Dominic LaRoche

9/9/14

Assignment 1

For this assignment we were required to read in a data set and calculate a new variable from the existing variables using a data step. I partitioned this problem into three steps:

Step 1: Read in raw data

**data** class.risk;

infile 'C:\Classes\SASandDataMgmt\Risk\_data.csv' dsd dlm=',' firstobs=**2** missover;

input Case\_ID DIAGNOSIS $ los DialAtTx $ LipidDis $ CbV\_DS $ PVD $ Neurpthy $ RespProb $ CancrPre $ CBYPASS $ MI $;

run;

Step 2: Define new variables with the appropriate weights necessary for calculating the CCI score. By defining new numeric variables I ensure that anything other than a Y or N will be coded missing.

**data** class.risk1;

set class.risk;

if los < **0** then delete;

if MI='Y' then MI\_c=**2**;

Else if MI='N' then MI\_c=**0**;

if CBYPASS='Y' then CBYPASS\_c=**1**;

Else if CBYPASS='N' then CBYPASS\_c=**0**;

if DIAGNOSIS='TYPE1\_DM' then DIAG\_c=**1**;

else if DIAGNOSIS='TYPE2\_DM' then DIAG\_c=**0**;

if CbV\_DS='Y' then CbV\_DS\_c=**1**;

Else if CbV\_DS='N' then CbV\_DS\_c=**0**;

if PVD='Y' then PVD\_c=**1**;

else if PVD='N' then PVD\_c=**0**;

if Neurpthy='Y' then neur\_c=**1**;

else if Neurpthy='N' then neur\_c=**0**;

if RespProb='Y' then resp\_c=**1**;

else if RespProb='N' then resp\_c=**0**;

if LipidDis='Y' then lipid\_c=**2**;

else if LipidDis='N' then lipid\_c=**0**;

if DialAtTx='Y' then dial\_c=**3**;

else if DialAtTx='N' then dial\_c=**0**;

if CancrPre='Y' then canc\_c=**3**;

else if CancrPre='N' then canc\_c=**0**;

run;

Step 3: Calculate the CCI by summing the weights from the new variables just defined. This will create missing CCI if any of the weight variables are missing. Since we don’t know the status of the missing weights I feel this is appropriate. An alternative would have been to impute the missing weights as 0 so that the CCI could still be calculated with the available information.

**data** class.risk3;

set class.risk1;

CCI= MI\_c + CBYPASS\_c + DIAG\_c + CbV\_DS\_c + PVD\_c + neur\_c + resp\_c + lipid\_c + dial\_c + canc\_c;

run;

We were then required to calculate the correlation coefficient between CCI and length of stay.

**proc** **corr** data=class.risk3;

var CCI los;

**run**;

This resulted in a positive and statistically significant (alpha<0.05) correlation of 0.138 (p=0.0003).

| **Simple Statistics** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** |
| **CCI** | 673 | 4.89004 | 2.06150 | 3291 | 1.00000 | 12.00000 |
| **los** | 829 | 11.69602 | 9.51996 | 9696 | 1.00000 | 86.00000 |

| **Pearson Correlation Coefficients  Prob > |r| under H0: Rho=0  Number of Observations** | | |
| --- | --- | --- |
|  | **CCI** | **los** |
| **CCI** | |  | | --- | | 1.00000 | |  | | 673 | | |  | | --- | | 0.13829 | | 0.0003 | | 673 | |
| **los** | |  | | --- | | 0.13829 | | 0.0003 | | 673 | | |  | | --- | | 1.00000 | |  | | 829 | |

67 /\*Problem 1 9-4-14 CPH576D\_LaRoche\*/

68 data class.risk;

69 infile 'C:\Classes\SASandDataMgmt\Risk\_data.csv' dsd dlm=',' firstobs=2 missover;

70 input Case\_ID DIAGNOSIS $ los DialAtTx $ LipidDis $ CbV\_DS $ PVD $ Neurpthy $ RespProb $

70 ! CancrPre $ CBYPASS $ MI $;

71 run;

NOTE: The infile 'C:\Classes\SASandDataMgmt\Risk\_data.csv' is:

Filename=C:\Classes\SASandDataMgmt\Risk\_data.csv,

RECFM=V,LRECL=256,File Size (bytes)=28771,

Last Modified=04Sep2014:11:39:55,

Create Time=04Sep2014:11:39:55

NOTE: 830 records were read from the infile 'C:\Classes\SASandDataMgmt\Risk\_data.csv'.

The minimum record length was 27.

The maximum record length was 35.

NOTE: The data set CLASS.RISK has 830 observations and 12 variables.

NOTE: DATA statement used (Total process time):

real time 0.02 seconds

cpu time 0.03 seconds

72 /\* For each variable included in the CCI score creatre a numeric variable with the

72 ! appropriate weight\*/

73 data class.risk1;

74 set class.risk;

75 if los < 0 then delete; /\*remove cases with negative los\*/

76 if MI='Y' then MI\_c=2;

77 Else if MI='N' then MI\_c=0;

78 if CBYPASS='Y' then CBYPASS\_c=1;

79 Else if CBYPASS='N' then CBYPASS\_c=0;

80 if DIAGNOSIS='TYPE1\_DM' then DIAG\_c=1;

81 else if DIAGNOSIS='TYPE2\_DM' then DIAG\_c=0;

82 if CbV\_DS='Y' then CbV\_DS\_c=1;

83 Else if CbV\_DS='N' then CbV\_DS\_c=0;

84 if PVD='Y' then PVD\_c=1;

85 else if PVD='N' then PVD\_c=0;

86 if Neurpthy='Y' then neur\_c=1;

87 else if Neurpthy='N' then neur\_c=0;

88 if RespProb='Y' then resp\_c=1;

89 else if RespProb='N' then resp\_c=0;

90 if LipidDis='Y' then lipid\_c=2;

91 else if LipidDis='N' then lipid\_c=0;

92 if DialAtTx='Y' then dial\_c=3;

93 else if DialAtTx='N' then dial\_c=0;

94 if CancrPre='Y' then canc\_c=3;

95 else if CancrPre='N' then canc\_c=0;

96 run;

NOTE: There were 830 observations read from the data set CLASS.RISK.

NOTE: The data set CLASS.RISK1 has 829 observations and 22 variables.

NOTE: DATA statement used (Total process time):

real time 0.04 seconds

cpu time 0.06 seconds

97 /\*calculate the CCI from the numeric variables just created\*/

98 data class.risk3;

99 set class.risk1;

100 CCI= MI\_c + CBYPASS\_c + DIAG\_c + CbV\_DS\_c + PVD\_c + neur\_c + resp\_c + lipid\_c + dial\_c +

100! canc\_c;

101 run;

NOTE: Missing values were generated as a result of performing an operation on missing values.

Each place is given by: (Number of times) at (Line):(Column).

2 at 100:11 10 at 100:23 130 at 100:51 5 at 100:60 5 at 100:69

4 at 100:88

NOTE: There were 829 observations read from the data set CLASS.RISK1.

NOTE: The data set CLASS.RISK3 has 829 observations and 23 variables.

NOTE: DATA statement used (Total process time):

real time 0.02 seconds

cpu time 0.01 seconds

102 /\*Correlate CCI with length of stay\*/

103 proc corr data=class.risk3;

104 var CCI los;

105 run;

NOTE: PROCEDURE CORR used (Total process time):

real time 0.19 seconds

cpu time 0.04 seconds