**Instructions:**

**You can use Word, Excel, Power Point and SAS to answer the questions in this exam. There are a total of six (6) multi-part questions, with point values noted for each question.**

**Please show your calculations, or the details of your program(s) for each problem. Include your SAS programs and output in your submission. The SAS programs should be commented so that each step is clearly explained.**

**Combine all your answers/files into a single zipped file and post the zipped file to “Final Submissions” in Moodle.**

**Problem #1: (15 points)**

X number of high school students are scored on various tests, such as science, math, and social studies (**socst**). The variable **female** is a dichotomous variable, coded 1 if the student was female and 0 if male. Using the multiple regression analysis results below, answer the following questions:

* How many students were scored?

Sol: N= P+E+1 = 195+4+1=200

* Is the overall model significant?

Sol: Yes,the Pr >F is <.0001 which is less than 0.05, therefore the overall model is significant.

* What is the F-value (1-?)?

Sol: Fstat = MSR / MSE = 2385.93019/51.09630 = 46.694774181300799

* What is the R-square for this model (2-?)?

Sol: R-square = SSR/(SSE+SSR)= 9543.72074/19508=0.489220870412139

* What is the formula for this model?

Sol: the formula for this model is Y= a0+a1x0+a2x1+a3x2+a4x3 = 12.32529 + 0.38931 \* math -2.00976 \*female + 0.04984 \* socst + 0.33530 \* read

* Is this a good model? Why or why not?

Sol:

No, it is not a good model, because the There are 2 features “female” and “socst” cannot reject the null hypothesis, because their P-values of the t-statistic are greater than the threshold 0.05.

* Would you change the model? If yes, How?

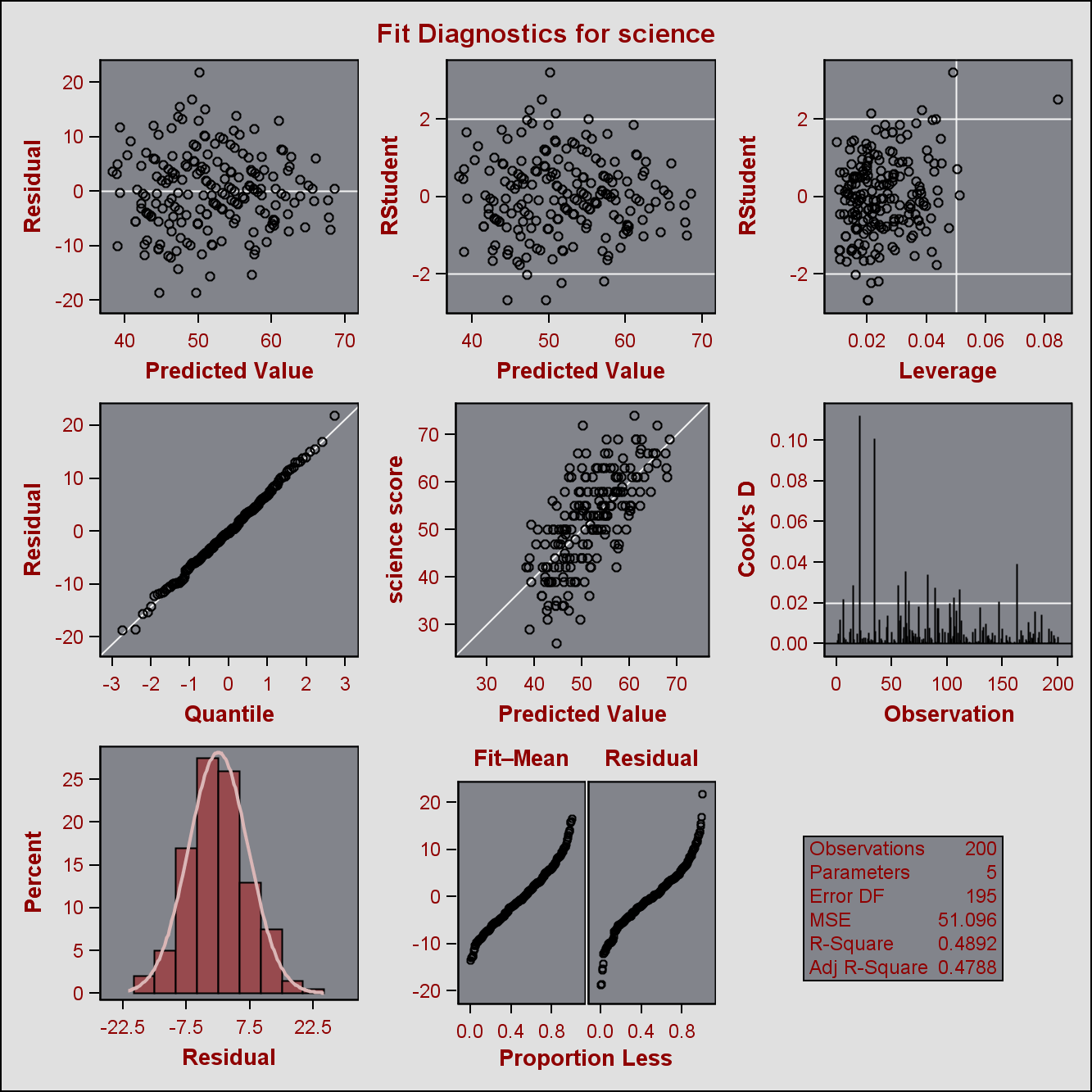
Sol:

Yes, I would change the model. I will rule out the socst parameter and female , because its p value is 0.4241 and 0.0508, The p value of the parameter significant is not at 5% level of significance and then reevaluate the model.

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 4 | 9543.72074 | 2385.93019 | **1-?** | <.0001 |
| **Error** | 195 | 9963.77926 | 51.09630 |  |  |
| **Corrected Total** | 199 | 19508 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 7.14817 | **R-Square** | **2-?** |
| **Dependent Mean** | 51.85000 | **Adj R-Sq** | 0.4788 |
| **Coeff Var** | 13.78624 |  |  |

| **Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **95% Confidence Limits** | |
| **Intercept** | Intercept | 1 | 12.32529 | 3.19356 | 3.86 | 0.0002 | 6.02694 | 18.62364 |
| **math** | math score | 1 | 0.38931 | 0.07412 | 5.25 | <.0001 | 0.24312 | 0.53550 |
| **female** |  | 1 | -2.00976 | 1.02272 | -1.97 | 0.0508 | -4.02677 | 0.00724 |
| **socst** | social studies score | 1 | 0.04984 | 0.06223 | 0.80 | 0.4241 | -0.07289 | 0.17258 |
| **read** | reading score | 1 | 0.33530 | 0.07278 | 4.61 | <.0001 | 0.19177 | 0.47883 |



**Problem #2: select one (5 points)**

A software package has produced the following output for a regression model estimating the nutritional ratings of cereals, based on the location of the cereal on a super market shelf (shelf1, shelf2). Is this model a good regression model?

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | **1** | 45.22003 | 2.23245 | 20.26 | <.0001 |
| **shelf1** | **1** | 0.92541 | 3.73561 | 0.25 | 0.8050 |
| **shelf2** | **1** | -10.24721 | 3.67798 | -2.79 | 0.0068 |

1. The model is NOT a good model because variable shelf2 and “Intercept” are not significant at 5%
2. The model is NOT a good model because variable shelf1 is not significant at 5%
3. The model is NOT a good model because the location of cereal (“shelf1 vs. shelf2) has nothing to do with ratings and cannot cause a change in cereal ratings.
4. Both I and III

Sol: I would like to choose II

**Problem #3: (20 points)**

1. Use the Lung dataset in CANVAS, and forward, backward, and stepwise selection methodologies to develop multiple regression models for “HEIGHT of Oldest Child” as dependent variable and “AGE of Oldest Child”, “WEIGHT of Oldest Child”, “HEIGHT of Mother”, “WEIGHT of Mother, “HEIGHT of Father” and “WEIGHT of Father” as independent variables. (Do not perform any data transformation).
2. Find the best subset of the three variables

Sol1:

/\*

selection=forward

selection=backward

selection=stepwise\*/

title " Multiple Regression Analysis";

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

model Height\_oldest\_child = Age\_oldest\_child Weight\_oldest\_child

Height\_mother Weight\_mother Height\_father Weight\_father

/ dwProb dw VIF selection=forward;

OUTPUT OUT=lung\_out PREDICTED=P\_HOC RESIDUAL=R\_HOC

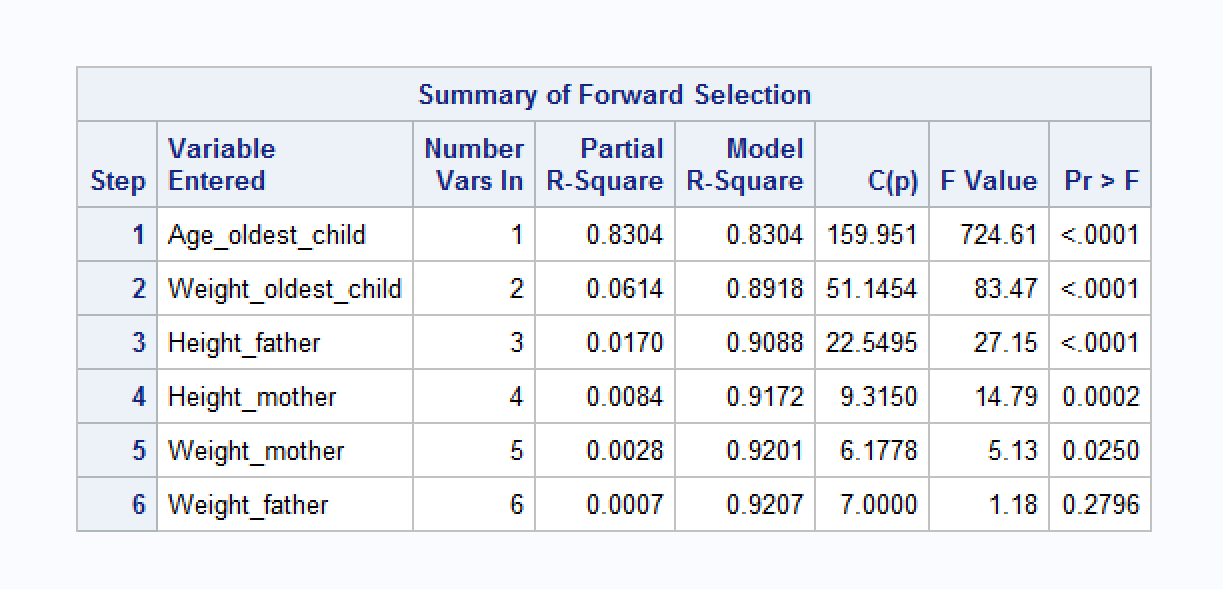
L95M=l95 U95M=U95M L95=L95 U95=U95

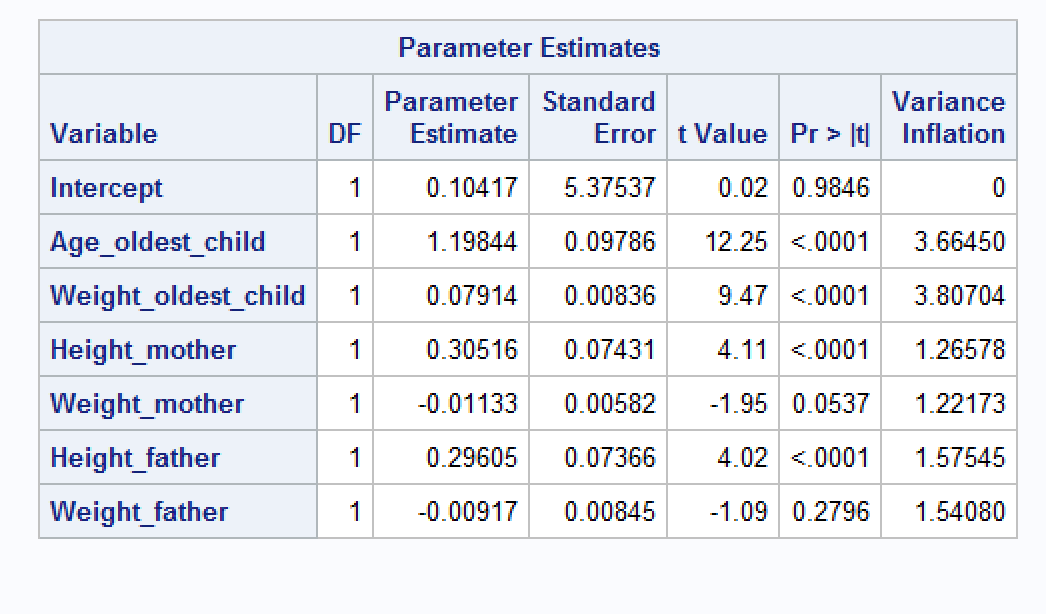
rstudent=rstudent h=lev cookd=cookd dffits=dffits

STDP=STDP STDR=STDR ;

**run**;

**quit**;





title " Multiple Regression Analysis";

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

model Height\_oldest\_child = Age\_oldest\_child Weight\_oldest\_child

Height\_mother Weight\_mother Height\_father Weight\_father

/ dwProb dw VIF selection=backward;

OUTPUT OUT=lung\_out PREDICTED=P\_HOC RESIDUAL=R\_HOC

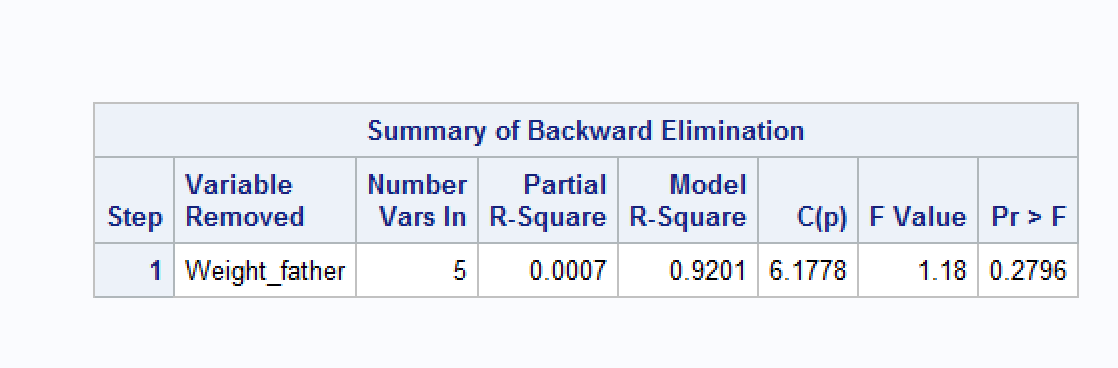
L95M=l95 U95M=U95M L95=L95 U95=U95

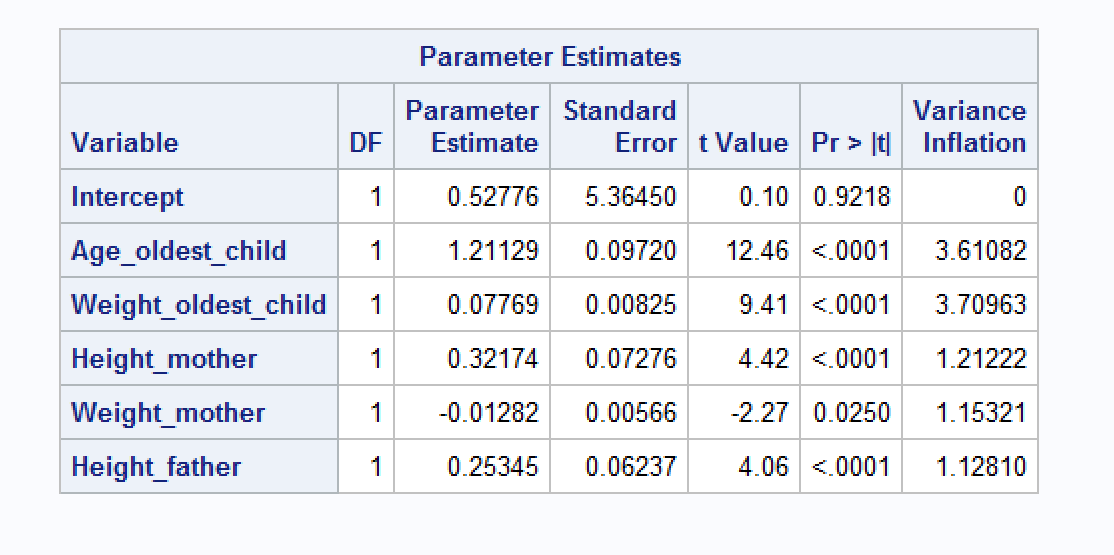
rstudent=rstudent h=lev cookd=cookd dffits=dffits

STDP=STDP STDR=STDR ;

**run**;

**quit**;





title " Multiple Regression Analysis";

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

model Height\_oldest\_child = Age\_oldest\_child Weight\_oldest\_child

Height\_mother Weight\_mother Height\_father Weight\_father

/ dwProb dw VIF selection=stepwise;

OUTPUT OUT=lung\_out PREDICTED=P\_HOC RESIDUAL=R\_HOC

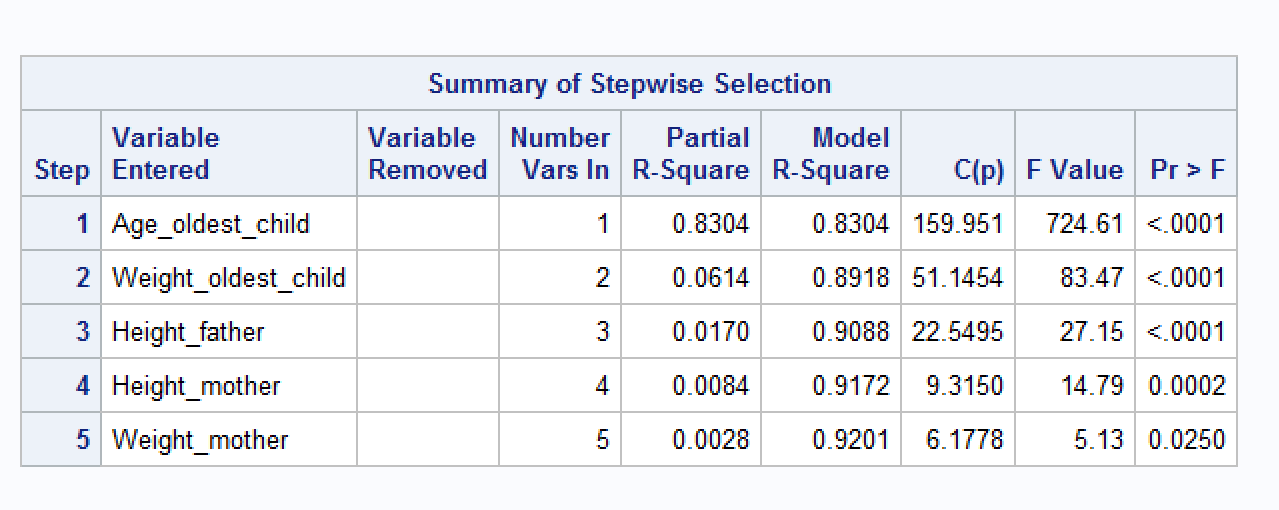
L95M=l95 U95M=U95M L95=L95 U95=U95

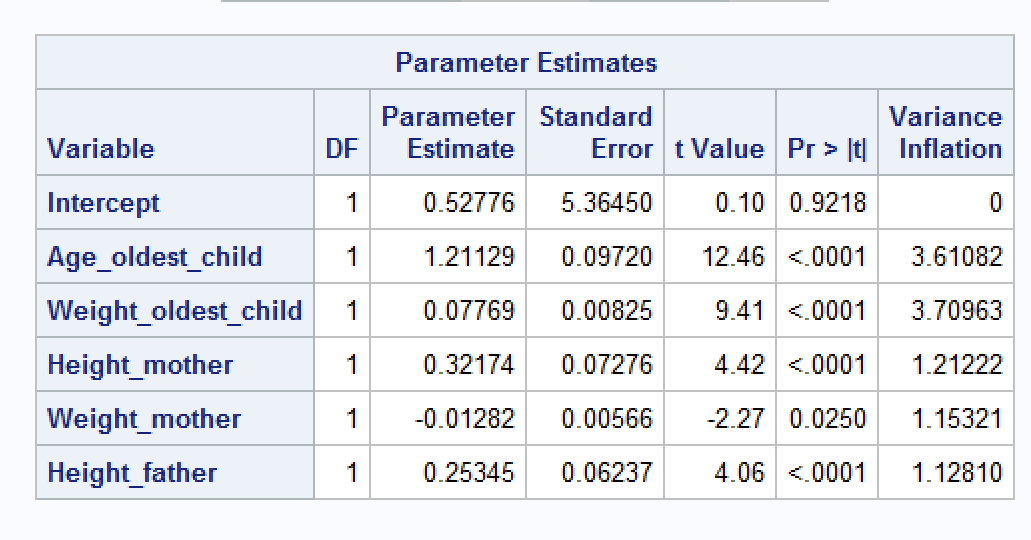
rstudent=rstudent h=lev cookd=cookd dffits=dffits

STDP=STDP STDR=STDR ;

**run**;

**quit**;





Sol2:

title " Multiple Regression Analysis";

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

model Height\_oldest\_child = Age\_oldest\_child Weight\_oldest\_child

Height\_mother Weight\_mother Height\_father Weight\_father

/ dwProb dw VIF selection=MaxR;

OUTPUT OUT=lung\_out PREDICTED=P\_HOC RESIDUAL=R\_HOC

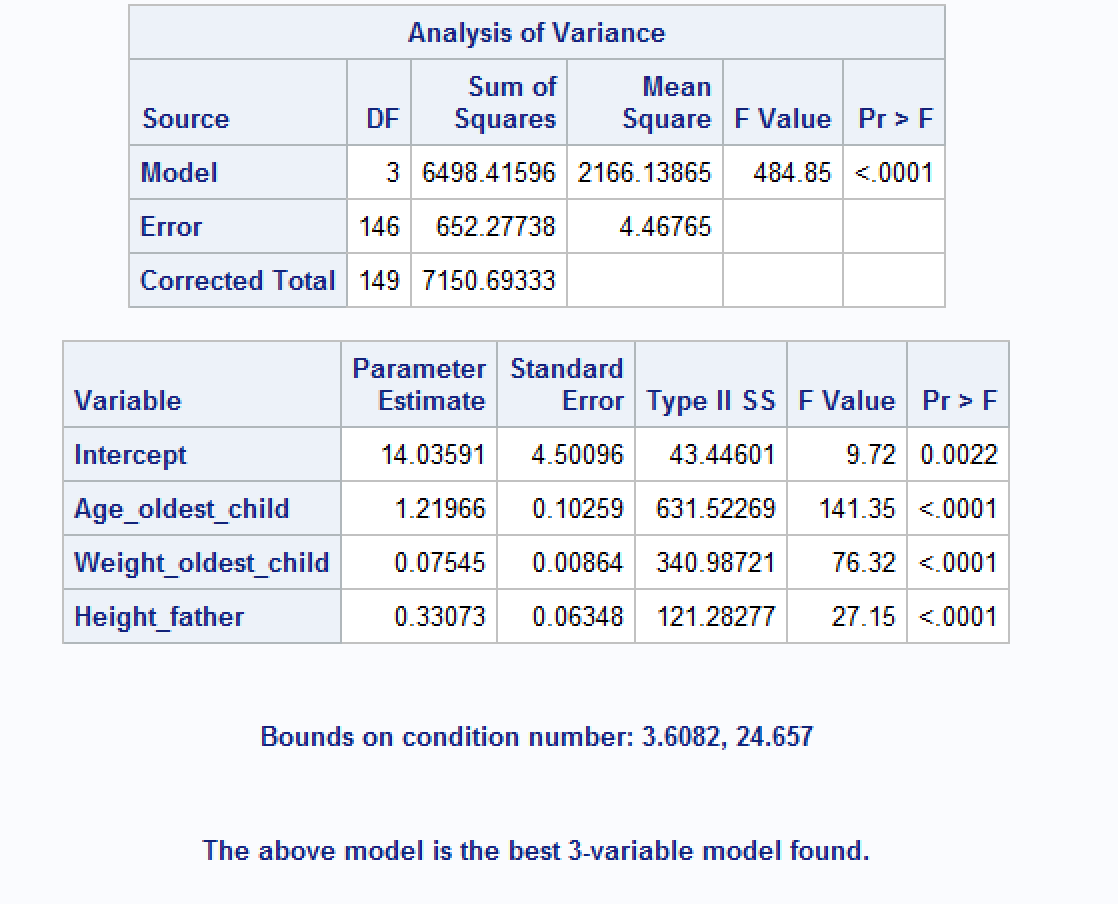
L95M=l95 U95M=U95M L95=L95 U95=U95

rstudent=rstudent h=lev cookd=cookd dffits=dffits

STDP=STDP STDR=STDR ;

**run**;

**quit**;



**From the table above , we can see that the best three variable mode is age\_oldest\_child, weight\_oldest\_child, height\_father.**

**Problem #4: (20 points)**

The “heart attack” dataset in CANVAS contain the records for twenty heat attack patients. The dependent variable (Heart\_Attack\_2) is an indicator showing whether the patient has had a second heart attack within 1 year (yes=1). The first independent variable “Anger Treatment”, indicates whether the patient completed an anger management treatment or not. The second independent variable (“Anxiety Treatment) shows the level of anxiety treatment of the patient.

1. Develop a logistic regression model for predicting the probability of the patient having s second heart attack (show your development steps)

Code:

**proc** **logistic** data=Heart\_attack descending;

class Anger\_Treatment(ref='0')

/ param=ref ;

model Heart\_Attack\_2 = Anger\_Treatment Anxiety\_Treatment;

**quit**;



From the table we can see Pr>ChiSq is 0.3818 , which is bigger than 0.05, then we will remove this parameter out:

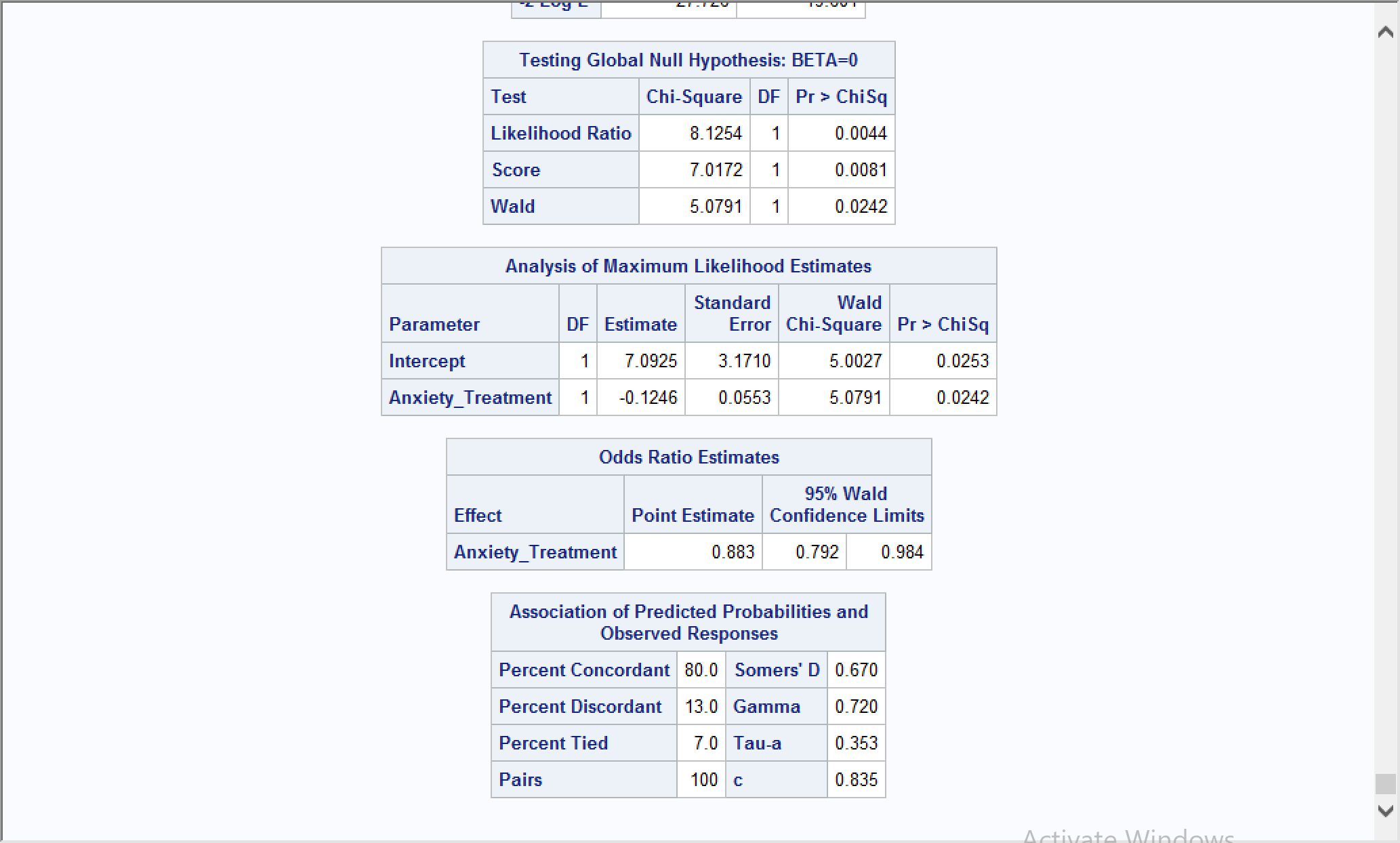
The we run this code:

\* delete the Anger\_Treatment from the model ;

**proc** **logistic** data=heart\_attack;

model Heart\_Attack\_2 = Anxiety\_Treatment;

**quit**;



**Then we got the model is**

**Heart\_Attack\_2 = -0.1246 \* Anxiety\_Treatment + 7.0925**

1. Using your model:
   1. Predict the probabilities of the following two patients (A and B) having a heart attack within the next year?

|  |  |  |
| --- | --- | --- |
| Patient | Anger Treatment | Anxiety Treatment |
| A | 1 | 40 |
| B | 0 | 70 |
|  |  |  |

Sol:

Patient A: g(x)=-0.1246\*40+7.0925=2.1085

π(x)== 0.1082

Patient B: g(x)= -0.1246\*70+7.0925=-1.6295

π(x)== 0.836

* 1. What are the odds for patient A and patient B?

Sol:

Patient A: odds=0.1082/(1-0.1082)=0.1213

Patient B: odds=0.836/(1-0.836)=5.0976

* 1. What is the odds ratio of A over B?

**Sol:**

Odds ratio=0.1213/5.0976=0.0238

**Problem #5: (20 points)**

The Breast Cancer dataset in CANVAS includes some of the features that are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei in the image. (Source: UCI). Perform PCA analysis on the following 10 variables.

1. How many principal components should be used to explain at least 85 percent of the variability in data?

Sol:

Code:

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

select breast\_cancer\_data;

**run**;

**PROC** **STANDARD**

DATA=breast\_cancer\_data MEAN=**0** STD=**1**

OUT= breast\_cancer\_data\_z ;

VAR radius\_mean texture\_mean perimeter\_mean area\_mean smoothness\_mean

compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean

fractal\_dimension\_mean;

**RUN**;

**proc** **princomp** data= breast\_cancer\_data\_z out=pca\_breast\_cancer\_data ;

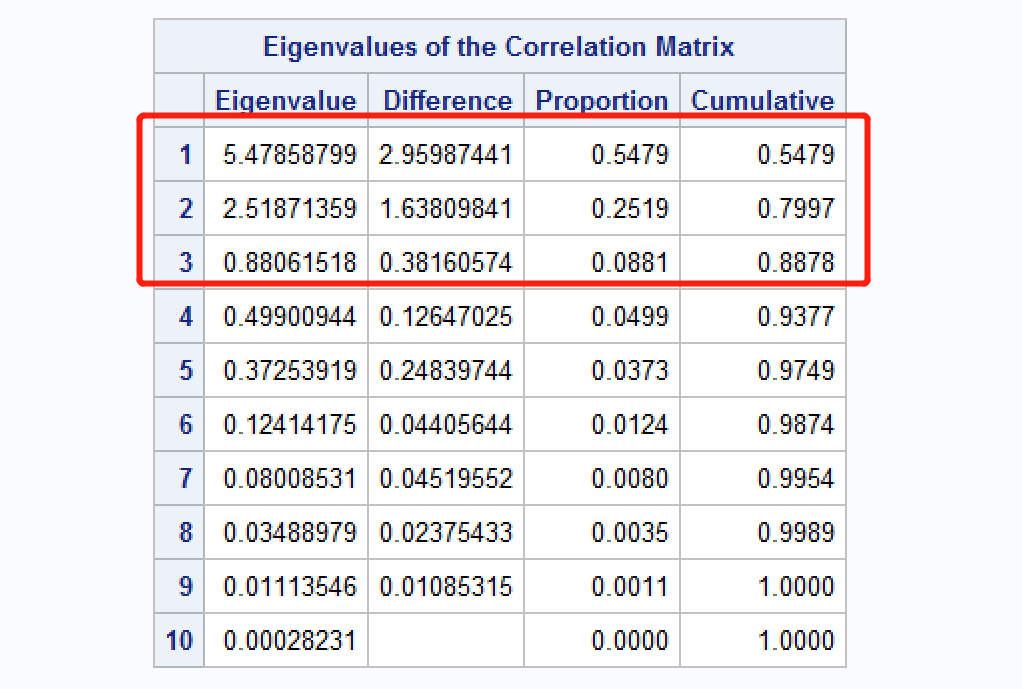
VAR radius\_mean texture\_mean perimeter\_mean area\_mean smoothness\_mean

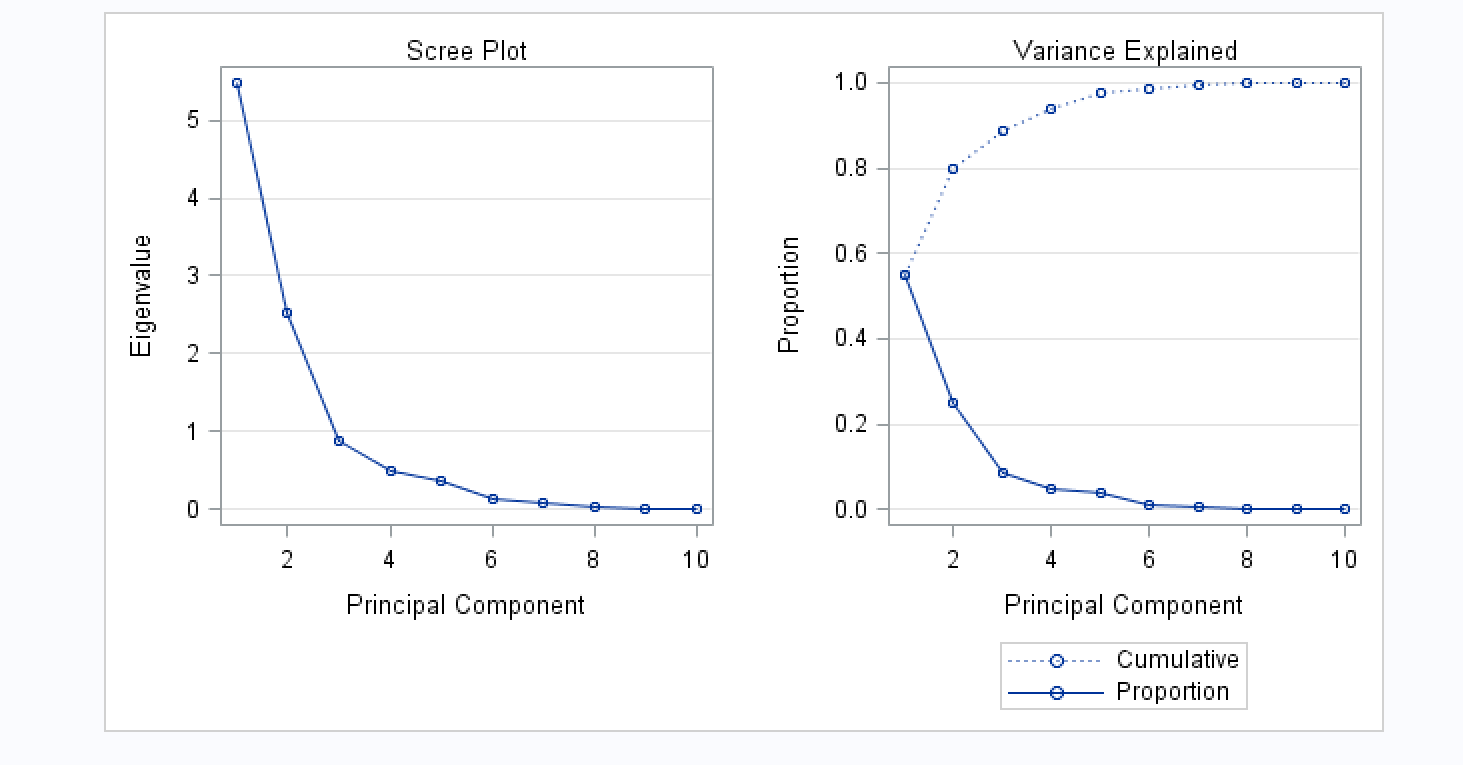
compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean

fractal\_dimension\_mean;

**run**;

Output result:

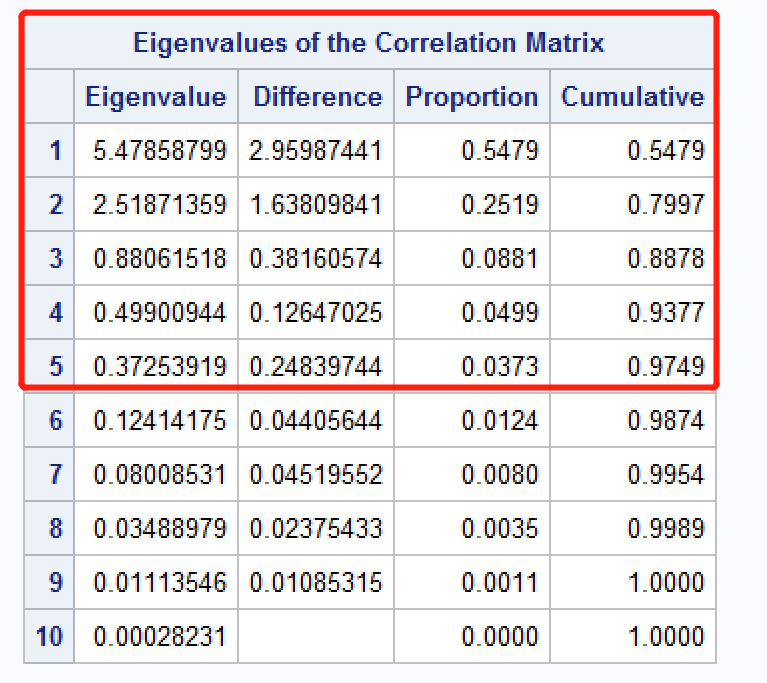




We can see that the cumulative value is 0.8878 when the first 3 principal components are selected, which is over at least 85 percent of the variability in data.

1. What if the study requires more than 95 percent of variability to be explained, how many variables do you use?
2. radius\_mean
3. texture\_mean
4. perimeter\_mean
5. area\_mean
6. smoothness\_mean
7. compactness\_mean
8. concavity\_mean
9. concave\_points\_mean
10. symmetry\_mean
11. fractal\_dimension\_mean

**Sol:**



From the table above, we know that the first 5 principal components should be used to explain at least 95 percent of the variability in data.

**Problem #6: (20 points)**

Assuming the following web structure, calculate the page rank of nodes A through F.

**data** Arcs;

infile datalines;

input Node $ A B C D E F G;

datalines;

A 0 1 1 0 0 1 0

B 1 0 0 1 0 0 1

C 0 0 0 1 0 1 0

D 0 1 0 0 1 0 0

E 0 0 1 0 0 0 0

F 0 0 0 0 1 0 0

G 0 1 0 0 0 0 0

;

**run**;

**proc** **sql**;

create table matrix\_1 as

select a/sum(a) as x1

,b/sum(b) as x2

,c/sum(c) as x3

,d/sum(d) as x4

,e/sum(e) as x5

,f/sum(f) as x6

,g/sum(g) as x7

from Arcs

;

**quit**;

**data** rank\_p;

x1=**1**/**7**;

x2=**1**/**7**;

x3=**1**/**7**;

x4=**1**/**7**;

x5=**1**/**7**;

x6=**1**/**7**;

x7=**1**/**7**;

output;

**run**;

**proc** **iml**;

use matrix\_1;

read all var { x1 x2 x3 x4 x5 x6 x7 } into M;

print M;

use rank\_p;

read all var { x1 x2 x3 x4 x5 x6 x7 } into rank\_p1;

rank\_p = t(rank\_p1);

print rank\_p ;

rank\_p2=(M\*rank\_p);

print rank\_p2 ;

rank\_p3=(M\*rank\_p2 );

print rank\_p3 ;

rank\_p3b=(M\*\***2**)\*rank\_p;

print rank\_p3b ;

rank\_p50=(M\*\***50**)\*rank\_p;

print rank\_p50 ;

**quit**;

|  |
| --- |
| Multiple Regression Analysis |

| **M** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | 0.3333333 | 0.5 | 0 | 0 | 0.5 | 0 |
| 1 | 0 | 0 | 0.5 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0.5 | 0 | 0.5 | 0 |
| 0 | 0.3333333 | 0 | 0 | 0.5 | 0 | 0 |
| 0 | 0 | 0.5 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0.5 | 0 | 0 |
| 0 | 0.3333333 | 0 | 0 | 0 | 0 | 0 |

| **rank\_p** |
| --- |
| 0.1428571 |
| 0.1428571 |
| 0.1428571 |
| 0.1428571 |
| 0.1428571 |
| 0.1428571 |
| 0.1428571 |

| **rank\_p2** |
| --- |
| 0.1904762 |
| 0.3571429 |
| 0.1428571 |
| 0.1190476 |
| 0.0714286 |
| 0.0714286 |
| 0.047619 |

| **rank\_p3** |
| --- |
| 0.2261905 |
| 0.297619 |
| 0.0952381 |
| 0.1547619 |
| 0.0714286 |
| 0.0357143 |
| 0.1190476 |

| **rank\_p3b** |
| --- |
| 0.2261905 |
| 0.297619 |
| 0.0952381 |
| 0.1547619 |
| 0.0714286 |
| 0.0357143 |
| 0.1190476 |

| **rank\_p50** |
| --- |
| 0.1849831 |
| 0.3821913 |
| 0.0853217 |
| 0.1520627 |
| 0.0433784 |
| 0.0213122 |
| 0.1307505 |

So the page rank through A to F:

| **rank\_p50** |
| --- |
| A=0.1849831 |
| B=0.3821913 |
| C=0.0853217 |
| D=0.1520627 |
| E=0.0433784 |
| F=0.0213122 |