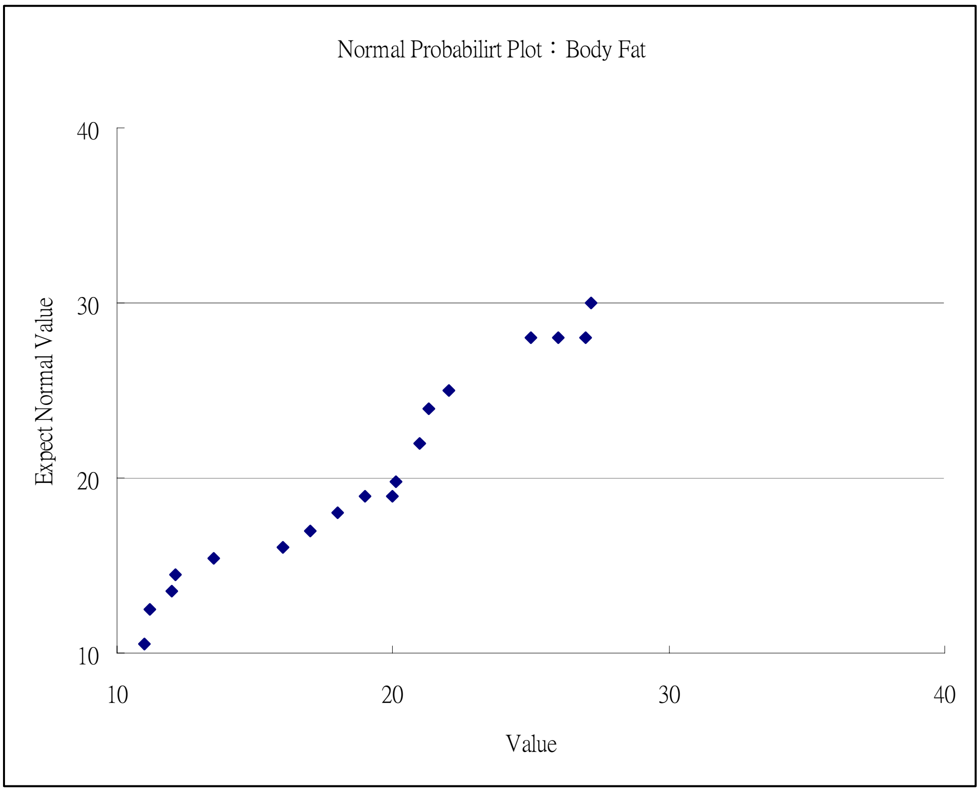
X_2：Thigh circumference

X_3：Midarm circumference

Analysis

According to the assumption of the regression analysis, the dependent variable's distribution is normal distribution. Q-Q plot (Fig. 1) shows that the variable(Y) is fitted normail distribution, so we can use regression analysis.

Figure. 1body fat Q-Q plot



From Table 1 the correlation between X_1 (triceps) and X_2 (thigh) exceeds 0.90, the problem of multicollinearity maybe exist in the model. In order to modify multicollinearity, the regression model needs to drop one or more predictor variables that are not fitted for model. Because the correlation between $X_1$and X_2 is high, I just try to remove one or both of the variables (X_1,X_2) from the model, and use the regression analysis.

Table1 Correlations between two variables

	Y	X_1	X_2	X_3
Y	1			
X_1	0.843265	1		
X_2	0.87809	0.923843	1	
X_3	0.14244	0.457777	0.084667	1

Suppose the regression functions are as follow：

Model 1 $Y=\beta_0+\beta_2X_2+\beta_3X_3$ (remove X_1)

Y=body fat

X_2=Thigh

X_3=Midarm

Hypothesis H_0：$\beta_2=\beta_3=0$

Ha：$\beta_2\neq0$ or $\beta_3\neq0$

Statistic method ：F Test

ANOVA

Model #1	df	SS	MS	F	Singificant
Regression	2	384.2797	192.199	29.39775	3.03E-0.6
Residual	17	111.1098	6.535869		
Total	19	495.3895			

Conclusion

$F=MSR/MSE=29.39775 > F_{0.05}(2,17)$，p-value<0.05

The result rejects Ho. The regression function is：

$Y=-25.997+0.850882X_2+0.0960292X_3$

Model #1	Coefficient β	Std. Error	t	P-Value
Constant	-25.997	6.997321	-3.71527	0.00172
X_2	0.850882	0.112448	7.566874	7.72E-07
X_3	0.096029	0.161393	0.595005	0.559678

Model #1	
R	0.880745
R Square	0.775712
Adjusted R Square	0.749325
Std. Error	2.556535

Model #2 $Y=\beta_0+\beta_1X_1+\beta_3X_3$ (remove X_2)

Y=Body fat

X_1=Triceps

X_3=Mid arm

Hypothesis H_0 : $\beta_1=\beta_3=0$

H_a : $\beta_1\neq0$ or $\beta_3\neq0$

Statistic method : F Test

ANOVA

Model #2	df	SS	MS	F	Singificant
Regression	2	389.4553	194.7277	31.24932	2.02E-0.6
Residual	17	105.9342	6.231422		
Total	19	495.3895			

Conclusion

$F=MSR/MSE=31.249325 > F_{0.05}(2,17)$, p-value<0.05

The result rejects H_0. The regression function is :

$Y=6.791627+1.00058X_1-0.43144X_3$

Model #2	Coefficient β	Std. Error	t	P-Value
Constant	6.791627	4.488287	1.513189	0.1486
X_1	1.000585	0.128232	7.802921	5.12E-07
X_3	-0.43144	0.176616	-2.44283	0.025786

Model #2	
R	0.886657
R Square	0.78616
Adjusted R Square	0.761002
Std. Error	2.496282

Model #3 $Y=\beta_0+\beta_3X_3$ (remove X_1 and X_2)

Y=Body fat

X_3=Midarm

Hypothesis H_0 : $\beta_3=0$

H_a : $\beta_3\neq0$

Statistic method　: F Test

ANOVA

Model #3	df	SS	MS	F	Singificant
Regression	1	10.0516	10.0516	0.372789	0.54912
Residual	18	485.3379	26.96322		
Total	19	495.3895			

Conclusion

$F=MSR/MSE=0.372789 < F_{0.05}(2,17)$，p-value>0.05

The result don't reject H_0. There is no linear relationship between X_3 and Y.

Model #3	Coefficient β	Std. Error	t	P-Value
Constant	14.68678	9.095926	1.614655	0.123778
X_3	0.199429	0.32663	0.610565	0.54912

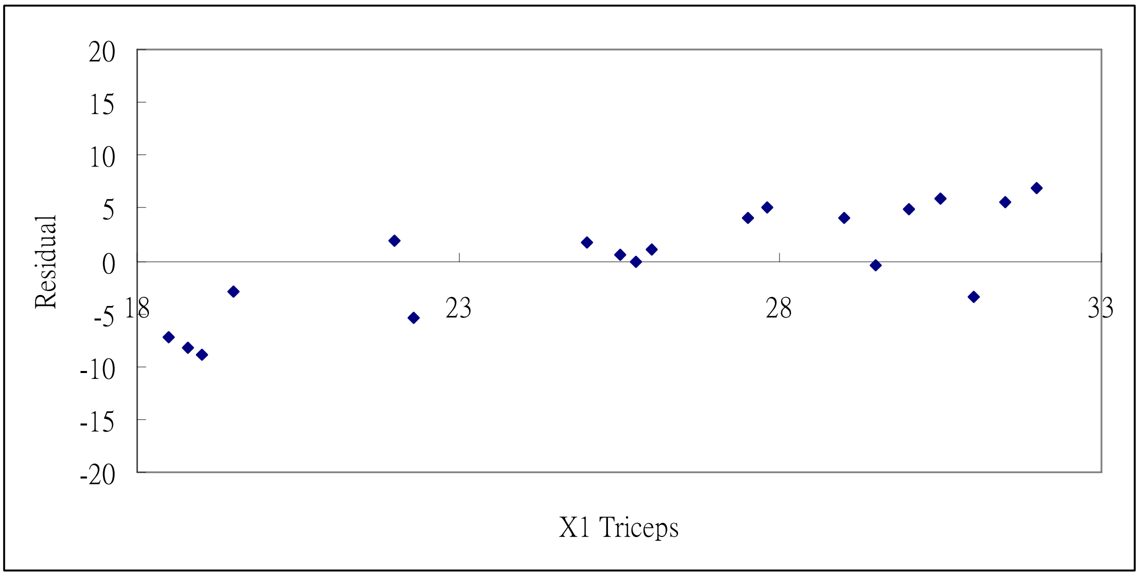
Model #3	
R	0.142444
R Square	0.02029
Adjusted R Square	-0.03414
Std. Error	5.192612

Comparison

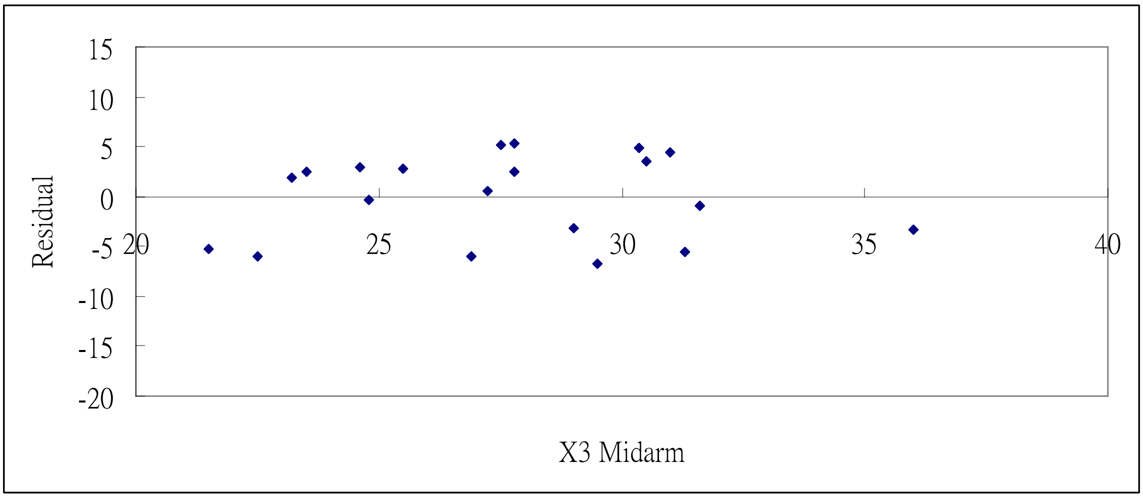
From the conclusion of the regression analysis for Model #3, there is no linear relationship between X_3 and Y. Therefore, Model #3 is not a good model to predict th body fat, so I just compare the remaining models (Models #1 and Model #2) to determine which one is better to predict the body fat. At first, I find both R-Square and Adj R-Sq of Model #2 are greater than Model #1. Furthermore, the parameters of the regression function for Model #2 are both significant, but the parameter of the regression function for Model #1 are not. Therefore, it can state that the regression Model #2 is better.

<u>Diagnostic</u>

At residual to independent variable X_1 Plot, the residuals are randomly scattered alone with the zero axis and the deviation all fall into the interval (-10,10).It shows the residual is independent to variableX_1.



At residual to independent variable X_2 Plot, the residuals are randomly scattered alone with the zero axis and the deviation all fall into the interval (-10,10).It shows the residual is independent to variable X_2.



Conclusion

According to the introduction, triceps skinfold thickness, thigh circumference and midarm circumference may affect the body fat. However, in the regression analysis, the problem of multicollinearity may exist. So, we try to remove one or more variables from the model and to modify multicollinearity. Then do some comparisons and diagnostic. The final result suggests that we can use the regression Model #2(as shows bellow) to predict the body fat.

Y(Body fat)=6.791627+1.000585X_1(Triceps)-0.43144X_3(Midarm)