Stanwrick Daniels, Harini Mittal, Despina Skartados

STA 9705 Multivariate Statistical Methods

Final Project

## **General Introduction**

We intended to investigate entrepreneurial behaviors and attitudes using data collected from the Global Entrepreneurship Monitor (GEM). The data collected is through GEM’s Adult Population Survey, which is given to at least 2000 adults between the ages of 18 and 64 to complete and is designed to collect information about the important indicators. The results from the survey consisted of many years, countries, and variables. We narrowed down our data set by identifying variables of interest that may influence a business and proceeded by choosing to analyze the 2017 and 2018 data. For example, we removed Panama because it lacked data for the chosen variable of entrepreneurial employee activity. Deleting the countries of missing observations helps to reduce the chance of making biased conclusions. Then, we classified the level of income in each country using the World Bank data as either high, middle, or low, where middle consisted of the upper middle class and low consisted of lower middle class and low-income level as well. To ensure each year had the same number of observations for each income level, we also balanced out the data set by removing extra observations. We decided to keep 30 observations of countries that are considered high-income for 2017 and 2018. Likewise, we kept 11 middle-income countries and 6 low-income countries for each year. Our data consists of 47 countries for each year, having a total of 94 observations with some overlap of countries between the two years. Each year does not consist of the same countries because they did not all respond to the survey in both years, which is why we decided to simply balance each year with the same number of observations per income level. After having finalized our data set, we started experimenting with the techniques learned throughout the semester on the real data we found. We performed multivariate analysis of variance, classification analysis, multivariate regression, and canonical correlation.

## **Objective(s) of Analysis**

Our primary goal was to look at the difference in income levels amongst countries and see how this affects the establishment of a business. Additionally, we decided to use two years worth of data to have more observations and be able to analyze entrepreneurial behaviors and attitudes more precisely. We performed multivariate analysis of variance to test if the means amongst the three income levels are different. Classification analysis was used to check the probability of misclassification of the countries in each respective income level using apparent and hold-out error rates for various methods. Multivariate regression was used to examine if the predictor variables were useful in response to the total early-stage entrepreneurial activity and established business ownership. We also explored testing the significance of the reduced models to see if any of the chosen predictor variables we found of interest can be removed from the model to improve the results. The purpose of looking at the canonical correlations was to measure the linear relationship between the predictor and response variables. Each method will provide us with distinct results for our analysis on entrepreneurs in diverse countries based on the three income levels.

## **Data Description**

Our data includes 2 y-variables and 5 x-variables, which include:

X1=Perceived Opportunities (OPP)

X2=Perceived Capabilities (CAP)

X3=Fear of failure rate (FAIL)

X4= Entrepreneurial Intentions (INT)

X5=Entrepreneurial Employee Activity (EMPACTV)

Y1=Total early-stage Entrepreneurial Activity (TEA)

Y2=Established Business Ownership (EBA)

As per the Global Entrepreneurship Monitor’s Adult Population Survey, perceived opportunities depict the rate of the people in each country who see the area in which they live to be a good place to start a business. Perceived capabilities is a percentage rate of people who believe they have the necessary skills and knowledge to start a business. The fear of failure rate shows the percentage of people that would not open a business because of fear. Entrepreneurial intentions show the rate of entrepreneurs who plan on starting a business within three years. Total early-stage entrepreneurial activity is the rate of people who are entrepreneurs or owner-managers of a new business. Established business ownership shows the rate of current owner-managers of an established business. According to the World Bank Classification of income levels, we let:

1=High Income

2=Middle Income

3=Low Income

## **Multivariate Analysis of Variance**

Objectives for the One Way MANOVA analysis are to determine if there is statistically significant evidence of a difference in the means for the three levels of income (Low, Middle, and High) with 2017 and 2018. In other words, we want to recognize if these 5 predictor variables are distinct.

Test: Is there a difference in the performance of businesses based on countries with low, middle, and high income as it relates to start-up and established businesses in 2017-2018?

Null Hypothesis- H0: μ1=μ2=μ3 = 0 (No difference in the means of the three levels of income as it relates to entrepreneurial behavior, attitude, and the establishment of businesses.)

Alternative Hypothesis-Ha: At least one μn is an inequality.

Implementation:

(1) We extracted the data from the file using Excel, and narrowed it down to make sure the sub-data was applicable to the analysis of the hypothesis.

(2) We defined all parts of the code to match the data set. The Class and the Model statements were defined. The variable economy (Income) was inserted into the CLASS statement then the variables x1, x2, x3, x4, x5 y1 y2 were inserted on the left side of the equal sign of the model. The variable *economy (Income)* was inserted on the right-hand side.

(3) We ran MANOVA code after linking the data file to the code to the extracted output in SAS. See Appendix A for code and Appendix B for output.

(4) We extracted the results and analyzed the data.

Limitations to MANOVA Analysis:

(1) The MANOVA method did not work well with the unbalanced dataset since each level of income had a different number of observations. You may need to extract or delete some data to make it applicable so as to be executed using this technique.

(2) The sample size for each economy (level of income) group was unequal. This may cause the data not to have sufficient robustness. We created a smaller sample of the data set which consisted of 6 observations for each level of income per year. This smaller sample of our data consists of 36 observations in total.

(3) If there are outliers in the data set the results will be affected.

Assumptions to MANOVA Analysis:

(1) The observation for each variable is independent.

(2) The observation for each variable is normally distributed.

(3) Assume you have a balanced data set. We balanced it in a way so that each level has the same number of observations. We removed some of the 2018 data to balance it with the 2017 data.

(4) The four tests are not equivalent in general when Ho is false.

(5) Apply MANOVA to account for correlation between multiple variables.

(6) Rencher (2002) discussed the idea that robustness is affected by unequal groups (p. 177). Now, since the sample size was not equal for each income level in this data set, we were hoping that the MANOVA tests will have sufficient data to be robust with respect to the heterogeneity of the covariance matrix. However, we were able to extract a sample data set with equal groups.

(8) The response variables in the 2017-2018 data set are all continuous.

(9) Assume a significance level of 0.05.

Interpretation of Results:

The SAS output gives us promising results about the type of variables that will affect or influence whether a business is considered a startup or an established business. There are three income levels we deduced as per the World Bank measure (criteria) for evaluating a business based on the income levels using the 5 x-variables and 2 y-variables.

In this technique, we are testing to see if there is a significant difference among the means in terms of perceived opportunities, perceived capabilities, fear of failure rate, entrepreneurial intentions, entrepreneurial employee activity, and three income levels using data from 2017 and 2018. We assume a significant level of 0.05.

The H matrix shows two characteristic roots. There are only two nonzero eigenvalues.

The correlation matrix shows the correlation between the variables in the data set.

The MANOVA test for the hypothesis of No Overall income effect shows the exact p-values are all less than 0.05.

The eigenvalue matrix shows that there are 2 non-zero eigenvalues. The first eigenvalue, λ1, accounted for 95.63% λ2 and sums up to 100% (95.63 and 4.37). It means that the essential dimensionality of the mean vector space is 1. Therefore, this is a linear model.

Based on the eigenvalue matrix, the first two eigenvalues contribute 100% to the variance of the model. Since the model is the full model, the power of the MANOVA test follows Roy’s Test > Lawley Hotelling Test>, Pillai Trace> Test then Wilks’ Test.

In this case, the more appropriate and powerful test is Roy’s test.

The overall MANOVA tests show that there are statistically significant differences in the means for perceived opportunities, perceived capabilities, fear of failure rate, entrepreneurial intention, and entrepreneurial employee activity among the income levels for 2017- 2018. The overall p-values are 0.0006, 0.0028, 0.003, and 0.0002 for all 4 tests. Since the p-values are all <0.05, we reject Ho and go with the alternative. We have convincing evidence that there is a significant difference in the mean in terms of perceived opportunities, perceived capabilities, fear of failure rate, entrepreneurial intentions, entrepreneurial employee activity, and the three income levels (low, middle, and high) in 2017 and 2018.

The greatest root is Roy’s with an eigenvalue of 1.54247657. This is the dominant lambda. The other nonzero eigenvalue is 0.07046427. These values are found in the characteristic roots section in the E Inverse H matrix. The (θ) theta value is calculated by doing lambda 1 divided by 1+lambda 1 and this is equal to 1.54247657/ (1+1.54247657) which is equal to 0.6066827078(θ) theta.

Calculations:

α=0.05, n=36, p=5, νH =k-1=3-1=2,

νE=k(n-1) = 3(12-1) =33.

s=min (p, νH ) = min(5,2)=2,

m=1/2(|νH -p|-1) = 1/2(|3-5|-1) = 1/2(|2|-1) = 1/2(2-1) = 1/2(1) =0.5≈1,

N=1/2(νE -p-1) = 1/2(33-5-1) = 1/2 (27) = 13.5

λ1=1.54247657, λ2=0.07046427

| Wilks’ test statistic: Λ=Product (1/1+lambda i) = (1/(1+1.61294) (1/(1+0.07046427) ≈0.3674  Reject H0 if Λ <Λ α(p, νH, νE)=  Λ0.05(5,2,33) =Λ0.05(5,2,40) ==0.617  Since 0.36740 < 0.617, reject H0. | Roy’s test statistic: θ=lambda/(1+lambda)=1.54247657/(1+1.54247657) =0.60668≈0.60668  Reject H0 if θ> θα (s,m,N)  θ0.05(2,1,13.5) = θ0.05(2,1,15) =0.348  Since 0.60668 >0.348 reject H0. |
| --- | --- |
| Pillai’s test statistic: V(s)=i=Sum (lambda i/1+1+lambda i+1) = (1/(1+1.5424)+(1/(1+0.07046427)=(0.3933+0.93417)≈1.3274  Reject H0 if V(s)>V(s)α(s,m,N)  V(s)0.05(2,1,13.5) = 0.455  Since 1.3274 > 0.455, reject H0. | Lawley-Hotelling’s test statistic: U(s)= sum of lambda 1+lambda 2=1.542447657+0.0704627=1.61294  Reject H0 if (E/H) U(s)>U(s)α (p, νH, νE)  (E/H) U(s)= (33/2) (1.61294) =53.227/2 = 26.6135  U(s)0.05(2,5,2,5,33) =U(s)0.05(2,5,33), since, p >Vh, we will use (Vh, P,Ve+Vh-P) for U(s)α (p,νH,νE) , hence U(s)0.05(2,5,33)  = U(s)0.05(5,2,33).  =26.6135 (U(s)0.05(5,2,33))  Since 26.6135 > 4.3743, reject H0. |

Conclusion For One Way MANOVA:

The overall MANOVA test shows that there is a statistically significant difference in means in terms of perceived opportunities, perceived capabilities, fear of failure rate, entrepreneurial intentions, and entrepreneurial employee activity among the three income levels for 2017- 2018. As stated earlier, there is convincing evidence in favor of the alternative hypothesis since the p-value <0.05.

## **Classification Analysis**

Objective: To verify if the 94 countries belonging to three income groups have similar classification categories for entrepreneurship variables as well. Classification analysis has been used to predict the income categories to which each of the 94 countries will belong to, based on the entrepreneurship variables, and calculate the error of misclassification.

Implementation:

(1) Using the method of PROC DISCRIM in SAS, we implemented the apparent error rates and hold-out error rates for the linear, quadratic, and K nearest neighbors classification analysis using the 5 x-variables to see which rate would be the lowest.

(2) Next, we compared the apparent and hold-out error rates using all 7 variables in our data, consisting of 2 y-variables and 5 x-variables, for linear, quadratic, and K nearest neighbors classification methods.

(3) See Appendix A for Classification codes and Appendix B for outputs for the respective error rates.

Methodology: We use linear classification, quadratic classification, hold out, and K nearest neighbors methods of classification to analyze the data in line with our objective. First, we perform these methods for only the five independent variables. Then we repeat these methods including all seven variables, both dependent and independent.

Linear Classification Analysis Method:

Linear classification analysis predicts the classification of each country into one of the three income groups using a linear combination of the five variables mentioned below.

The five variables are: X1 = Perceived opportunities

X2= Perceived capabilities

X3= Fear of failure rate

X4= Entrepreneurial intentions

X5= Entrepreneurial Employee Activity

Assumption:

(1) The main assumption is all the entrepreneurship variables have equal variances (i.e.) Σ1 = Σ 2 = Σ 3 = Σ 4 = Σ 5 ……= Σ k

The number of observations in each of the three groups is not equal. Hence, we take prior probabilities as proportional to get the following linear functions.

The linear combinations for each of the three groups are:

*L*1*(****x****)* = -26.43090 -0.04002 *x*1 +0.52478*x*2 +0.66159*x*3 -0.09389 *x4* +1.26456*x5*

*L*2*(****x****)* = -25.07033 -0.01622 *x*1 +0.50415*x*2 +0.61653*x*3 -0.00817*x4* +0.38170*x5*

*L*3*(****x****)* = -30.84779 +0.00901 *x*1 +0.46397*x*2 +0.73185*x*3+0.05433 *x4* +0.29802*x5*

Based on the above linear classification, x5 seems to make the most contribution to the classification method. The overall error rate of classification is 0.1809.

Limitations:

(1) One of the limitations of the linear classification method is its assumption that all the groups have equal variance.

(2) This may not be true in all cases, specifically in cases that are based on income classification such as this. The resulting classification rules are sensitive to the heterogeneity of covariance matrices.

(3) Observations tend to be classified too frequently into groups whose covariance matrices have larger variances on the diagonal, which may result in higher error rates. Therefore, we compare this error rate with the error rates of other methods of classification to determine which of the methods works best.

Quadratic Classification Method:

In the quadratic classification method, we do not assume equal variance. Instead, we use

*D*2*i (***y***)* = *(***y** − **y***i )’***S**−1*i (***y** − **y***i ), i* = 1*,* 2*, . . . , k,*

With **S***i* in place of **S**pl (pooled variance used in linear classification), the above function cannot be reduced to a linear function of **y,** but remains a quadratic function. Hence rules based on **S***i* are called *quadratic classification rules*.

The error rate using this method is 0.1809 as given below. It is noteworthy that the error rate when we use only 5 independent variables remains the same for both linear and quadratic classification.

Limitations of the quadratic method of classification:

(1) It can be sensitive to outliers. When we have many variables, the quadratic dimensionality makes it complex to interpret the results as compared to linear classification.

(2) Hold out method, therefore, is a good alternative to quadratic and linear classification methods to reduce misclassification.

Hold-out method:

In the holdout procedure, all but one observation is used to compute the classification rule, and this rule is then used to classify the omitted observation. We repeat this procedure for each observation, so that, in a sample of size *N* =Σ*i ni*, each observation is classified by a function based on the other *N* − 1 observations. The computation load is increased because *N* distinct classification procedures must be constructed.

The error rate using this method is 0.2872 as given below. It is noteworthy that the error rate when we use only 5 independent variables is higher for the holdout method than for both linear and quadratic classification, contrary to the general assumption.

To further improve the classification error rate, we included the two dependent variables and looked at the classification error rates across all three methods including all 7 variables – 2 dependent variables, and 5 independent variables.

Linear classification for all seven variables:

The linear combinations of each of the three groups are:

*L*1*(****x****)* = -2705915 -0.02363*x* 1 +0.52887*x*2 +0.64904*x*3 -0.045 *x4* +1.37219*x5*-0.32595y1+0.18646y2

*L*2*(****x****)* = -25.61635 -0.00333 *x* 1 +0.47513*x*2 +0.60579*x*3 +0.00896*x4* +0.3997*x5-0.03026y1+0.21649y2*

*L*3*(****x****)* = -30.84779 +0.03723 *x* 1 +0.46571*x*2 +0.70936*x*3+0.11676*x4* +0.41562*x5-0.33059y1+0.39227y2*

The linear combinations are almost similar for all the five variables except for x4 and x5.

The error rate for linear classification including all seven variables is 0.1915, higher than what it was for just five independent variables.

Quadratic classification for all seven variables: The error rate is the least for quadratic classification including all seven variables is the least ie. 0.0745.

Hold-out method: The error rate for the hold-out method including all seven variables is 0.2553, lower than what it was for just five independent variables, but higher than both quadratic and linear classification methods.

We compare the holdout method using linear classification with the hold-out method using quadratic classification for all 7 variables. The error rate is lower at 0.2340 for holdout with quadratic than linear classification. However, it is higher than plain linear and quadratic methods.

KNN using the hold-out method:

K Nearest Neighbors is a non-parametric classification algorithm that predicts the class of a data point by considering the classes of its k nearest neighbors in the groups. KNN assumes equal variances for all groups. The classification is based on the smallest of an observation from its k nearest neighbors. The distance is calculated using the function

D2i (ynew) = (ynew - yi)’S-1pl (ynew -yi)

We computed the error for different values of K starting from 6 down to 3. The error for K=4 was 0.1809, which happened to be the lowest when we considered 7 variables.

However, when we considered 5 independent variables, K=3 gave the lowest error rate of 0.2340.

Classification Analysis Conclusion:

Since the error rate for quadratic classification using all 7 variables gives the least classification errors, it seems that the three groups do not have equal variances, and all 7 variables are important for the accurate classification of the groups.

## 

## 

## **Multivariate Regression**

Objective: To test if the model overall is significant in predicting total early-stage entrepreneurial activity and established business ownership using all 5 predictor variables. Also, we will test if any predictor variables can be dropped from the model to make it simpler yet provide sufficient results.

Implementation:

(1) PROC REG was used in SAS to test the overall significance of the model consisting of Y1 and Y2 on the left-hand side of the equation and X1, X2, X3, X4, and X5 on the right-hand side.

(2) Next partial tests were conducted using a variety of combinations of which predictor variables to include and which to remove to obtain useful results for our analysis.

(3) Test statistics and critical values were obtained to generate conclusions.

(4) See Appendix A for SAS code on Multivariate Regression and Appendix B for the corresponding output.

Assumptions:

(1) The n=94 observed values of the vector of y’s will have rows corresponding to the values of the p=2 dependent variables. It is assumed the matrix X, containing the values of the 5 x-variables from the 94 values, is fixed. Each y-variable will depend on the x-variables, portraying the need for a matrix column of the different slopes.

(2) The multivariate model will be Y=XB+Ξ, where Y is a nxp=94x2 matrix, X is nx(q+1)=94x6, and B is (q+1)xp=(6x2). The notation Ξ represents the matrix of error terms which will also be an nxp=94x2 matrix.

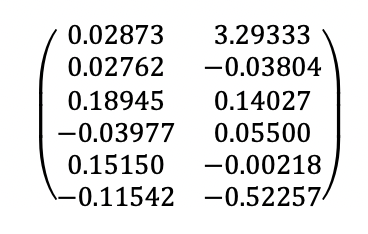
(3) We assume that the expectation of the matrix of random error terms is 0, which means the model is linear and no additional x-variables are necessary to predict the y-variables.

(4) Additionally, we assume the covariance of each observation equals Σ, depicting they all have the same variance-covariance matrix. Lastly, the covariance of each y-variables observation vector equals 0, meaning matrix Y contains uncorrelated rows (Rencher p. 337-339).

(5) Assume a significance level of 0.05 for all tests.

Some useful properties of the least squares estimator matrix include that it is unbiased because the means of repeated random samples have an expectation resembling that of the true mean. Also, each estimator has minimum variance and is all correlated with each other. Since the x-variables are correlated, each least squares estimator within each column of the matrix is also correlated. Similarly, since the y-variables are correlated, each least squares estimator between the columns of the matrix will also be correlated (Rencher p. 341-342).

The least squares estimate for the regression of (y1,y2) on (x1,x2,x3,x4,x5) is given by the following estimated coefficient matrix based on the SAS output:



The eigenvalues of E-1H reveal that the essential rank of 1 is equal to 1 because the largest eigenvalue λ1=1.3935 is the only dominant eigenvalue since 90.67%. The implications of this rank is that the relative power of the four tests is θ>U(s)>Λ>V(s).

Multivariate Regression Method 1 - Test of Overall Regression

We test H0: B1=0 (All βij are zeros, i.e. no x is a useful predictor for any y response) vs Ha: B10 (At least one βij0, i.e. at least one x is a useful predictor for some y response)

Calculations:

α=0.05, n=94, p=2, q=5, νH=q=5, νE=n-q-1=94-5-1=88, s=min(p,q)=min(2,5)=2,

m=(|q-p|-1)=(|5-2|-1)=(|3|-1)=(3-1)=(2)=1, N=(n-q-p-2)=(94-5-2-2)=(85)=42.5

λ1=1.3935, λ2=0.1434

| Wilks’ test statistic: Λ==0.3654  Reject H0 if Λ<Λα(p, νH, νE)= Λα(p, q, n-q-1)  Λ0.05(2,5,88)Λ0.05(2,5,80)=0.798  Since 0.3654<0.798, reject H0. | Roy’s test statistic:  θ=  Reject H0 if θ>θα(s,m,N)  θ0.05(2,1,42.5) θ0.05(2,1,50)=0.130  Since 0.582>0.130, reject H0. |
| --- | --- |
| Pillai’s test statistic:  V(s)=  Reject H0 if V(s)>V(s)α(s,m,N)  Since p-value<0.0001<0.05=α, reject H0. | Lawley-Hotelling’s test statistic: U(s)==1.3935+0.1434=1.5369  Reject H0 if U(s)>U(s)α(p,νH,νE)  U(s)=(1.5369)27.0494  U(s)0.05(2,5,88)U(s)0.05(2,5,100)=3.8557  Since 27.0494>3.8557, reject H0. |

Conclusion for the test of overall regression:

All 4 MANOVA tests reject H0 given α=0.05. Thus, the model is statistically useful. This shows that at least one predictor variable out of Perceived Opportunities, Perceived Capabilities, Fear of failure rate, Entrepreneurial Intentions, or Entrepreneurial Employee Activity is useful for some response variable in our model.

Multivariate Regression Method 2 - Test on Various Subsets

There are 5 partial tests we can consider, including testing the significance of X1 (or X2, or X3, or X4, or X5) adjusted for the other four X’s.

We test H0: Bd=0 vs Ha: Bd0, where Bd has slopes related to X1 (or X2, or X3, or X4, or X5).

When testing the significance of X1 (Perceived opportunities):

Full Model: X1, X2, X3, X4, X5 & Reduced Model (H0): X1

When testing the significance of X2 (Perceived capabilities):

Full Model: X1, X2, X3, X4, X5 & Reduced Model (H0): X2

When testing the significance of X3 (Fear of failure rate):

Full Model: X1, X2, X3, X4, X5 & Reduced Model (H0): X3

When testing the significance of X4 (Entrepreneurial intentions):

Full Model: X1, X2, X3, X4, X5 & Reduced Model (H0): X4

When testing the significance of X5 (Entrepreneurial Employee Activity):

Full Model: X1, X2, X3, X4, X5 & Reduced Model (H0): X5

Calculations:

α=0.05, n=94, p=2, q=5, νH=h=1, νE=n-q-1=94-5-1=88, s=min(p,h)=min(2,1)=1,

m=(|h-p|-1)=(|1-2|-1)=(|-1|-1)=(1-1)=(0)=0, N=(n-q-p-2)=(94-5-2-2)=(85)=42.5

Since νH=h=1, all four tests are equivalent to each other, which means they have exact F transformations and the SAS output provides us with exact p-values that can be used to draw conclusions for each significance test.

Below is a summary using Wilks’ Λ test for each significance test which is calculated using the formula Λ=

Reject H0 if Λ<Λα(p, νH, νE)= Λα(p, h, n-q-1) = Λ0.05(2,1,88)Λ0.05(2,1,80)=0.927

|  | Eigenvalue (λ1) | Λ | F | p-value |
| --- | --- | --- | --- | --- |
| X1 | X2, X3, X4, X5 | 0.02567 | 0.97400331 | 1.16 | 0.3180 |
| X2 | X1, X3, X4, X5 | 0.1498 | 0.86971604 | 6.52 | 0.0023 |
| X3 | X1, X2, X4, X5 | 0.0353 | 0.96593022 | 1.53 | 0.2214 |
| X4 | X1, X2, X3, X5 | 0.2720 | 0.78613784 | 11.83 | < 0.0001 |
| X5 | X1, X2, X3, X4 | 0.0461 | 0.95596441 | 2.00 | 0.1410 |

Conclusion for testing each individual partial subset test:

Given the significance level of α=0.05, variables perceived capabilities and entrepreneurial intentions are significant predictors while perceived opportunities, fear of failure rate, and entrepreneurial employee activity are not, after being adjusted for the other predictor variables. X2 and X4 conclude in favor of the full model, while the rest conclude in favor of their respective reduced models. This suggests X1, X3, and X5 could be dropped from the model to make it simpler and still sufficient.

We wanted to further investigate the significance of different predictor variables.

Testing the significance of (X1, X3) adjusted for the other X’s.

We test H0: Bd=0 vs Ha: Bd0, where Bd has slopes related to (X1, X3).

Full Model: X1, X2, X3, X4, X5 & Reduced Model (H0): X1, X3

Calculations:

α=0.05, n=94, p=2, q=5, νHd=h=2, νE=n-q-1=94-5-1=88, s=min(p,h)=min(2,2)=2,

m=(|h-p|-1)=(|2-2|-1)=(|0|-1)=(-1)=,

N=(n-q-p-2)=(94-5-2-2)=(85)=42.5

λ1= 0.0584, λ2= 0.0000

| Wilks’ test statistic: Λ==0.945  Reject H0 if Λ<Λα(p, νH, νE)= Λα(p, h, n-q-1)  Λ0.05(2,2,88)Λ0.05(2,2,80)=0.888  Since 0.945>0.888, do NOT reject H0. | Roy’s test statistic: θ=  Reject H0 if θ>θα(s,m,N)  θ0.05(2,-0.5,42.5) θ0.05(2,0,50)=0.099  Since 0.055<0.099, do NOT reject H0. |
| --- | --- |
| Pillai’s test statistic: V(s)=  Reject H0 if V(s)>V(s)α(s,m,N)  Since p-value=0.2922>0.05=α, do NOT reject H0. | Lawley-Hotelling’s test statistic: U(s)==0.0584+0.0000 = 0.0584  Reject H0 if U(s)>U(s)α(p,νH,νE)  U(s)=(0.0584)=2.5696  U(s)0.05(2,2,88)U(s)0.05(2,2,100)=4.9628  Since 2.5696<4.9628, do NOT reject H0. |

Conclusion for testing the subset: We do NOT reject H0 for all the tests and conclude in favor of the reduced model containing perceived opportunities and fear of failure rate.

Testing the significance of (X2, X4, X5) adjusted for the other X’s.

We test H0: Bd=0 vs Ha: Bd0, where Bd has slopes related to (X2, X4, X5).

Full Model: X1, X2, X3, X4, X5 & Reduced Model (H0): X2, X4, X5

Calculations:

α=0.05, n=94, p=2, q=5, νHd=h=3, νE=n-q-1=94-5-1=88, s=min(p,h)=min(2,3)=2,

m=(|h-p|-1)=(|3-2|-1)=(|1|-1)=(0)=0,

N=(n-q-p-2)=(94-5-2-2)=(85)=42.5

λ1= 0.8635, λ2= 0.0828

| Wilks’ test statistic: Λ==0.496  Reject H0 if Λ<Λα(p, νH, νE)= Λα(p, h, n-q-1  Λ0.05(2,3,88)Λ0.05(2,3,80)=0.854  Since 0.496<0.854, reject H0. | Roy’s test statistic: θ=  Reject H0 if θ>θα(s,m,N)  θ0.05(2,0,42.5) θ0.05(2,0,50)=0.099  Since 0.463>0.099, reject H0. |
| --- | --- |
| Pillai’s test statistic: V(s)=  Reject H0 if V(s)>V(s)α(s,m,N)  Since p-value<0.0001<0.05=α, reject H0. | Lawley-Hotelling’s test statistic: U(s)==0.8635 +0.0828 = 0.9463  Reject H0 if U(s)>U(s)α(p,νH,νE)  U(s)=(0.9463)=27.7581  U(s)0.05(2,3,88)U(s)0.05(2,3,100)=4.4030  Since 27.7581>4.4030, reject H0. |

Conclusion for testing the subset:

We reject H0 for all the tests and conclude in favor of the full model containing all predictor variables, rather than using the model with just perceived capabilities, entrepreneurial intentions, and entrepreneurial employee activity.

## **Canonical Correlation**

Objective: We want to test the significance of each canonical correlation and be able to identify which ones contribute the most in the model. We also want to see in how many dimensions there exists a significant linear relationship amongst the canonical correlations.

Implementation:

(1) Using PROC CANCORR in SAS, all information about canonical correlations can be obtained.

(2) The input variables will consist of the 5 predictor variables and the yield variables will show the contributions of the 2 y-variables to the canonical correlations.

(3) The calculations and critical values to compare the test statistics were also used to draw conclusions.

(4) See Appendix A for the SAS output used in Canonical Correlation and Appendix B for the respective output.

Assumptions:

(1) One property of canonical correlations is they are scale-invariant on the y-variables or the x-variables.

(2) Another property is that the first canonical correlation is r1, which is considered the maximum correlation when measuring the linear relationship between y and x. This tells us that r1 exceeds the simple correlations and the multiple correlations (Rencher p. 366).

(3) When testing the significance of canonical correlations, it is important to note that the four tests (Wilks’, Pillai’s, Roy’s, and Lawley-Hotelling’s) in general are not equivalent.

(4) Assume a significance level of 0.05.

Limitations:

(1) A limitation is that for all of the canonical correlations excluding the first one, the only test that can be used is Wilks’.

(2) Testing the independence of two sets of variables, the significance of overall multivariate regression, or even the significance of the canonical correlations, are all equivalent tests (Alvin Rencher pgs. 368-369).

(3) Canonical correlation is used to measure the linear relationship between multiple y variables and multiple x variables. They are non-redundant because each one contributes no information to the others.

The canonical correlations between (y1,y2) and (x1,x2,x3,x4,x5) are r1=0.763019 and r2=0.354125.

The standardized coefficients, denoted by ci and di, show the contribution of each variable to the canonical correlations as well as removing the scaling effect of the variables denoted by ai and bi in ui = a′iy and vi = b′ix. The standardized version will correspond to ci = Dyai and di = Dxbi, where D consists of the diagonals of the standard deviations for the respective variables ​​(Rencher p. 371-372). The standardized coefficients for the canonical variates are shown below. The variables that contribute most to the correlation between u1 and v1 are y1 and x4, while the correlation between u2 and v2 is because of y2 and x2, x5.

|  | c1 | c2 |
| --- | --- | --- |
| y1 | 1.1291 | -0.3224 |
| y2 | -0.3173 | 1.1306 |

|  | d1 | d2 |
| --- | --- | --- |
| x1 | 0.1472 | -0.3981 |
| x2 | 0.4578 | 0.6620 |
| x3 | -0.1295 | 0.3510 |
| x4 | 0.5404 | -0.3492 |
| x5 | 0.0267 | -0.6980 |

Testing the significance of each canonical correlation:

When k=1, we test H0: ρ1=ρ2=0 vs Ha: At least ρ10, i.e., test whether there is any significant linear association between y and x.

Calculations:

α=0.05, n=94, p=2, q=5, νH=q=5, νE=n-q-1=94-5-1=88, s=min(p,q)=min(2,5)=2,

m=(|q-p|-1)=(|5-2|-1)=(|3|-1)=(3-1)=(2)=1,

N=(n-q-p-2)=(94-5-2-2)=(85)=42.5

r1=0.763019 and r2=0.354125

| Wilks’ test statistic:  Λ1==(1-(0.763019)2)(1-(0.354125)2)0.365  Reject H0 if Λ1<Λα(p, νH, νE)= Λα(p, q, n-q-1)  Λ0.05(2,5,88)Λ0.05(2,5,80)= 0.798  Since 0.365<0.798, reject H0. | Roy’s test statistic:  θ=r12=(0.763019)2  Reject H0 if θ>θα(s,m,N)  θ0.05(2,1,42.5) θ0.05(2,1,50)=0.130  Since 0.528>0.130, reject H0. |
| --- | --- |
| Pillai’s test statistic:  V(s)=  Reject H0 if V(s)>V(s)α(s,m,N)  Since p-value<0.0001<0.05=α, reject H0. | Lawley-Hotelling’s test statistic:  U(s)=  Reject H0 if U(s)>U(s)α(p,νH,νE)  U(s)=(1.537)27.0512  U(s)0.05(2,5,88)U(s)0.05(2,5,100)=3.8557  Since 27.0512>3.8557, reject H0. |

Conclusion:

Thus, it can be concluded that there is a significant relationship between total early-stage entrepreneurial activity and established business ownership with the predictor variables perceived opportunities, perceived capabilities, fear of failure rate, entrepreneurial intentions, and entrepreneurial employee activity in at least one dimension.

When H0: ρ1=ρ2=0 is rejected, ρ10. This means we will now check what occurs when k=2. We test H0: ρ2=0 vs Ha: ρ20, where the only test statistic that can be calculated is through the use of the Wilks’ test.

Λ2==(1-(0.354125)2)0.875

Reject H0 if Λ2<Λα(p-1, q-1, n-q-2)

Λ0.05(2-1,5-1,94-5-2)=Λ0.05(1,4,87)Λ0.05(1,4,80)=0.889

Since 0.875<0.889, reject H0.

Conclusion for Canonical Correlation:

Thus, conclude that the variables total early-stage entrepreneurial activity and established business ownership and the predictor variables perceived opportunities, perceived capabilities, fear of failure rate, entrepreneurial intentions, and entrepreneurial employee activity are significantly linearly related in two dimensions, which means a significant relationship exists in exactly two dimensions.

## **Overall Conclusions:**

The Global Entrepreneurship Monitor provided us with an intriguing data set that allowed us to extract information about the five predictor variables regarding entrepreneurial behaviors and attitudes, and two response variables regarding the entrepreneurial activity of 47 countries in 2017 and another 47 countries from 2018 data. The World Bank’s data allowed us to classify each country with its respective income level, which was either high, middle, or low. The types of analysis we applied to our data set were MANOVA, Multivariate Regression, Canonical Correlation, and Classification Analysis. For the MANOVA method in the analysis, we gathered that there is a statistically significant difference in the means of the three income levels for 2017 and 2018 in relation to perceived opportunities, perceived capabilities, fear of failure rate, entrepreneurial intentions, and entrepreneurial employee activity. After performing the Classification Analysis on the different income levels, the quadratic classification method using all 7 response and predictor variables in our model had the lowest probability of misclassification. In relation to Multivariate Regression, it can be concluded the overall model is statistically useful, but while looking at the partial tests only perceived capabilities and entrepreneurial intentions appear to be significant predictors of the response variables - Total early-stage Entrepreneurial Activity and Established Business Ownership. The Canonical Correlation method showed there is a significant linear relationship in exactly two dimensions between all independent and dependent variables. Overall, we are able to say the model consisting of the predictor variables Perceived Opportunities, Perceived Capabilities, Fear of failure rate, Entrepreneurial Intentions, and Entrepreneurial Employee Activity, and the response variables Total early-stage Entrepreneurial Activity and Established Business Ownership are significant.

## **Works Cited**

*Gem Global Entrepreneurship Monitor*, www.gemconsortium.org/data. Accessed 6 Apr. 2023.

“How Does the World Bank Classify Countries?” *How Does the World Bank Classify Countries? – World*

*Bank Data Help Desk*, datahelpdesk.worldbank.org/knowledgebase/articles/378834-how-

does-the-world-bank-classify-countries. Accessed 6 Apr. 2023.

Rencher, Alvin C. (2002). *Methods of Multivariate Analysis*. Second Edition. Brigham Young

John Wiley& Sons, Inc., 605 Third Avenue, New York.

## **Appendix A. SAS Codes**

Method 1 MANOVA

TITLE 'MANOVAECONOMYINCOME.DAT EXAMPLE 6.1.7';

DATA Income;

Infile"/home/u63155071/sample2017\_2018.DAT.txt/";

INPUT COUNTRY YEAR INCOME X1 X2 X3 X4 X5 Y1 Y2;

PROC GLM;

CLASS INCOME;

MODEL X1 X2 X3 X4 X5= INCOME;

MANOVA H= INCOME/PRINTE PRINTH MSTAT = EXACT;

Run;

/\* PROJECT CLASSIFICATION LINEAR\*/

TITLE 'CLASSIFICATION 2017 AND 2018';

DATA GEMDATA;

INFILE '/home/u63173211/STA 9705/my\_shared\_file\_links/u63173211/PROJECT/2017-2018 analysis/2017\_2018\_data.DAT';

INPUT ECO YR INCCODE X1 X2 X3 X4 X5;

PROC DISCRIM LIST POOL=yes;

CLASS INCCODE;

VAR X1 X2 X3 X4 X5;

PRIORS proportional;

RUN;

TITLE 'CLASSIFICATION 2017 AND 2018';

DATA GEMDATA;

INFILE '/home/u63173211/STA 9705/my\_shared\_file\_links/u63173211/PROJECT/2017-2018 analysis/2017\_2018\_data.DAT';

INPUT ECO YR INCCODE X1 X2 X3 X4 X5 Y1 Y2;

PROC DISCRIM LIST POOL=yes;

CLASS INCCODE;

VAR X1 X2 X3 X4 X5 Y1 Y2;

PRIORS proportional;

RUN;

/\* PROJECT CLASSIFICATION QUADRATIC\*/

TITLE 'CLASSIFICATION 2017 AND 2018';

DATA GEMDATA;

INFILE '/home/u63173211/STA 9705/my\_shared\_file\_links/u63173211/PROJECT/2017-2018 analysis/2017\_2018\_data.DAT';

INPUT ECO YR INCCODE X1 X2 X3 X4 X5 Y1 Y2;

PROC DISCRIM LIST POOL=no;

CLASS INCCODE;

VAR X1 X2 X3 X4 X5;

PRIORS proportional;

RUN;

\*This is no difference between this method and linear method for just x variables.

TITLE 'CLASSIFICATION 2017 AND 2018';

DATA GEMDATA;

INFILE '/home/u63173211/STA 9705/my\_shared\_file\_links/u63173211/PROJECT/2017-2018 analysis/2017\_2018\_data.DAT';

INPUT ECO YR INCCODE X1 X2 X3 X4 X5 Y1 Y2;

PROC DISCRIM LIST POOL=no;

CLASS INCCODE;

VAR X1 X2 X3 X4 X5 Y1 Y2;

PRIORS proportional;

RUN;

\*This is better than linear method

/\* PROJECT CLASSIFICATION HOLDOUT \*/

TITLE 'CLASSIFICATION 2017 AND 2018';

DATA GEMDATA;

INFILE '/home/u63173211/STA 9705/my\_shared\_file\_links/u63173211/PROJECT/2017-2018 analysis/2017\_2018\_data.DAT';

INPUT ECO YR INCCODE X1 X2 X3 X4 X5 Y1 Y2;

PROC DISCRIM DATA=GEMDATA POOL=YES LIST CROSSVALIDATE;

CLASS INCCODE;

VAR X1 X2 X3 X4 X5;

PRIORS proportional;

TITLE 'Discriminant Analysis of GEM Data for only X variables';

RUN;

/\* PROJECT CLASSIFICATION \*/

TITLE 'CLASSIFICATION 2017 AND 2018';

DATA GEMDATA;

/\* MULTIVARIATE REGRESSION \*/

DATA INCOME;

INFILE '/home/u49642016/STA 9705/2017\_2018\_data.DAT';

INPUT CODE YEAR ECONOMY X1 X2 X3 X4 X5 Y1 Y2;

RUN;

PROC REG;

MODEL Y1 Y2 = X1 X2 X3 X4 X5;

OVERALL: MTEST /PRINT CANPRINT MSTAT=EXACT;

TITLE 'Multivariate Regression';

RUN;

PROC REG;

MODEL Y1 Y2 = X1 X2 X3 X4 X5;

PartialX1: MTEST X1/PRINT CANPRINT MSTAT=EXACT;

PartialX2: MTEST X2/PRINT CANPRINT MSTAT=EXACT;

PartialX3: MTEST X3/PRINT CANPRINT MSTAT=EXACT;

PartialX4: MTEST X4/PRINT CANPRINT MSTAT=EXACT;

PartialX5: MTEST X5/PRINT CANPRINT MSTAT=EXACT;

PartialX13: MTEST X1,X3/PRINT CANPRINT MSTAT=EXACT;

PartialX245: MTEST X2,X4,X5/PRINT CANPRINT MSTAT=EXACT;

RUN;

## 

/\* CANONICAL CORRELATION \*/

FILENAME INCOME '/home/u49642016/STA 9705/2017\_2018\_data.DAT';

TITLE 'Canonical Correlation';

DATA INCOME;

INFILE INCOME;

INPUT CODE YEAR ECONOMY X1 X2 X3 X4 X5 Y1 Y2;

PROC CANCORR ALL

VPREFIX = INPUT VNAME = 'INPUT VARIABLES' MSTAT=EXACT

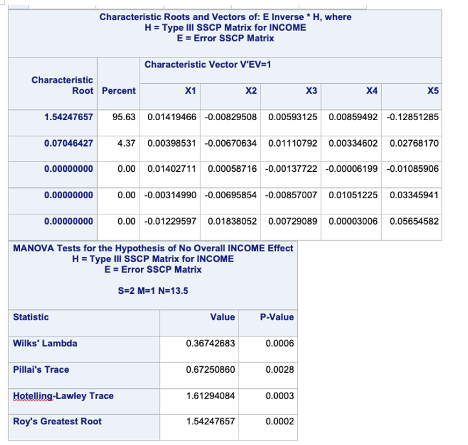
WPREFIX = YIELD WNAME = 'YIELD VARIABLES' MSTAT=EXACT;

WITH Y1 Y2;

VAR X1 X2 X3 X4 X5;

RUN;

## **Appendix B. Relevant Outputs**



| **Total Sample Size** | 94 | **DF Total** | 93 |
| --- | --- | --- | --- |
| **Variables** | 5 | **DF Within Classes** | 91 |
| **Classes** | 3 | **DF Between Classes** | 2 |

**CLASSIFICATION 2017 AND 2018**

**The DISCRIM Procedure**

| **Number of Observations Read** | 94 |
| --- | --- |
| **Number of Observations Used** | 94 |

| **Class Level Information** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **INCCODE** | **Variable Name** | **Frequency** | **Weight** | **Proportion** | **Prior Probability** |
| **1** | 1 | 60 | 60.0000 | 0.638298 | 0.638298 |
| **2** | 2 | 22 | 22.0000 | 0.234043 | 0.234043 |
| **3** | 3 | 12 | 12.0000 | 0.127660 | 0.127660 |

| **Linear Discriminant Function for INCCODE** | | | |
| --- | --- | --- | --- |
| **Variable** | **1** | **2** | **3** |
| **Constant** | -26.43090 | -25.07033 | -30.84779 |
| **X1** | -0.04002 | -0.01622 | 0.00901 |
| **X2** | 0.52478 | 0.50415 | 0.46397 |
| **X3** | 0.66159 | 0.61653 | 0.73185 |
| **X4** | -0.09389 | -0.00817 | 0.05433 |
| **X5** | 1.26456 | 0.38170 | 0.29802 |

**CLASSIFICATION 2017 AND 2018**

**The DISCRIM Procedure**

**Classification Summary for Calibration Data: WORK.GEMDATA**

**Resubstitution Summary using Linear Discriminant Function**

| **Number of Observations and Percent Classified into INCCODE** | | | | |
| --- | --- | --- | --- | --- |
| **From INCCODE** | **1** | **2** | **3** | **Total** |
| **1** | 53  88.33 | 6  10.00 | 1  1.67 | 60  100.00 |
| **2** | 5  22.73 | 16  72.73 | 1  4.55 | 22  100.00 |
| **3** | 0  0.00 | 4  33.33 | 8  66.67 | 12  100.00 |
| **Total** | 58  61.70 | 26  27.66 | 10  10.64 | 94  100.00 |
| **Priors** | 0.6383 | 0.23404 | 0.12766 |  |

| **Error Count Estimates for INCCODE** | | | | |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **Total** |
| **Rate** | 0.1167 | 0.2727 | 0.3333 | 0.1809 |
| **Priors** | 0.6383 | 0.2340 | 0.1277 |  |

**CLASSIFICATION 2017 AND 2018**

**The DISCRIM Procedure**

**Classification Summary for Calibration Data: WORK.GEMDATA**

**Resubstitution Summary using Quadratic Discriminant Function**

| **Number of Observations and Percent Classified into INCCODE** | | | | |
| --- | --- | --- | --- | --- |
| **From INCCODE** | **1** | **2** | **3** | **Total** |
| **1** | 53  88.33 | 6  10.00 | 1  1.67 | 60  100.00 |
| **2** | 3  13.64 | 16  72.73 | 3  13.64 | 22  100.00 |
| **3** | 2  16.67 | 2  16.67 | 8  66.67 | 12  100.00 |
| **Total** | 58  61.70 | 24  25.53 | 12  12.77 | 94  100.00 |
| **Priors** | 0.6383 | 0.23404 | 0.12766 |  |

| **Error Count Estimates for INCCODE** | | | | |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **Total** |
| **Rate** | 0.1167 | 0.2727 | 0.3333 | 0.1809 |
| **Priors** | 0.6383 | 0.2340 | 0.1277 |  |

**Discriminant Analysis of GEM Data for only X variables**

**The DISCRIM Procedure**

**Classification Summary for Calibration Data: WORK.GEMDATA**

**Cross-validation Summary using Linear Discriminant Function**

| **Number of Observations and Percent Classified into INCCODE** | | | | |
| --- | --- | --- | --- | --- |
| **From INCCODE** | **1** | **2** | **3** | **Total** |
| **1** | 51  85.00 | 7  11.67 | 2  3.33 | 60  100.00 |
| **2** | 7  31.82 | 13  59.09 | 2  9.09 | 22  100.00 |
| **3** | 2  16.67 | 7  58.33 | 3  25.00 | 12  100.00 |
| **Total** | 60  63.83 | 27  28.72 | 7  7.45 | 94  100.00 |
| **Priors** | 0.6383 | 0.23404 | 0.12766 |  |

| **Error Count Estimates for INCCODE** | | | | |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **Total** |
| **Rate** | 0.1500 | 0.4091 | 0.7500 | 0.2872 |
| **Priors** | 0.6383 | 0.2340 | 0.1277 |  |

| **Linear Discriminant Function for INCCODE** | | | |
| --- | --- | --- | --- |
| **Variable** | **1** | **2** | **3** |
| **Constant** | -27.05915 | -25.61635 | -32.31683 |
| **X1** | -0.02363 | -0.00333 | 0.03723 |
| **X2** | 0.55887 | 0.47513 | 0.46571 |
| **X3** | 0.64904 | 0.60579 | 0.70936 |
| **X4** | -0.04500 | 0.00896 | 0.11676 |
| **X5** | 1.37219 | 0.39997 | 0.41562 |
| **Y1** | -0.32595 | -0.03026 | -0.33059 |
| **Y2** | 0.18646 | 0.21649 | 0.39227 |

**CLASSIFICATION 2017 AND 2018**

**The DISCRIM Procedure**

**Classification Summary for Calibration Data: WORK.GEMDATA**

**Resubstitution Summary using Linear Discriminant Function**

| **Number of Observations and Percent Classified into INCCODE** | | | | |
| --- | --- | --- | --- | --- |
| **From INCCODE** | **1** | **2** | **3** | **Total** |
| **1** | 55  91.67 | 4  6.67 | 1  1.67 | 60  100.00 |
| **2** | 5  22.73 | 14  63.64 | 3  13.64 | 22  100.00 |
| **3** | 1  8.33 | 4  33.33 | 7  58.33 | 12  100.00 |
| **Total** | 61  64.89 | 22  23.40 | 11  11.70 | 94  100.00 |
| **Priors** | 0.6383 | 0.23404 | 0.12766 |  |

| **Error Count Estimates for INCCODE** | | | | |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **Total** |
| **Rate** | 0.0833 | 0.3636 | 0.4167 | 0.1915 |
| **Priors** | 0.6383 | 0.2340 | 0.1277 |  |

**CLASSIFICATION 2017 AND 2018**

**The DISCRIM Procedure**

**Classification Summary for Calibration Data: WORK.GEMDATA**

**Resubstitution Summary using Quadratic Discriminant Function**

| **Number of Observations and Percent Classified into INCCODE** | | | | |
| --- | --- | --- | --- | --- |
| **From INCCODE** | **1** | **2** | **3** | **Total** |
| **1** | 56  93.33 | 4  6.67 | 0  0.00 | 60  100.00 |
| **2** | 1  4.55 | 21  95.45 | 0  0.00 | 22  100.00 |
| **3** | 1  8.33 | 1  8.33 | 10  83.33 | 12  100.00 |
| **Total** | 58  61.70 | 26  27.66 | 10  10.64 | 94  100.00 |
| **Priors** | 0.6383 | 0.23404 | 0.12766 |  |

| **Error Count Estimates for INCCODE** | | | | |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **Total** |
| **Rate** | 0.0667 | 0.0455 | 0.1667 | 0.0745 |
| **Priors** | 0.6383 | 0.2340 | 0.1277 |  |

**Discriminant Analysis of GEM Data**

**The DISCRIM Procedure**

**Classification Summary for Calibration Data: WORK.GEMDATA**

**Cross-validation Summary using Linear Discriminant Function**

| **Number of Observations and Percent Classified into INCCODE** | | | | |
| --- | --- | --- | --- | --- |
| **From INCCODE** | **1** | **2** | **3** | **Total** |
| **1** | 54  90.00 | 4  6.67 | 2  3.33 | 60  100.00 |
| **2** | 5  22.73 | 12  54.55 | 5  22.73 | 22  100.00 |
| **3** | 3  25.00 | 5  41.67 | 4  33.33 | 12  100.00 |
| **Total** | 62  65.96 | 21  22.34 | 11  11.70 | 94  100.00 |
| **Priors** | 0.6383 | 0.23404 | 0.12766 |  |

| **Error Count Estimates for INCCODE** | | | | |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **Total** |
| **Rate** | 0.1000 | 0.4545 | 0.6667 | 0.2553 |
| **Priors** | 0.6383 | 0.2340 | 0.1277 |  |

**Discriminant Analysis of GEM Data**

**The DISCRIM Procedure**

**Classification Summary for Calibration Data: WORK.GEMDATA**

**Cross-validation Summary using Quadratic Discriminant Function**

| **Number of Observations and Percent Classified into INCCODE** | | | | |
| --- | --- | --- | --- | --- |
| **From INCCODE** | **1** | **2** | **3** | **Total** |
| **1** | 55  91.67 | 5  8.33 | 0  0.00 | 60  100.00 |
| **2** | 6  27.27 | 14  63.64 | 2  9.09 | 22  100.00 |
| **3** | 6  50.00 | 3  25.00 | 3  25.00 | 12  100.00 |
| **Total** | 67  71.28 | 22  23.40 | 5  5.32 | 94  100.00 |
| **Priors** | 0.6383 | 0.23404 | 0.12766 |  |

| **Error Count Estimates for INCCODE** | | | | |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **Total** |
| **Rate** | 0.0833 | 0.3636 | 0.7500 | 0.2340 |
| **Priors** | 0.6383 | 0.2340 | 0.1277 |  |

**Discriminant Analysis of GEM Data for X only**

**The DISCRIM Procedure**

**Classification Summary for Calibration Data: WORK.GEMDATA**

**Cross-validation Summary using 3 Nearest Neighbors**

| **Number of Observations and Percent Classified into INCCODE** | | | | |
| --- | --- | --- | --- | --- |
| **From INCCODE** | **1** | **2** | **3** | **Total** |
| **1** | 56  93.33 | 2  3.33 | 2  3.33 | 60  100.00 |
| **2** | 5  22.73 | 15  68.18 | 2  9.09 | 22  100.00 |
| **3** | 2  16.67 | 4  33.33 | 6  50.00 | 12  100.00 |
| **Total** | 63  67.02 | 21  22.34 | 10  10.64 | 94  100.00 |
| **Priors** | 0.6383 | 0.23404 | 0.12766 |  |

| **Error Count Estimates for INCCODE** | | | | |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **Total** |
| **Rate** | 0.0667 | 0.3182 | 0.5000 | 0.1809 |
| **Priors** | 0.6383 | 0.2340 | 0.1277 |  |

**Discriminant Analysis of GEM Data**

**The DISCRIM Procedure**

**Classification Summary for Calibration Data: WORK.GEMDATA**

**Cross-validation Summary using 4 Nearest Neighbors**

| **Number of Observations and Percent Classified into INCCODE** | | | | |
| --- | --- | --- | --- | --- |
| **From INCCODE** | **1** | **2** | **3** | **Total** |
| **1** | 54  90.00 | 3  5.00 | 3  5.00 | 60  100.00 |
| **2** | 7  31.82 | 13  59.09 | 2  9.09 | 22  100.00 |
| **3** | 4  33.33 | 3  25.00 | 5  41.67 | 12  100.00 |
| **Total** | 65  69.15 | 19  20.21 | 10  10.64 | 94  100.00 |
| **Priors** | 0.6383 | 0.23404 | 0.12766 |  |

| **Error Count Estimates for INCCODE** | | | | |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **Total** |
| **Rate** | 0.1000 | 0.4091 | 0.5833 | 0.2340 |
| **Priors** | 0.6383 | 0.2340 | 0.1277 |  |

