**Problem 1 - (20 points)**

Open the Cereals dataset in SAS and use SAS to perform the following.

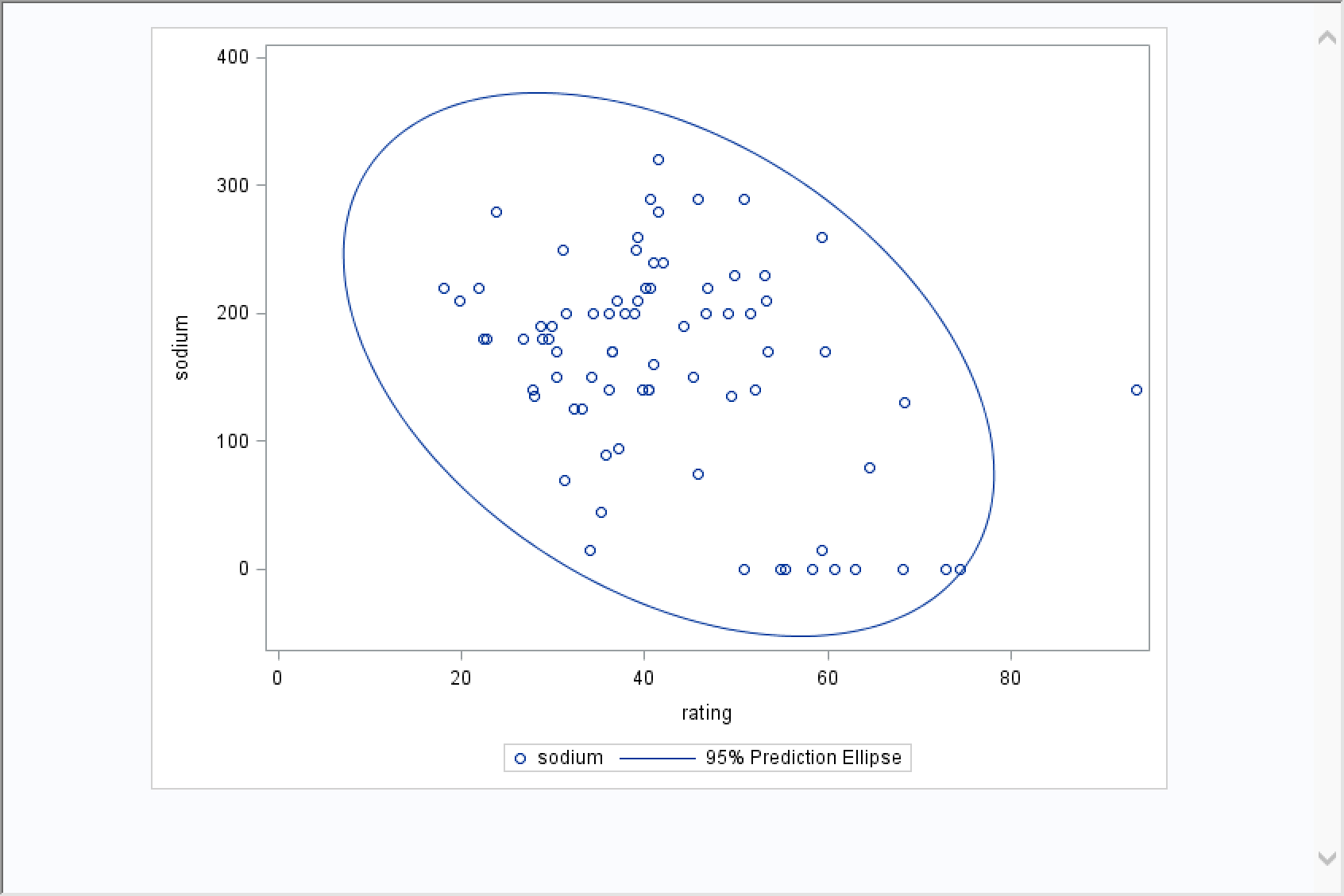
1. We are interested in predicting the rating based on sodium content. Construct the appropriate scatter plot.

**proc** **sgplot** data=cereal\_ds;

scatter x=rating y=sodium;

ellipse x=rating y=sodium;

**run**;



1. Based on the scatter plot, is there strong evidence of a linear relationship between the variables.

According the scatter plot ,for example, when the rating is 40, the sodium’s range is like between 0 and 300, that shows there is not a strong evidence of a linear relationship between the variables.

1. Perform the appropriate regression.

title2 "Simple Regression Analysis";

**proc** **reg** data = cereal\_ds outest = est\_cereal\_ds;

model rating = sodium / stb vif dwProb dw;

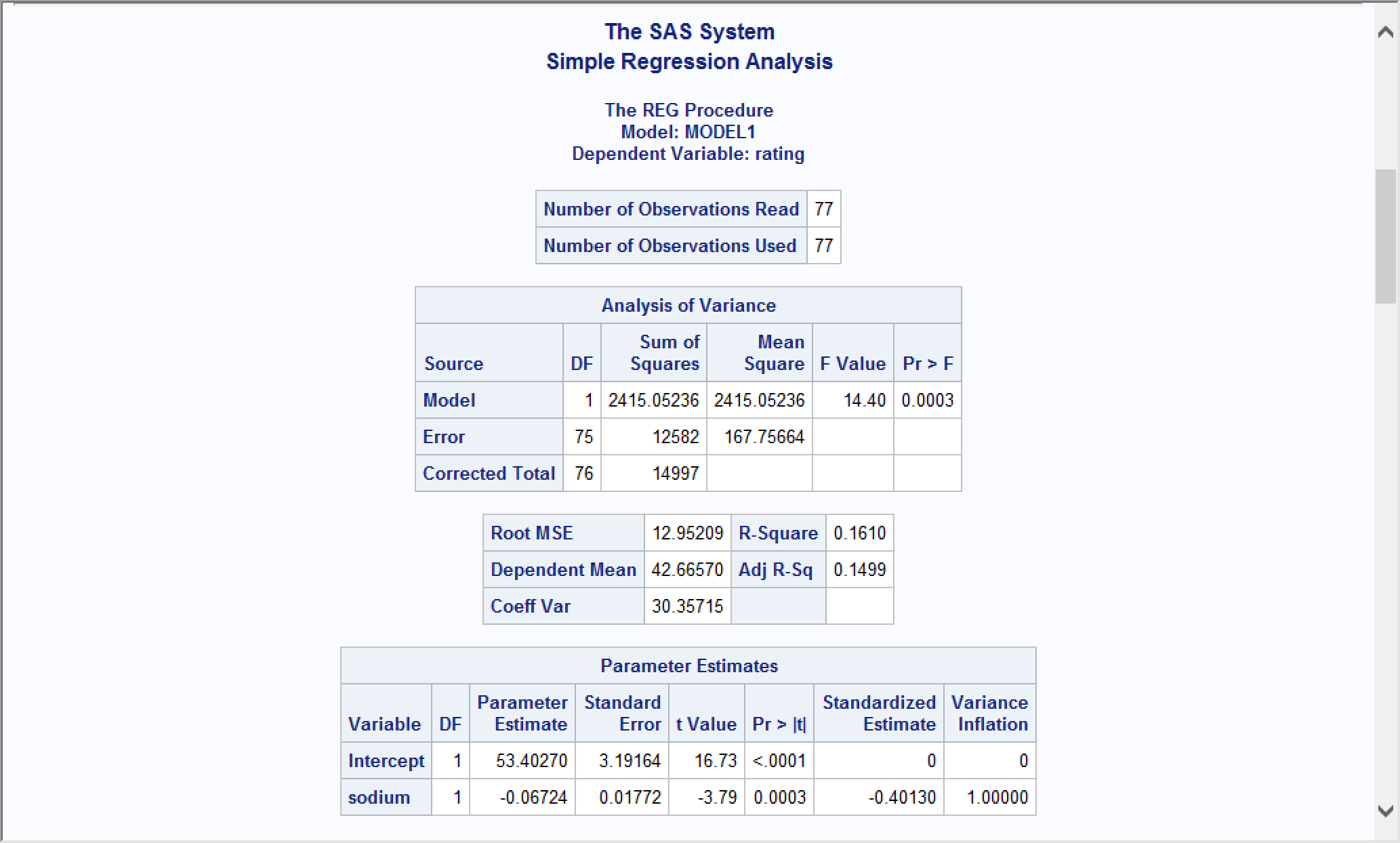
output out = reg\_ceralds\_OUT predicted = P\_rating residual = R\_rating

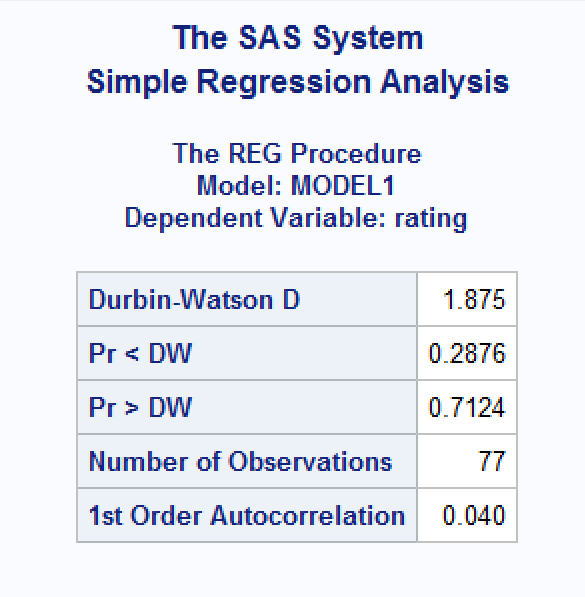
L95M = l95 U95M = U95M L95 = L95 U95 = U95

h = lev cookd = Cookd dffits = dffit;

**quit**;

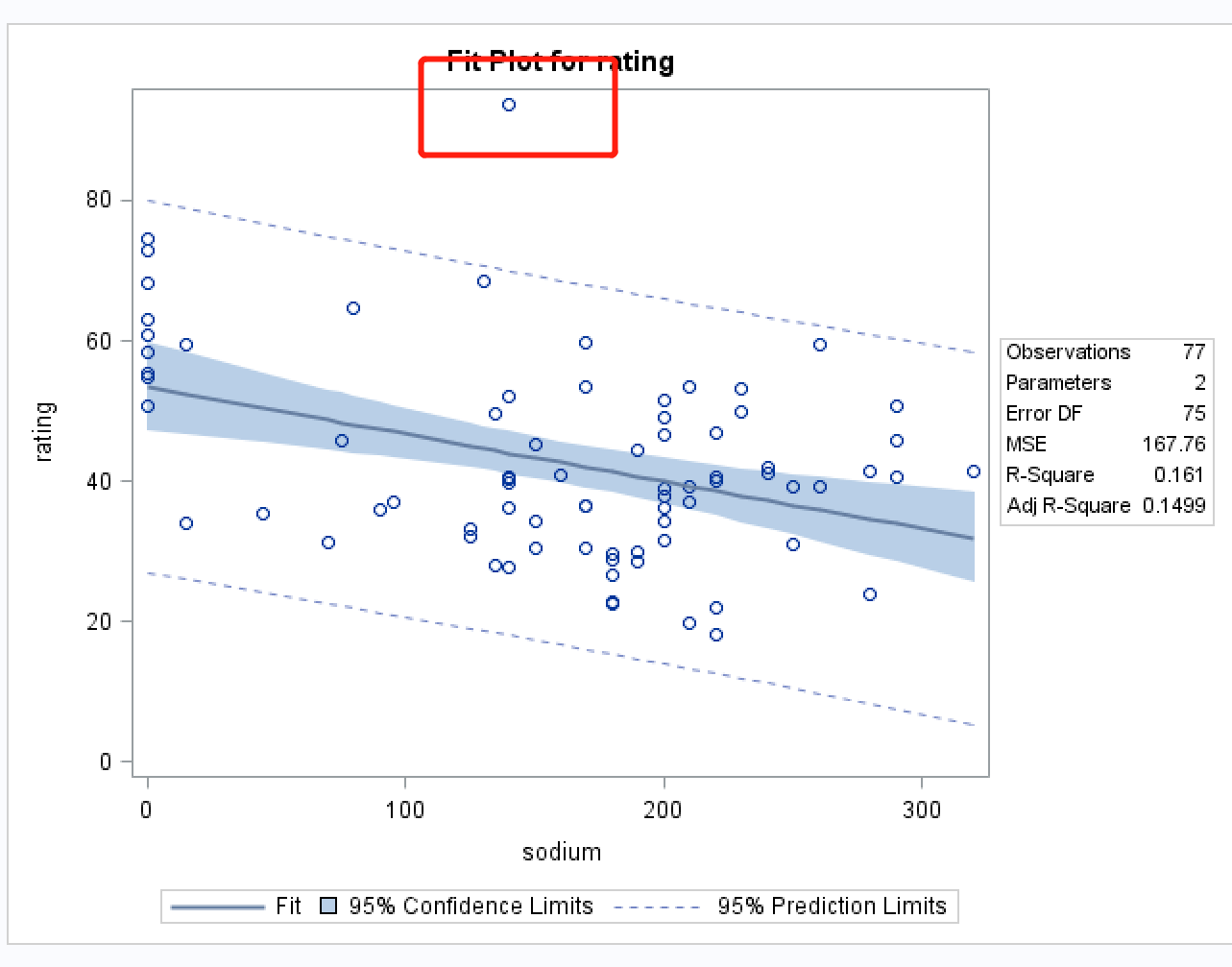
The result report:





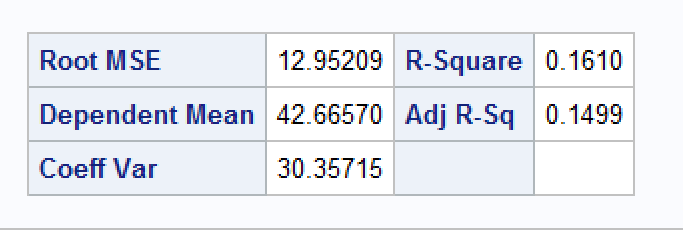
Therefore, the simple regression mode should be:

1. Which cereal is an outlier?



According the picture, the one’s rating beyond 90% is outlier, that is **All-Bran\_with\_Extra\_Fiber**

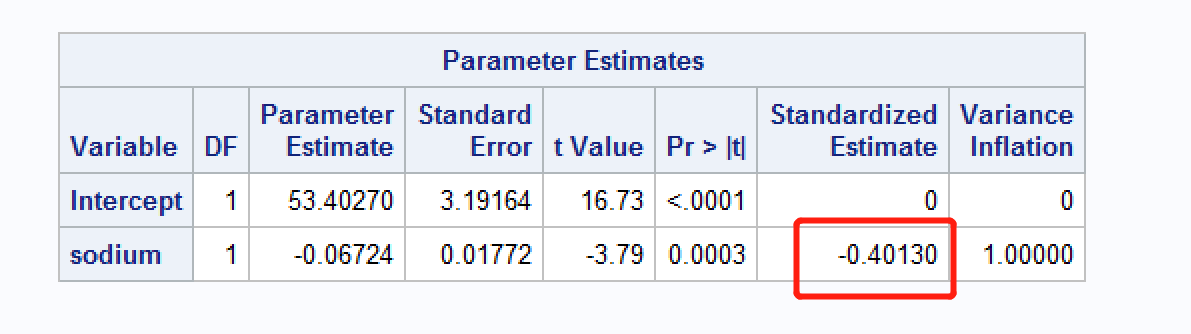
1. What is the typical error in predicting rating based on sodium content?

The typical error in predicting rating based on sodium content is root MSE, the value is 12.95209.

1. Interpret the y- intercept.

The y-intercept means that the estimated value for the response variable when the predictor variable equals zero. if sodium is zero, the rating is 53.40270. and the **Pr > |t|** <.0001, which means the y-intercept value of rating is significant.

1. Is there a significant relationship between the two variables? Why?



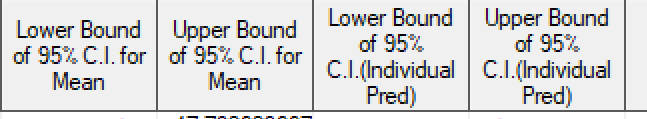
The standardized estimate value of the rating and sodium is -0.40130, which shows there is a weak negative relationship between these two variables .

1. What is the meaning of the slope?

The Slope is -0.06724, which means that when the sodium increase for every 100, the rating will decrease for 6.7 overall.

1. Construct a 95% confidence interval for the average nutrition rating of all cereals with sodium content of 100.

With the parameter “L95M = l95 U95M = U95M L95 = L95 U95 = U95”, we got the value of following parameter:



And we build four regression models with the parameters and the sodium.

So for the average nutrition rating of all cereals with sodium:

Lower Bound of 95% C.I. for Mean = 48.49985 -0.06150 \* sodium

Upper Bound of 95% C.I. for Mean= 58.30556 - 0.07298 \*sodium

With the sodium content of 100, we got the confidence interval is from 42.35 to 51.

1. Construct a 95% confidence interval for the nutrition rating for a randomly chosen cereal with sodium content of 100.

According to the previous question , we got the regression model:

Lower Bound of 95% C.I.(Individual Pred) = 27.09109 -0.06613 \* sodium

Upper Bound of 95% C.I.(Individual Pred) = 79.71432 -0.06836 \* sodium

With the sodium content of 100, we got the confidence interval is from 20.48 to 72.88

**Problem 2 - (20 points)**

Use the following ANOVA table of a regression model, produced by a software package, to answer the questions below.

|  |  |  |
| --- | --- | --- |
| **Source** | **DF** | **Sum of** |
| **Squares** |
| **Model** | 2 | 253.274 |
| **Error** | 147 | 6897.42 |

From the table, we have SSR =253.274,SSE=6897.42, dfr=2, dfe=147

1. How many observations are used for this model?

N= P+E+1 = 147+2+1=150

1. How many parameters are used in this model?

Y= a0+a1x0+a2x1

There are three parameters are used in this model.

1. What is the R-square for this model?

R-square = SSR/(SSE+SSR)= 253.274/(253.274+6897.42)= 3.541%

1. What is the F value of the model?

MSR= 253.274/2=126.637

MSE=6897.42/147=46.9212

Fstat = MSR / MSE = 2.6989

1. Is the F value of the model significant at 5% level of significance? Why?

Critical value of F test : **3.05762065**

**Since Fstat < 3.05762065 so we fail to reject the null hypothesis.**

**Problem 3 - (30 points)**

The forced expiratory volume in one second (FEV1) measurement shows the amount of air a person can forcefully exhale in one second. Using the Lung dataset in CANVAS, develop a regression model for predicting FEV1 of the father using weight and height of the father as predictors.

The SAS code:

/\*question 3 code\*/

**proc** **copy** in=sasdata out=work;

select lung;

**run**;

title2 "Multiple Regression Analysis";

**proc** **reg** data = lung outest=lung\_est;

model FEV1\_father = Weight\_father Height\_father / stb vif dwProb dw;

OUTPUT OUT= reg\_lungOUT predicted = P\_FEV1father residual = R\_FEV1father

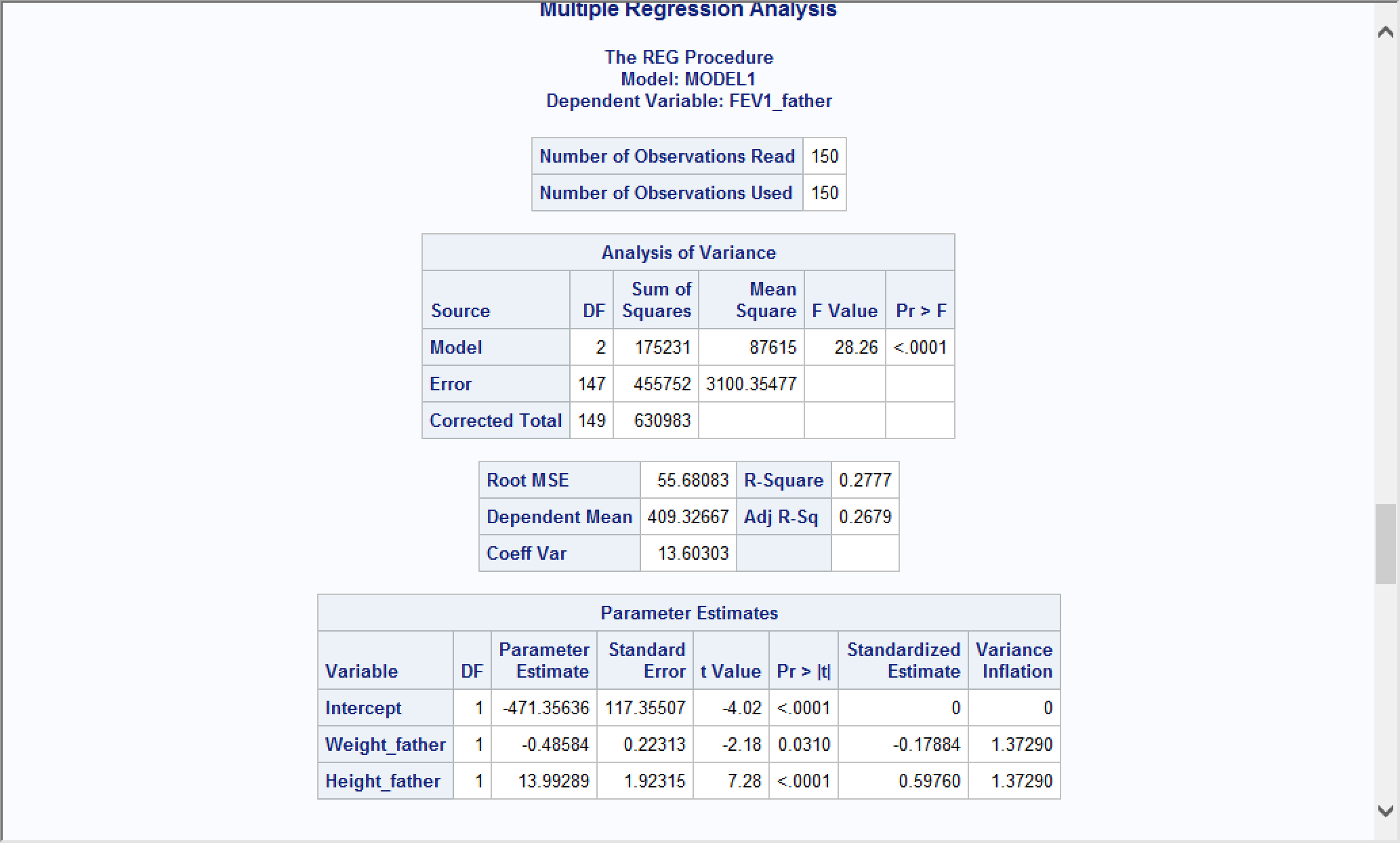
L95M = l95 U95M = U95M L95 = L95 U95 = U95 rstudent=rstudent

h= lev cookd= Cookd dffits= dffit;

**run**;

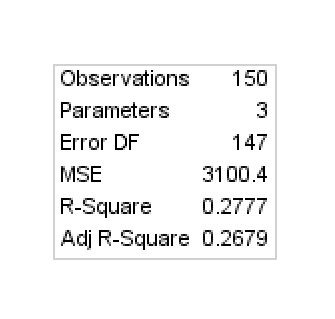
**quit**;

The output result:



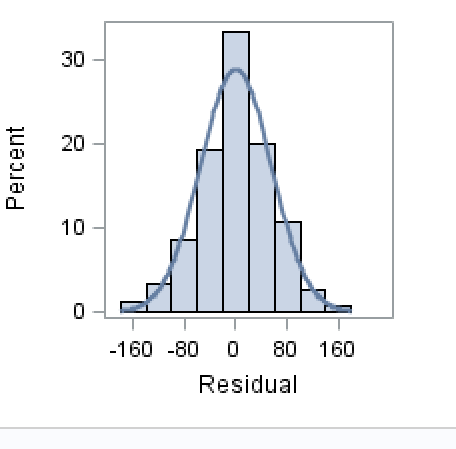
The Regression Mode:

1. Is this a good model? Why?



No, the R-square is only 0.2777 and the adj R-Square is 0.2679, both are even less the 50%., it does not look like a good model. but according to the null hypothesis, the p-value of the regression mode is < 0.0001, 0.0310, < 0.0001, they are satisfied by the threshold of the rejection of the null hypothesis 0.05,which means there is a relationship among variables, but their relationship is not strong.

1. Are the residuals normally distributed? Why?

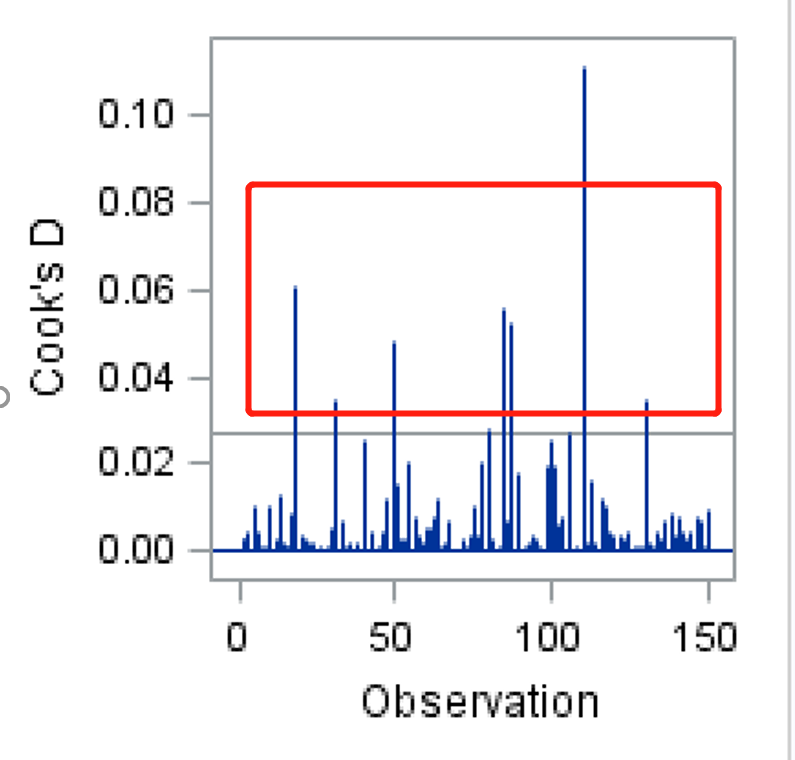


Yes, the residuals can be seen as a normally distributed graph.

1. What are the influential observations? Why?

Cook's distance or Cook's D is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis.

If an observation’s Cook Distance is great than 4 / (# sample – # predictor variable – 1), that means it is an influential observation. For this instance, we choose( 4/(150-1-1))



The Code:

\*Cookd;

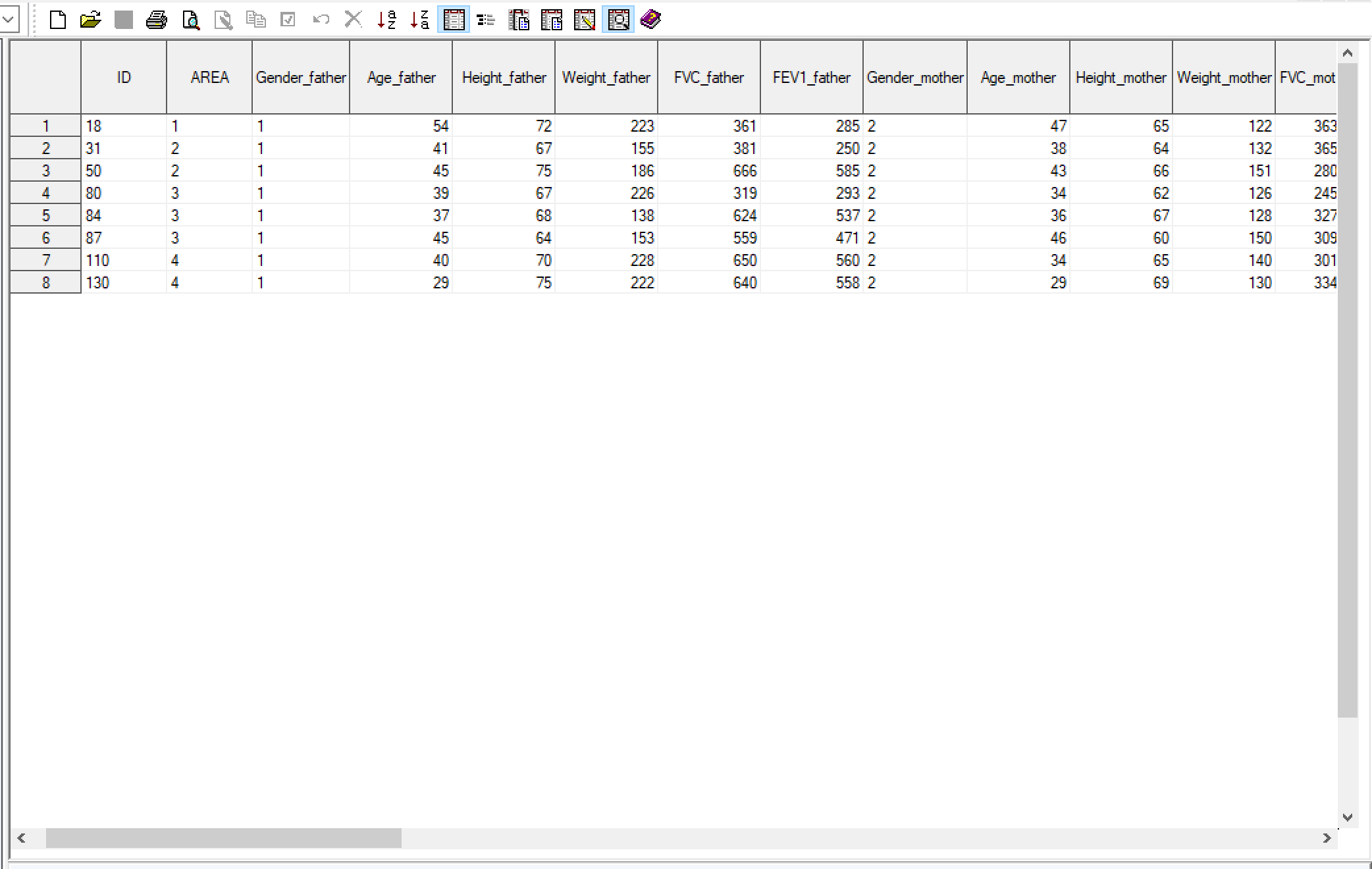
**data** cookd;

set reg\_lungOUT;

if Cookd < (**4** / (**150**-**1**-**1**)) then delete;

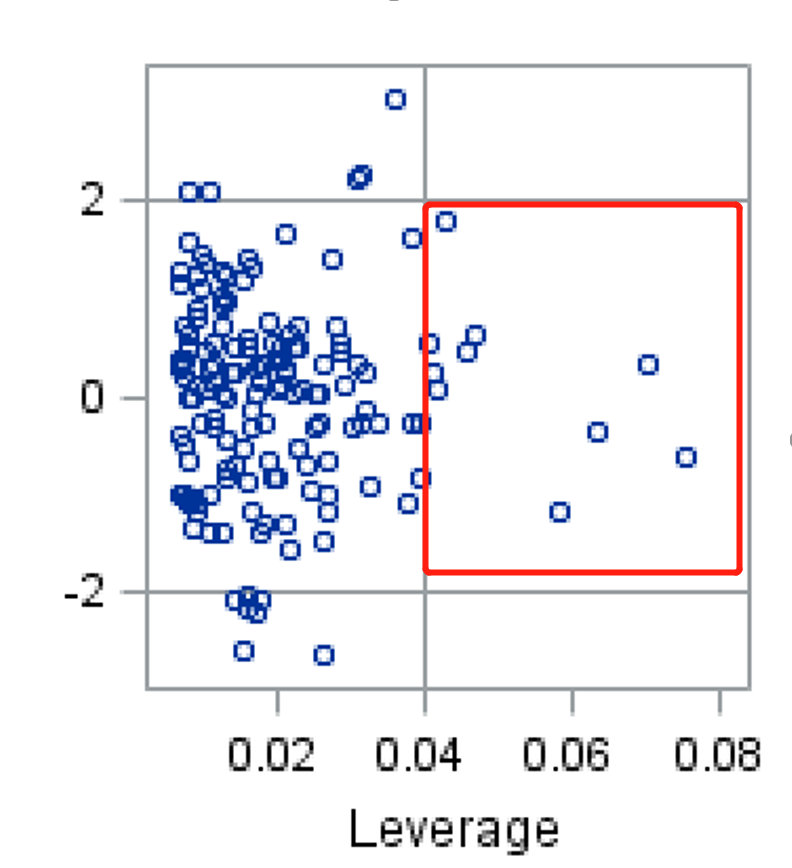
**run**;

The output result:



1. What are the high leverage observations? Why?

An observation with leverage greater than 3 (*m* + 1) */n* may be considered to have high leverage (where *m* indicates the number of predictors),we choose m=1,n=150, and it is should be great than 0.04, The high leverage observations is shown in the graph where the leverage is over 0.04.



The sas code:

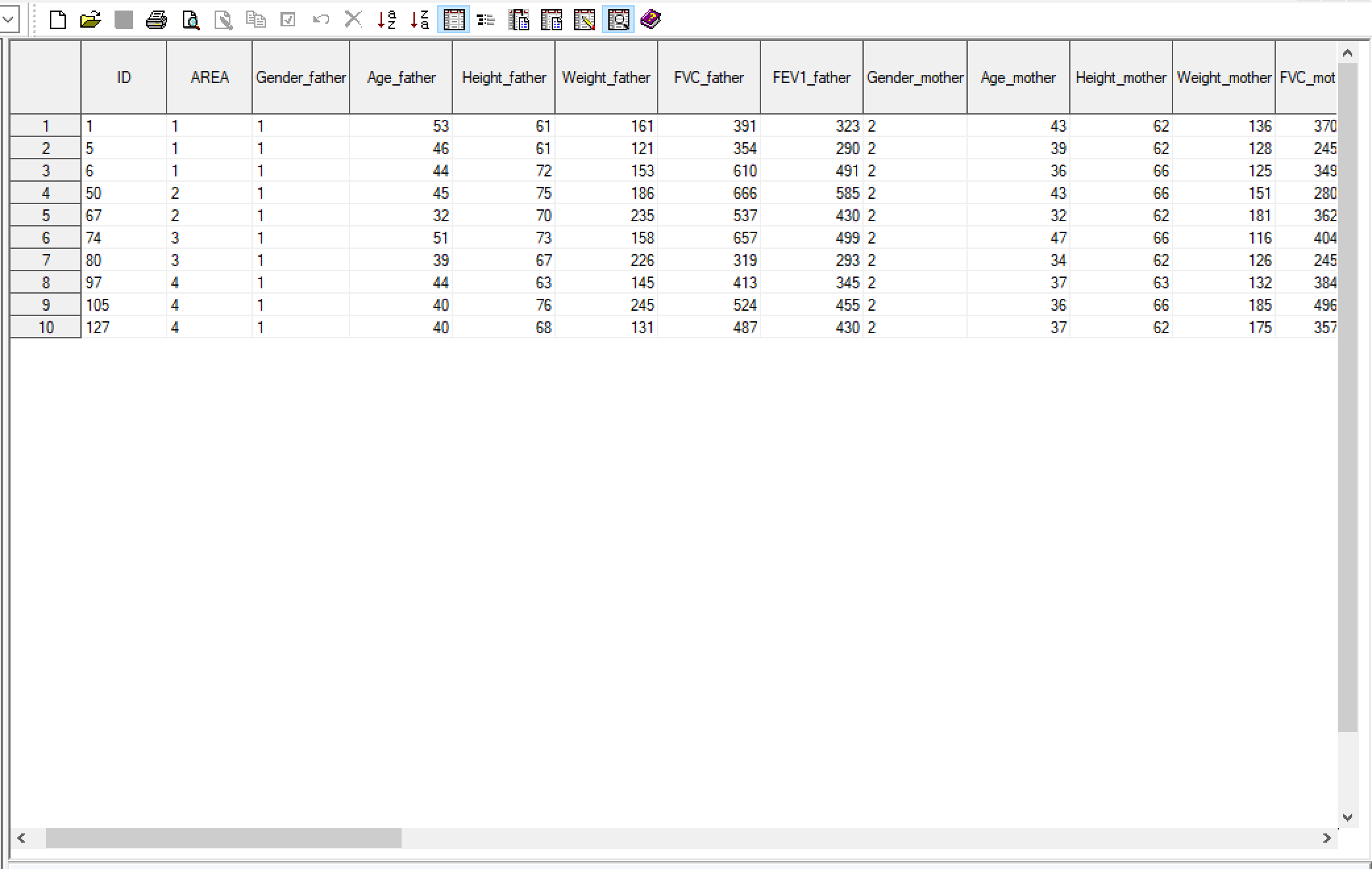
**data** HL;

set reg\_lungOUT;

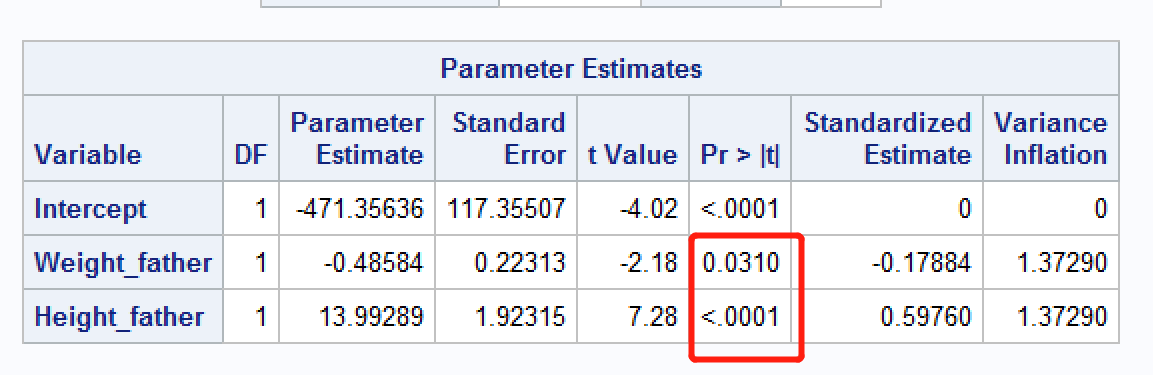
if lev < (**3** \* (**1** + **1**) / **150**) then delete;

**run**;

The output result:



1. For predicting FEV1 of the father, is height more important than weight? Why?



Yes, Height is more important than the weight, since the Pr > |t| value for the weight of father is 0.0310 and the Pr > |t| value for the height of father is <.0001. The lower the value , the more significant the factor.

**Problem 4 - (30 points)**

The Depression data set in CANVAS shows the result of a study on depression. Factors for predicting depression have been grouped into 20 categories (cat\_01 to cat\_20).

1. Normalize the “cat\_01” to “cat\_20” columns using the z transformation with mean=0 and mean=1.

**PROC** **STANDARD**

DATA=depression MEAN=**0** STD=**1**

OUT=depression\_z;

VAR cat\_01 cat\_02 cat\_03 cat\_04 cat\_05 cat\_06

cat\_07 cat\_08 cat\_09 cat\_10 cat\_11 cat\_12 cat\_13 cat\_14

cat\_15 cat\_16 cat\_17 cat\_18 cat\_19 cat\_20;

**RUN**;

1. Perform PCA analysis on the normalized cat\_01 to cat\_20;

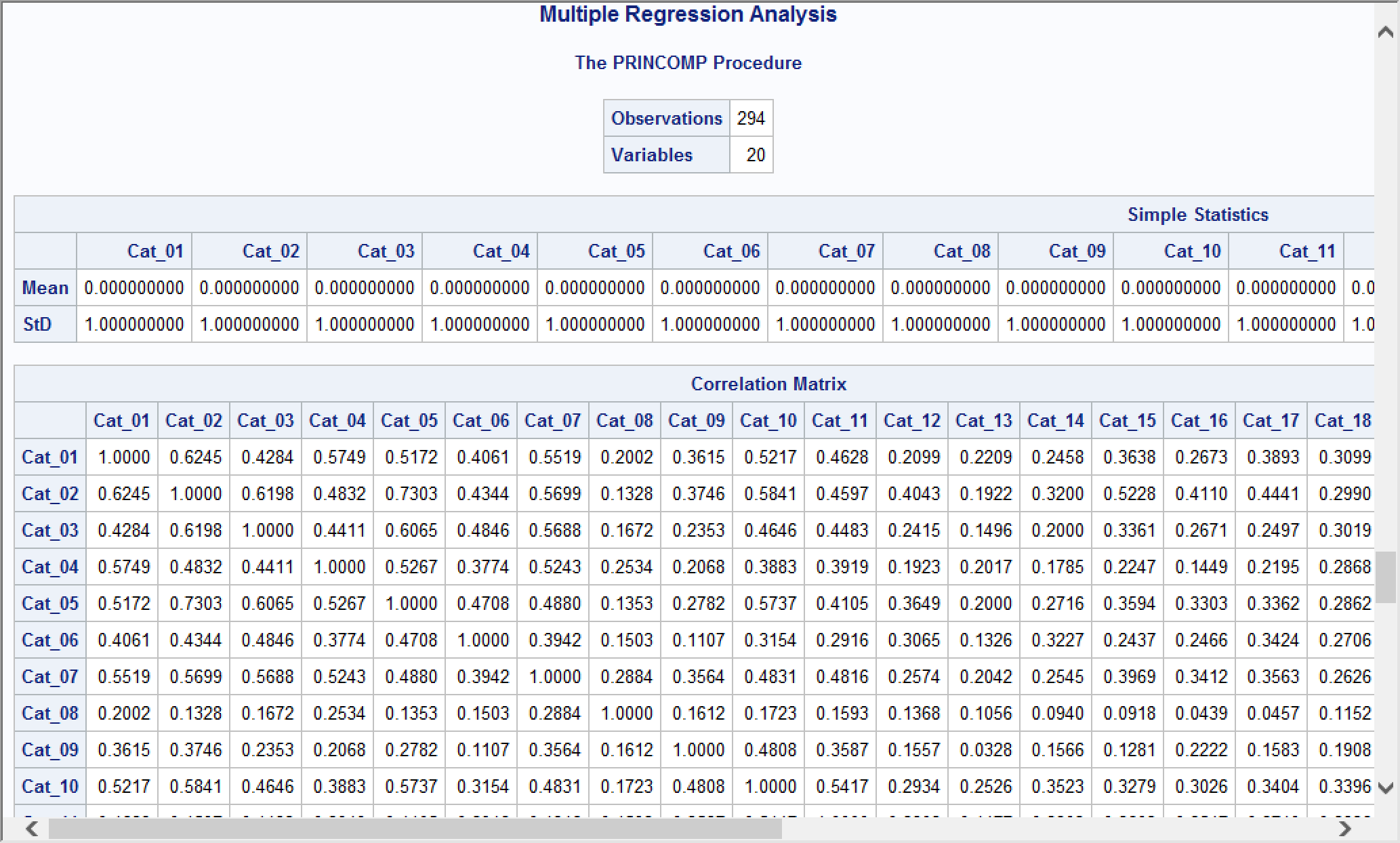
**proc** **princomp** data=depression\_z out=PCA\_depression\_z;

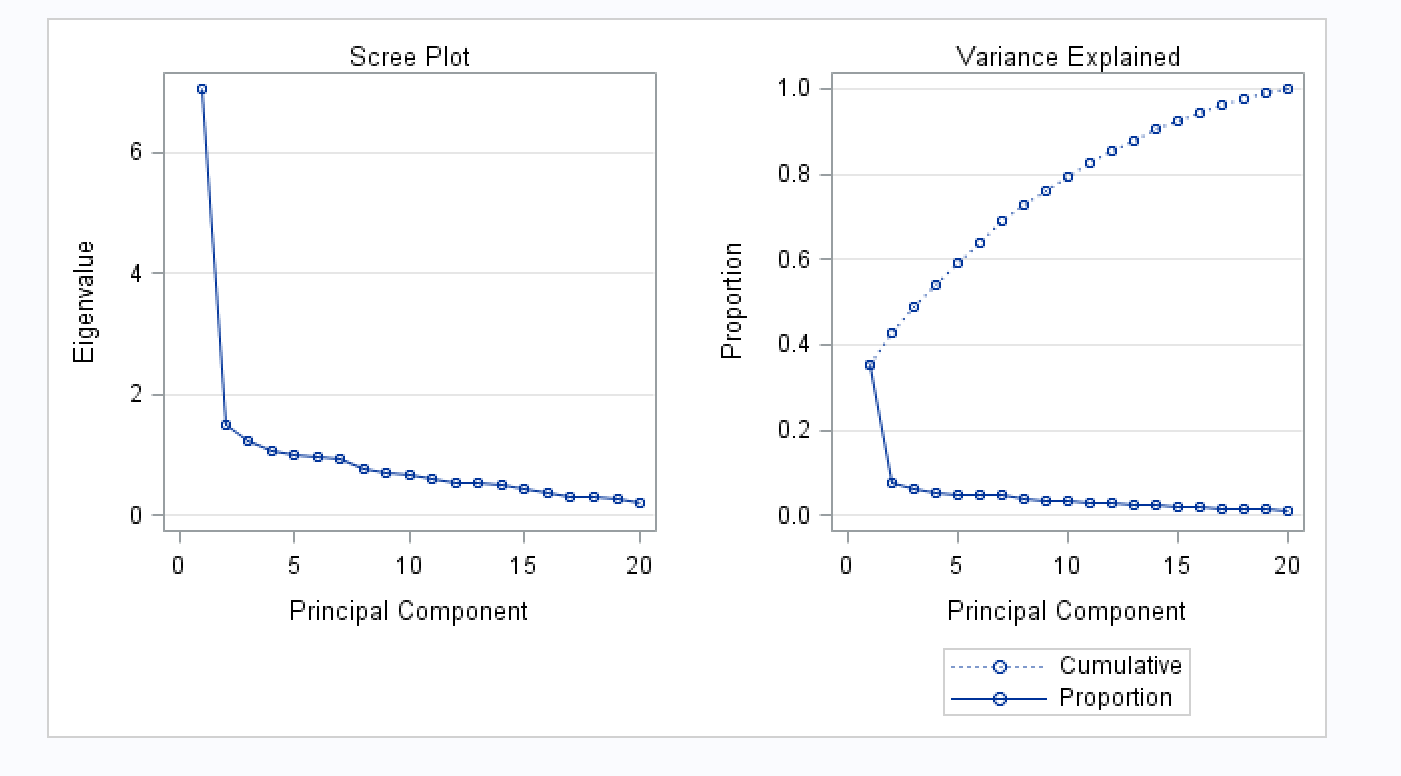
VAR cat\_01 cat\_02 cat\_03 cat\_04 cat\_05 cat\_06

cat\_07 cat\_08 cat\_09 cat\_10 cat\_11 cat\_12 cat\_13 cat\_14

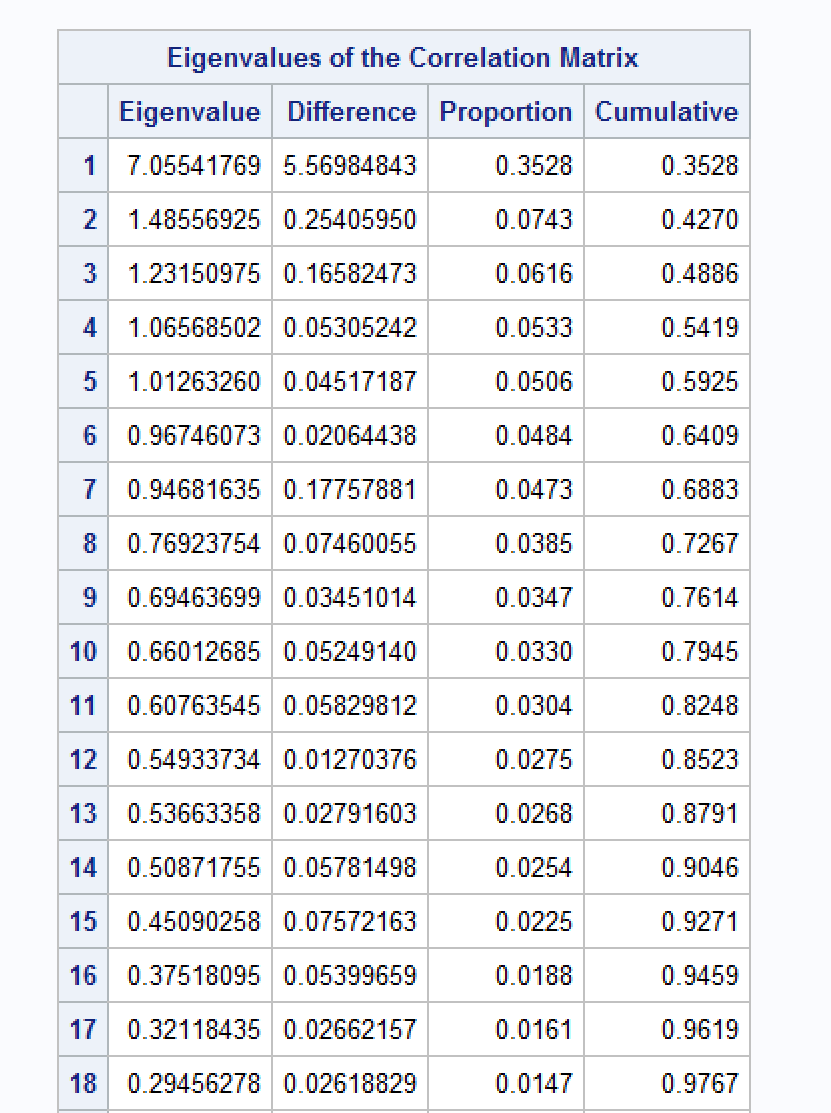
cat\_15 cat\_16 cat\_17 cat\_18 cat\_19 cat\_20;

**run**;





1. Find the best variables among the top 20 principal components that can predict depression.



According to the table, If the condition of the variability is 80%, it needs Prin1 to Prin10 to do the prediction.

If the condition of the variability is 90%, it needs 14 variables to predict the depression from Prin\_1 – Prin\_14.

Datasets: Cereals, Lung, Depression