Project 2 – Prediction Of Diabetes Status Using A Subset Of The Pima Indian Diabetes Data Set

Samuel Coyne, Hari Sanadhya, Joseph Denney

Problem Statement

29 million Americans are affected by diabetes, and their health care costs are 2.3 times higher than those of people without diabetes (1). The development of a way to accurately predict diabetes status using readily available clinical data would lead to earlier diagnosis and lower treatment costs for diabetics. 86 million adults in the United States have prediabetes, and 90% of them are unaware of their prediabetic status (2). Being able to empower clinicians to have a predictive model of diabetes risk would be helpful in early detection and management. While we won't be attaining that goal in this document, we do find value in working on a problem that has real world implications. Our goal here is to attempt to predict the diabetes status of a patient based on 8 clinical data points. The data used is the Pima Indians Diabetes Data Set from the University of California, Irvine Machine Learning Repository (3).

Data Set Description

The data used is a subset of a larger database originally collected by researchers at the National Institute of Diabetes and Digestive and Kidney Diseases. The subset of data we worked with are for female Native Americans of the Pima Tribe 21 years of age and older. The original data set has 768 observations with 8 component variables which were chosen because they have been found to be significant risk factors for diabetes among Pimas or other populations and 1 response variable. (4)

i. **pregnancies**: Number of times pregnant
ii. **gtt**: Plasma glucose concentration at 2 hours in an oral glucose tolerance test
iii. **bp**: Diastolic blood pressure (mm Hg)
iv. **skinfold**: Triceps skin fold thickness (mm)
v. **insulin**: 2-Hour serum insulin (mu U/ml)
vi. **bmi**: Body mass index (weight in kg/(height in m)^2)
vii. **pedigree**: Diabetes pedigree function
viii. **age**: Age of the patient (years)
ix. **diabetic**: class values 0 or 1. Patients with diabetes are indicated by '1', whereas those without diabetes are indicated by '0.'

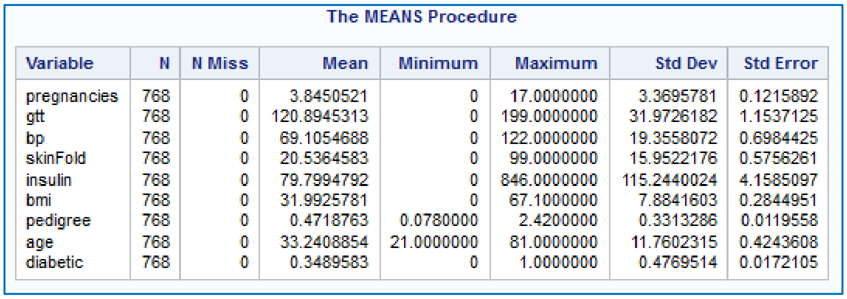
Exploratory Data Analysis

Data Validity and Missing Values

We began our analysis by looking for missing observations for each variable. In this data set, it appears missing values were coded with the value 0. Unfortunately, 0 is a valid observation in some of the variables. Examination of the variable descriptions and application of domain knowledge shows for variables pregnancies and diabetic, 0 is a valid observation. For all other variables, 0 is a missing observation. For example, a skinfold measurement of 0 is clearly erroneous. We re-coded the 0's in those fields to NA's in SAS. See lines 42-48 in the code appendix for the exact process used. Proc means (Figure 1) shows the number of missing values and the changes in variability after recoding to missing values.

Recoding 0 to NA presents us with other problems to now solve, specifically;

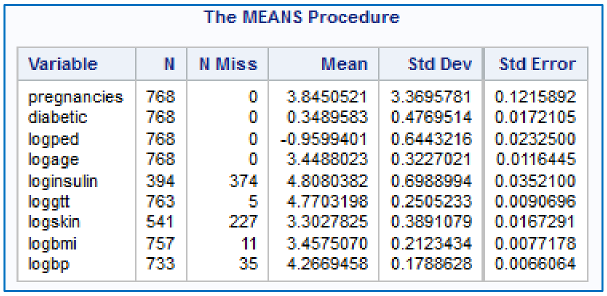
- a reduction in the sample size since SAS will ignore all the records having missing values and in our case almost 50% (376 out of 768) of the records have missing values and;
- a reduction in sample size causes an increase in standard error (figure 1) which may lead to these variables being less significant.



The MEANS Procedure

Variable	N	N Miss	Mean	Minimum	Maximum	Std Dev	Std Error
pregnancies	768	0	3.8450521	0	17.0000000	3.3695781	0.1215892
gtt	768	0	120.8945313	0	199.0000000	31.9726182	1.1537125
bp	768	0	69.1054688	0	122.0000000	19.3558072	0.6984425
skinFold	768	0	20.5364583	0	99.0000000	15.9522176	0.5756261
insulin	768	0	79.7994792	0	846.0000000	115.2440024	4.1585097
bmi	768	0	31.9925781	0	67.1000000	7.8841603	0.2844951
pedigree	768	0	0.4718763	0.0780000	2.4200000	0.3313286	0.0119558
age	768	0	33.2408854	21.0000000	81.0000000	11.7602315	0.4243608
diabetic	768	0	0.3489583	0	1.0000000	0.4769514	0.0172105

Figure 1 – PROC MEANS output original data with miscoded 0's



The MEANS Procedure

Variable	N	N Miss	Mean	Std Dev	Std Error
pregnancies	768	0	3.8450521	3.3695781	0.1215892
diabetic	768	0	0.3489583	0.4769514	0.0172105
logped	768	0	-0.9599401	0.6443216	0.0232500
logage	768	0	3.4488023	0.3227021	0.0116445
loginsulin	394	374	4.8080382	0.6988994	0.0352100
loggtt	763	5	4.7703198	0.2505233	0.0090696
logskin	541	227	3.3027825	0.3891079	0.0167291
logbmi	757	11	3.4575070	0.2123434	0.0077178
logbp	733	35	4.2669458	0.1788628	0.0066064

Figure 2 – PROC MEANS after conversion to NA

Based on these findings, we will need to impute the missing values. This process will be discussed fully in the imputation section following.

Testing Assumptions

Mulitvariate Normality -After completing the data validity steps, we then moved to testing the assumptions: multivariate normality and equality of covariance matrices. Our EDA process revealed the explanatory variables are not all normally distributed. Figure 3 indicates a few concerns: notice the skewness in the number of pregnancies, diabetes pedigree function, and age (years). Slight skewness does exist in several other variables such as insulin and plasma glucose concentration (gtt). Realizing our problematic variables are right-skewed and to meet the normality assumption, we applied a logarithmic transformation to our variables.

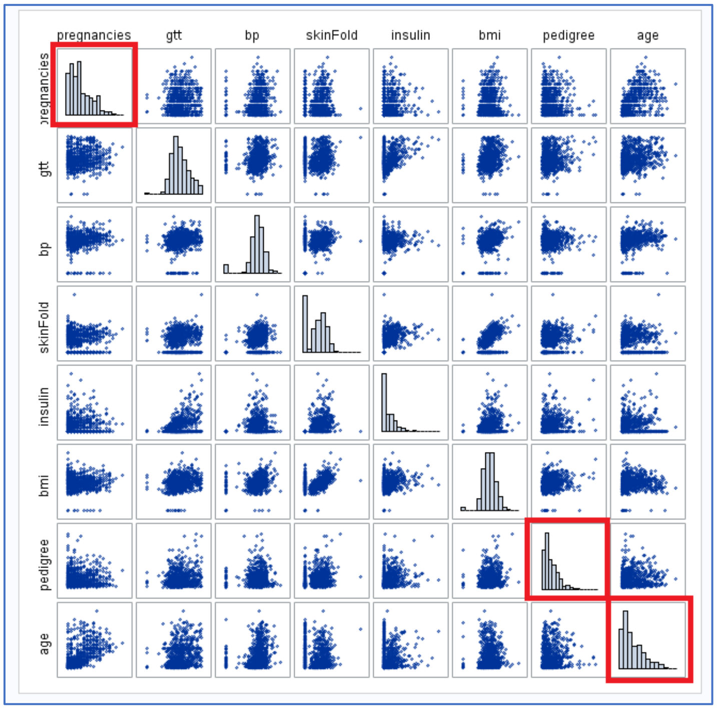
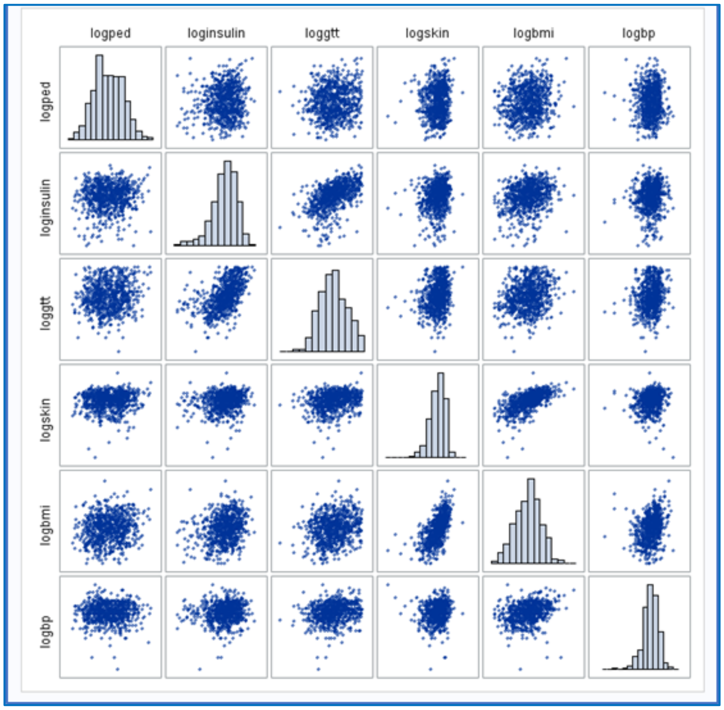


Figure 3 – Scatterplot of raw data with missing values showing normality issues

Furthermore, the number of pregnancies was deleted from the analysis due to the presence of '0' as a plausible and observed value. The age variable, however, is not aided by a logarithmic transformation. As with pregnancies, we will delete age (years).

Combining the issues of missing data and non-normality required us to decide regarding the order of operations. We compared the results of imputing the data then transforming the variables (Figure 4) to the inverse of transforming first, then imputing (Figure 5). After reviewing the results of both methods, we did the data imputation and then log transformed the resulting data set. As figure 4 indicates, this order of operations significantly improves the distribution of the biometric variables.



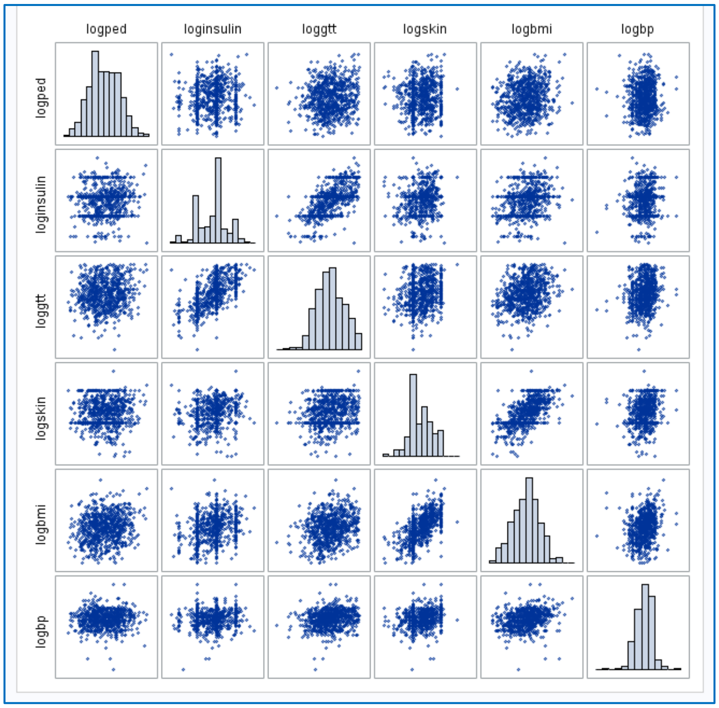


Figure 4 – Data imputed then log transformed

Figure 5 – Data log transformed then imputed

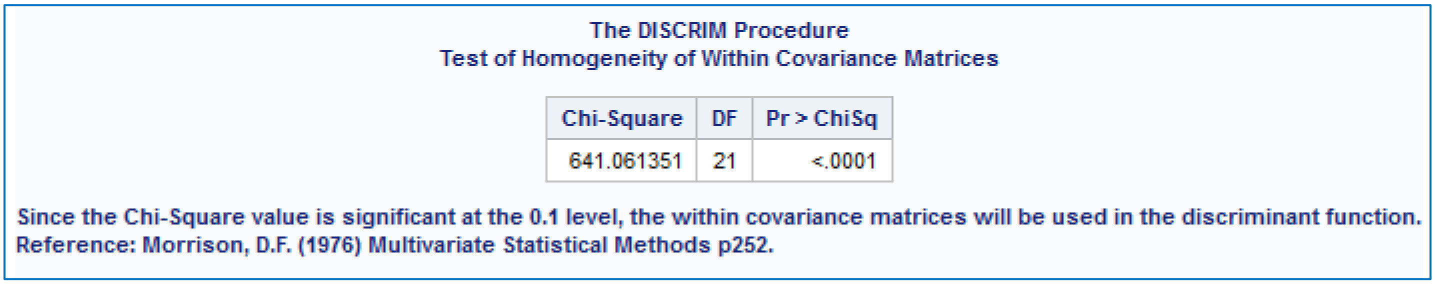
We can now conclude that the normality assumption is met. Deleting age and number of pregnancies did not have much of an effect on our prediction ability, so we left those variables out of the equation.

Equality of Covariance - To investigate the equal covariance matrix assumption, a test of homogeneity of within covariance matrices was conducted (Bartlett's test for homogeneity). For this test, we set up our hypotheses as follows:

H_0: covariance matrices are equal

H_a: covariance matrices are not equal for one or more pair.

We find the results to be significant at the 0.1 significance level (p-value < .0001). As such, we will reject the null hypothesis and conclude the covariance matrices are not equal. Figure 6 formalizes the results.



The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
641.061351	21	<.0001

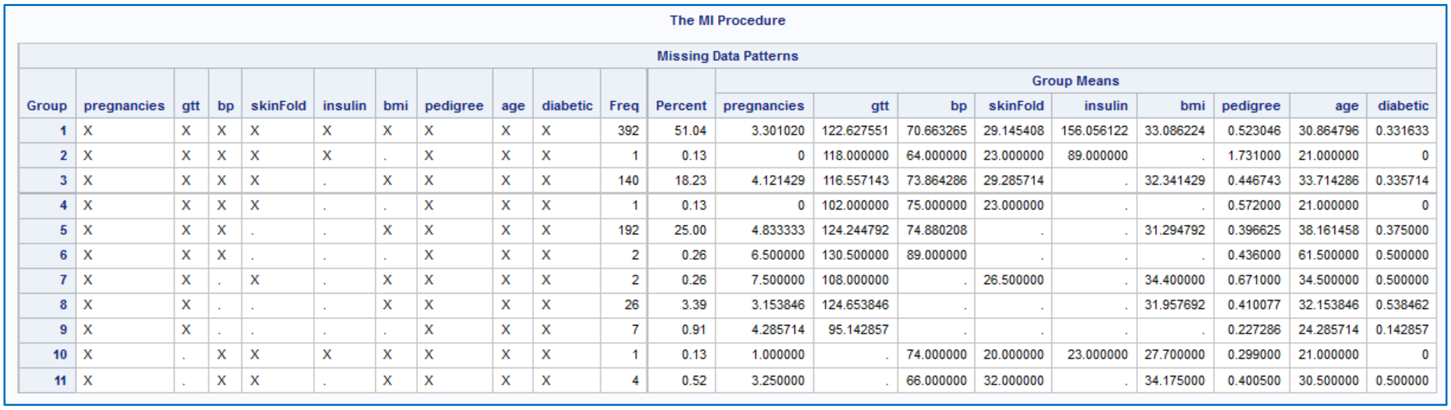
Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Figure 6 – Output of Proc Discrim showing rejection of null hypothesis

For our analysis, that finding suggests that we will implement quadratic discriminant analysis instead of linear discriminant analysis because we do not meet the equality of covariance matrices assumption.

Imputation

To solve the issues introduced by accurately coding missing values, we utilized proc mi (see code lines 65-68) to visualize any patterns in the missing data. Analyzing the variation in group means for various groups in figure 7 it appears each group has variation in group mean values for the complete variables indicating this to be a case of Missing at Random (MAR) where the missing values can be determined by using the complete variables in the study.



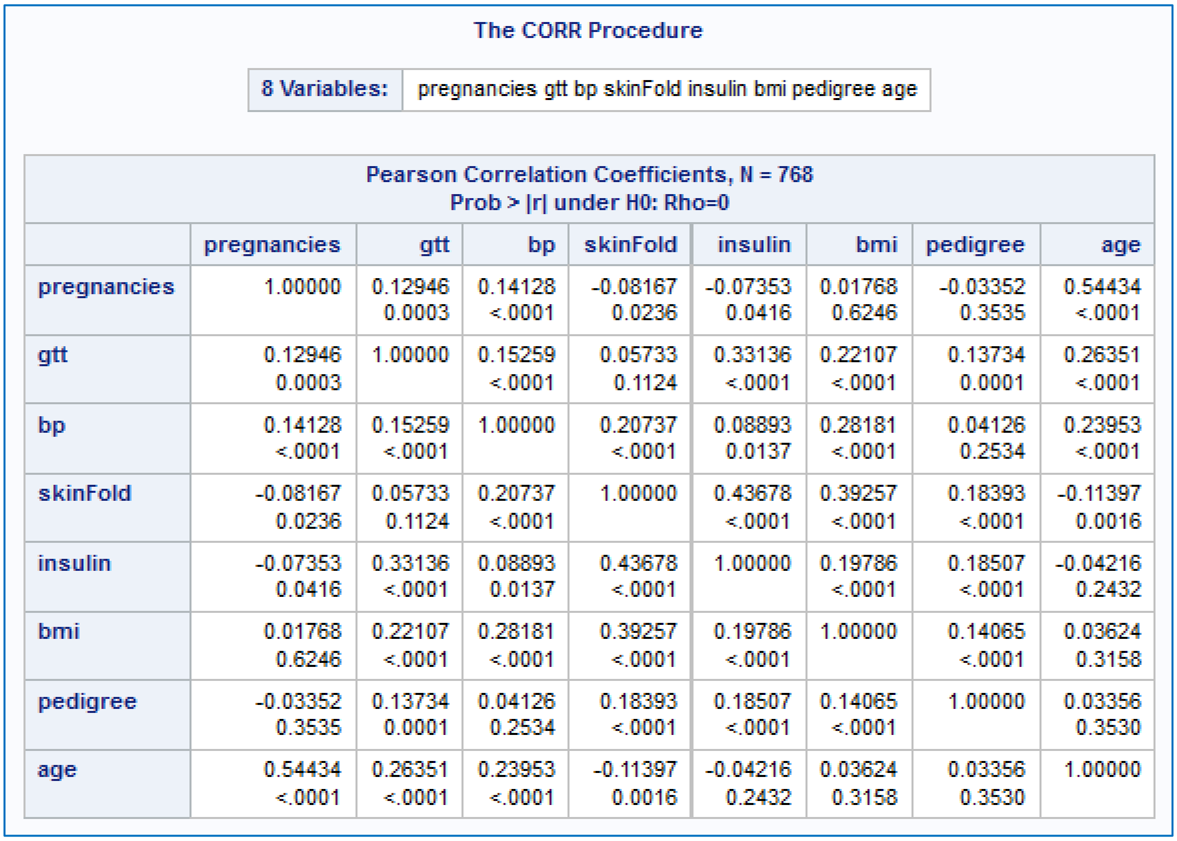
The MI Procedure

Missing Data Patterns

												Group Means								
Group	pregnancies	gtt	bp	skinFold	insulin	bmi	pedigree	age	diabetic	Freq	Percent	pregnancies	gtt	bp	skinFold	insulin	bmi	pedigree	age	diabetic
1	X	X	X	X	X	X	X	X	X	392	51.04	3.301020	122.627551	70.663265	29.145408	156.056122	33.086224	0.523046	30.864796	0.331633
2	X	X	X	X	X	.	X	X	X	1	0.13	0	118.000000	64.000000	23.000000	89.000000	.	1.731000	21.000000	0
3	X	X	X	X	.	X	X	X	X	140	18.23	4.121429	116.557143	73.864286	29.285714	.	32.341429	0.446743	33.714286	0.335714
4	X	X	X	X	.	.	X	X	X	1	0.13	0	102.000000	75.000000	23.000000	.	.	0.572000	21.000000	0
5	X	X	X	.	.	X	X	X	X	192	25.00	4.833333	124.244792	74.880208	.	.	31.294792	0.396625	38.161458	0.375000
6	X	X	X	.	.	.	X	X	X	2	0.26	6.500000	130.500000	89.000000	.	.	.	0.436000	61.500000	0.500000
7	X	X	.	X	.	X	X	X	X	2	0.26	7.500000	108.000000	.	26.500000	.	34.400000	0.671000	34.500000	0.500000
8	X	X	.	.	.	X	X	X	X	26	3.39	3.153846	124.653846	.	.	.	31.957692	0.410077	32.153846	0.538462
9	X	X	.	.	.	.	X	X	X	7	0.91	4.285714	95.142857	.	.	.	.	0.227286	24.285714	0.142857
10	X	.	X	X	X	X	X	X	X	1	0.13	1.000000	.	74.000000	20.000000	23.000000	27.700000	0.299000	21.000000	0
11	X	.	X	X	.	X	X	X	X	4	0.52	3.250000	.	66.000000	32.000000	.	34.175000	0.400500	30.500000	0.500000

Figure 7 – Results of Proc MI to analyze for patterns in missing data

Using proc mi again, we created a new data set on the log transformed data containing 19200 observations by running 25 imputations. We chose 25 imputations based on information provided by SAS regarding confidence and number of imputations (5). Figure 8 shows that insulin has strong correlation with gtt and skinfold. Also, skinfold has strong correlation with bp, bmi, pedigree, age and insulin. Due to the presence of the correlation, we can confirm the missing values in the data to be Missing at Random (MAR) type where the missing values can be accounted for by variables where there is complete information.



The CORR Procedure

8 Variables:	pregnancies gtt bp skinFold insulin bmi pedigree age

Pearson Correlation Coefficients, N = 768
Prob > |r| under H0: Rho=0

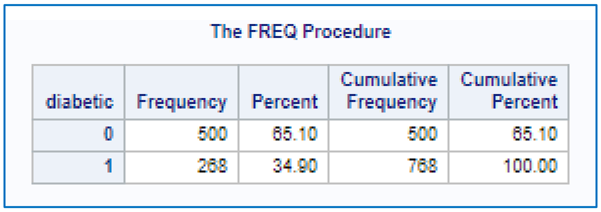
	pregnancies	gtt	bp	skinFold	insulin	bmi	pedigree	age
pregnancies	1.00000	0.12946 0.0003	0.14128 <.0001	-0.08167 0.0236	-0.07353 0.0416	0.01768 0.6246	-0.03352 0.3535	0.54434 <.0001
gtt	0.12946 0.0003	1.00000	0.15259 <.0001	0.05733 0.1124	0.33136 <.0001	0.22107 <.0001	0.13734 0.0001	0.26351 <.0001
bp	0.14128 <.0001	0.15259 <.0001	1.00000	0.20737 <.0001	0.08893 0.0137	0.28181 <.0001	0.04126 0.2534	0.23953 <.0001
skinFold	-0.08167 0.0236	0.05733 0.1124	0.20737 <.0001	1.00000	0.43678 <.0001	0.39257 <.0001	0.18393 <.0001	-0.11397 0.0016
insulin	-0.07353 0.0416	0.33136 <.0001	0.08893 0.0137	0.43678 <.0001	1.00000	0.19786 <.0001	0.18507 <.0001	-0.04216 0.2432
bmi	0.01768 0.6246	0.22107 <.0001	0.28181 <.0001	0.39257 <.0001	0.19786 <.0001	1.00000	0.14065 <.0001	0.03624 0.3158
pedigree	-0.03352 0.3535	0.13734 0.0001	0.04126 0.2534	0.18393 <.0001	0.18507 <.0001	0.14065 <.0001	1.00000	0.03356 0.3530
age	0.54434 <.0001	0.26351 <.0001	0.23953 <.0001	-0.11397 0.0016	-0.04216 0.2432	0.03624 0.3158	0.03356 0.3530	1.00000

Figure 8– Proc Corr output showing importance of skinfold and insulin

Analysis

With the assumptions addressed, the stage is set to proceed with a discriminant analysis. The goal is to predict whether a subject is diabetic (1) or not diabetic (0) as sampled from the female Pima Indian population. Discriminant analysis seeks to separate classes. In this analysis, the desired class separation should be between diabetic and not diabetic.

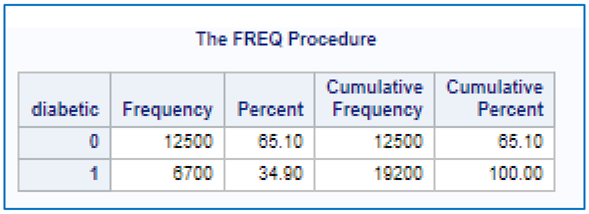
To ensure we have a robust solution, we will conduct the discriminant analysis as a comparison: model results will be presented that were derived using one imputation (n=768), referred to again as Model 1 and another derived from twenty-five imputations (n=19200), Model 2. This comparison is useful for interpretation of the model results and the plausibility of the imputed values. A basic frequency analysis reveals that we are not working with a balanced data set. In fact, the percent difference between the two classes is surprising: figures 9 and 10 tell us that 65.10% of the sample does not have diabetes and 34.90% are diabetic. This insight will be useful in model construction, as it will allow us to specify these probabilities in the discriminant analysis.



The FREQ Procedure

diabetic	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	500	65.10	500	65.10
1	268	34.90	768	100.00

Figure 9 – Proc Freq for Model 1



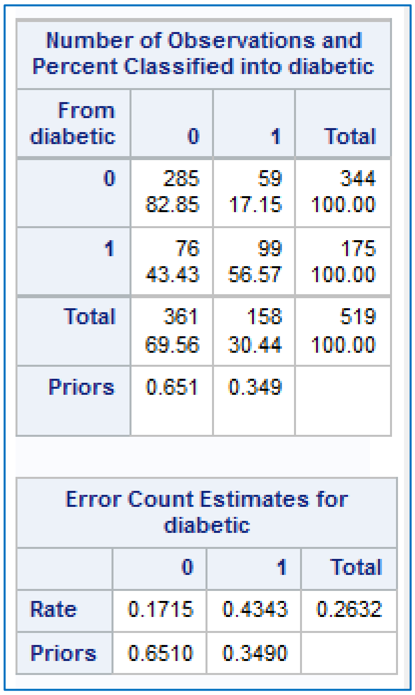
The FREQ Procedure

diabetic	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	12500	65.10	12500	65.10
1	6700	34.90	19200	100.00

Figure 10 – Proc Freq for Model 2

Both models are constructed by beginning with the transformed master data set. We conducted a random sample consisting of 70% of the master data observations; the remaining 30% of observations were selected for the test/hold-out data set. While we are equally interested in the train/test model development phases, the indication of predictive strength is demonstrated by the test data performance. This hold-out data is our best metric for how the model will perform when put into a production setting.

The discriminant analysis is performed first on the train data utilizing the previously found prior probabilities (figure11). The train data sets are comprised of n=519 and n=13354. The resulting confusion matrices can be seen in figure 12 The results are indicative of our unbalanced data set. We first notice the consistency between the data sets derived from one imputation and twenty-five imputations.



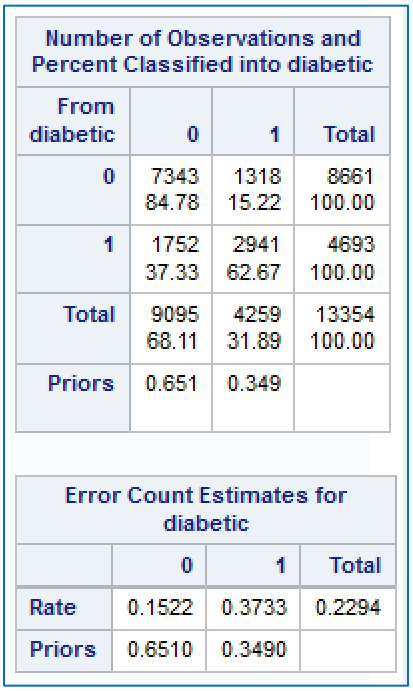
Number of Observations and Percent Classified into diabetic

From diabetic	0	1	Total
0	285 82.85	59 17.15	344 100.00
1	76 43.43	99 56.57	175 100.00
Total	361 69.56	158 30.44	519 100.00
Priors	0.651	0.349	

Error Count Estimates for diabetic

	0	1	Total
Rate	0.1715	0.4343	0.2632
Priors	0.6510	0.3490	

Figure 11 – Training Data Model 1



Number of Observations and Percent Classified into diabetic

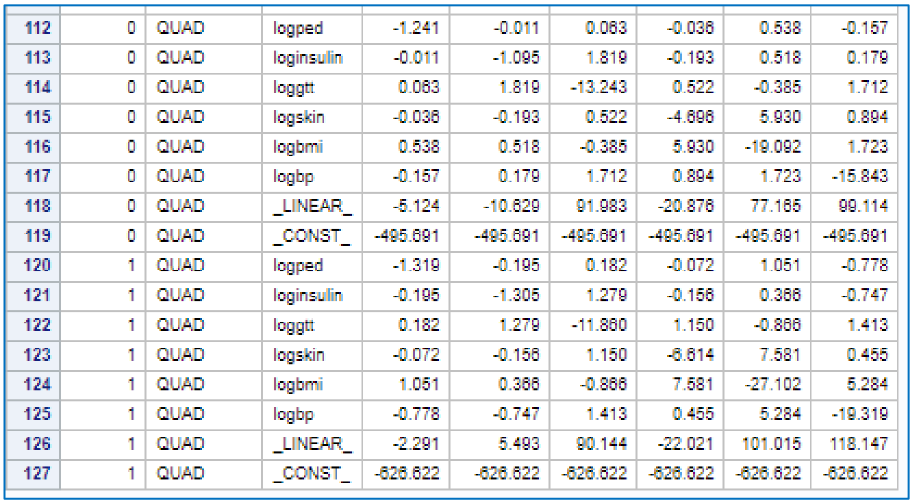
From diabetic	0	1	Total
0	7343 84.78	1318 15.22	8661 100.00
1	1752 37.33	2941 62.67	4693 100.00
Total	9095 68.11	4259 31.89	13354 100.00
Priors	0.651	0.349	

Error Count Estimates for diabetic

	0	1	Total
Rate	0.1522	0.3733	0.2294
Priors	0.6510	0.3490	

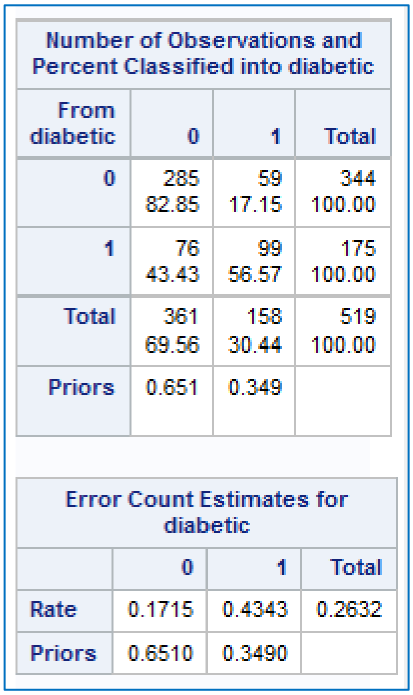
Figure 12– Training Data Model 2

We see that 82.85% (285 patients) of subjects are classified correctly as not diabetic (0) in Model 1 and 84.78% (7343 patients) of subjects are likewise correctly classified as not diabetic (0) in Model 2. On the other hand, 56.57% (99 patients) of observations are correctly classified as diabetic (1) in Model 1 and 62.67% (2941 patients) of observations are correctly classified as diabetic (1) in Model 2. While the diabetic classification may not be an overwhelming percent correct, these results are promising. The overall correct classification is also seen in figures 11 and 12. We conclude that 73.68% (100% - 26.32%) of our observations are correctly classified in Model 1 and 77.06% (100% - 22.94%) of the observations are correctly classified in Model 2. An analysis of the misclassified observations reveals that many of the missed classifications are associated with what one could potentially classify as ‘borderline,’ meaning the probabilities for classification into either category is close to equal probabilities. A table analysis of the test data classifications revealed that at approximately 400 observations are considered 'borderline' predictions, meaning the percent likelihood classification difference between not diabetic (0) and diabetic (1) fell into a predetermined category of minor (<5% difference) or moderate (<20% difference. The model coefficients of the quadratic function are seen in figure 13. Using these determined coefficients, and the promising results of the train data, the test data is presented next.



112	0	QUAD	logped	-1.241	-0.011	0.063	-0.036	0.538	-0.157
113	0	QUAD	loginsulin	-0.011	-1.095	1.819	-0.193	0.518	0.179
114	0	QUAD	loggtt	0.063	1.819	-13.243	0.522	-0.385	1.712
115	0	QUAD	logskin	-0.036	-0.193	0.522	-4.696	5.930	0.894
116	0	QUAD	logbmi	0.538	0.518	-0.385	5.930	-19.092	1.723
117	0	QUAD	logbp	-0.157	0.179	1.712	0.894	1.723	-15.843
118	0	QUAD	_LINEAR_	-5.124	-10.629	91.983	-20.876	77.165	99.114
119	0	QUAD	_CONST_	-495.691	-495.691	-495.691	-495.691	-495.691	-495.691
120	1	QUAD	logped	-1.319	-0.195	0.182	-0.072	1.051	-0.778
121	1	QUAD	loginsulin	-0.195	-1.305	1.279	-0.156	0.366	-0.747
122	1	QUAD	loggtt	0.182	1.279	-11.860	1.150	-0.866	1.413
123	1	QUAD	logskin	-0.072	-0.156	1.150	-6.614	7.581	0.455
124	1	QUAD	logbmi	1.051	0.366	-0.866	7.581	-27.102	5.284
125	1	QUAD	logbp	-0.778	-0.747	1.413	0.455	5.284	-19.319
126	1	QUAD	_LINEAR_	-2.291	5.493	90.144	-22.021	101.015	118.147
127	1	QUAD	_CONST_	-626.622	-626.622	-626.622	-626.622	-626.622	-626.622

Figure 13 – Model coefficients



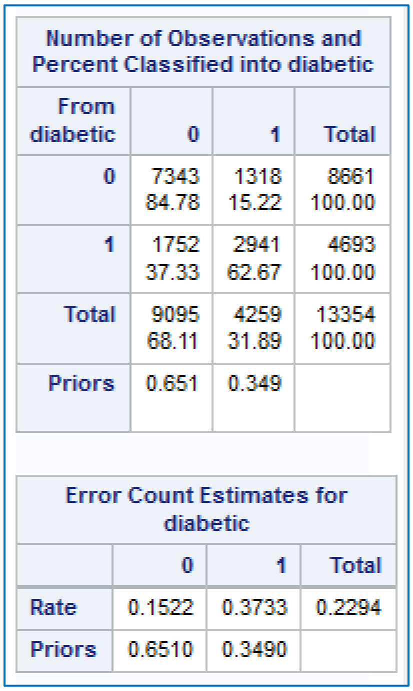
Number of Observations and Percent Classified into diabetic

From diabetic	0	1	Total
0	285 82.85	59 17.15	344 100.00
1	76 43.43	99 56.57	175 100.00
Total	361 69.56	158 30.44	519 100.00
Priors	0.651	0.349	

Error Count Estimates for diabetic

	0	1	Total
Rate	0.1715	0.4343	0.2632
Priors	0.6510	0.3490	

Figure 14 – Results for Cross Validation Set Model 1



Number of Observations and Percent Classified into diabetic

From diabetic	0	1	Total
0	7343 84.78	1318 15.22	8661 100.00
1	1752 37.33	2941 62.67	4693 100.00
Total	9095 68.11	4259 31.89	13354 100.00
Priors	0.651	0.349	

Error Count Estimates for diabetic

	0	1	Total
Rate	0.1522	0.3733	0.2294
Priors	0.6510	0.3490	

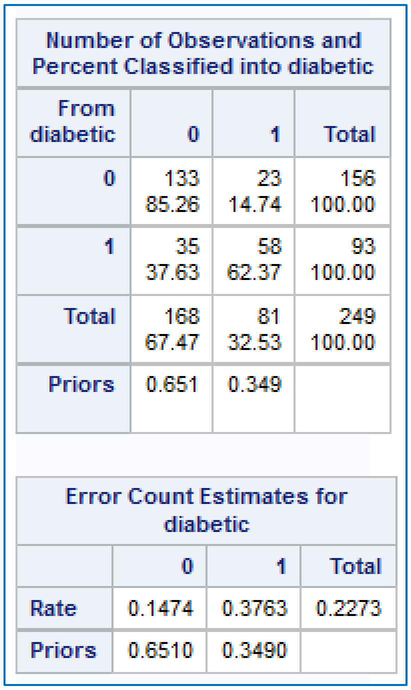
Figure 15- Results for Cross Validation Set Model 2

Before assessing the test data performance, we completed a cross validation of the training data. Cross validation was employed as a check against our training data. We sought to overcome the overfitting trap of model development. We achieved that goal first by utilizing cross validation, and then separately applying the discriminant functions to the independent test data set. The results of the cross validation can be seen in figure 14 for Model 1 and figure 15 for model2.

The results are mostly consistent with the model development train data set. The important takeaway from the cross-validation figures are twofold: first the highly consistent performance leads us to believe that the imputed values are plausible. Second, the results are like the train data set, indicating this model will likely not fall into the overfitting trap.

The confusion matrix for the test data (n=249) of Model 1 is seen in figure 16 and in figure 17 for Model 2. The model has overcome the overfitting trap; the independent sample of the test data holds promise for future studies.

We see that of the 93 subjects with diabetes, 58 were correctly classified. Likewise, of the 156 subjects without diabetes, 133 were classified correctly in Model 1. Model 2 displays a very similar pattern in classification with 84% correct classification for not diabetic and 60.93% correct classification for diabetic status. Overall, the test data results are like what we saw in the train results: 77.27 (100-22.73) % correct classification for Model 1 and 75.95 (100-24.05) % for Model 2.



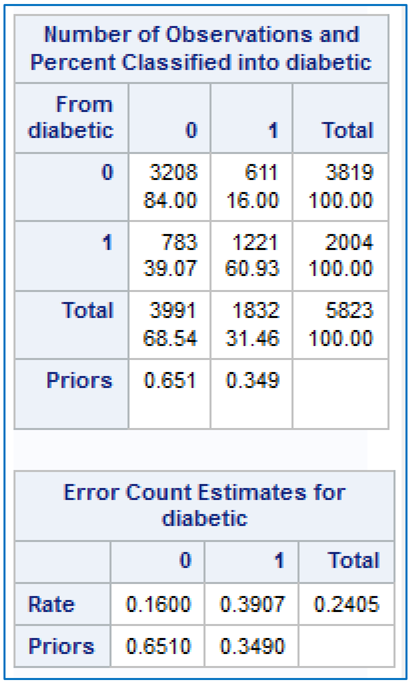
Number of Observations and Percent Classified into diabetic

From diabetic	0	1	Total
0	133 85.26	23 14.74	156 100.00
1	35 37.63	58 62.37	93 100.00
Total	168 67.47	81 32.53	249 100.00
Priors	0.651	0.349	

Error Count Estimates for diabetic

	0	1	Total
Rate	0.1474	0.3763	0.2273
Priors	0.6510	0.3490	

Figure 16- Model 1 Confusion Matrix



Number of Observations and Percent Classified into diabetic

From diabetic	0	1	Total
0	3208 84.00	611 16.00	3819 100.00
1	783 39.07	1221 60.93	2004 100.00
Total	3991 68.54	1832 31.46	5823 100.00
Priors	0.651	0.349	

Error Count Estimates for diabetic

	0	1	Total
Rate	0.1600	0.3907	0.2405
Priors	0.6510	0.3490	

Figure 17 Model 2 Confusion Matrix

Other Analysis Methods

While we sought to maximize our classification prediction power, we did seek alternatives to pure discriminant analysis. While it is not our intention to complete a deep-dive into principal component analysis, we feel a brief explanation of the high-level steps taken merits a discussion. Beginning after our imputation step on untransformed values, we applied PCA to this data set to investigate. We see from the scree/variance plot in Figure 18 that 5 principal components explain approximately 85% of the variation.

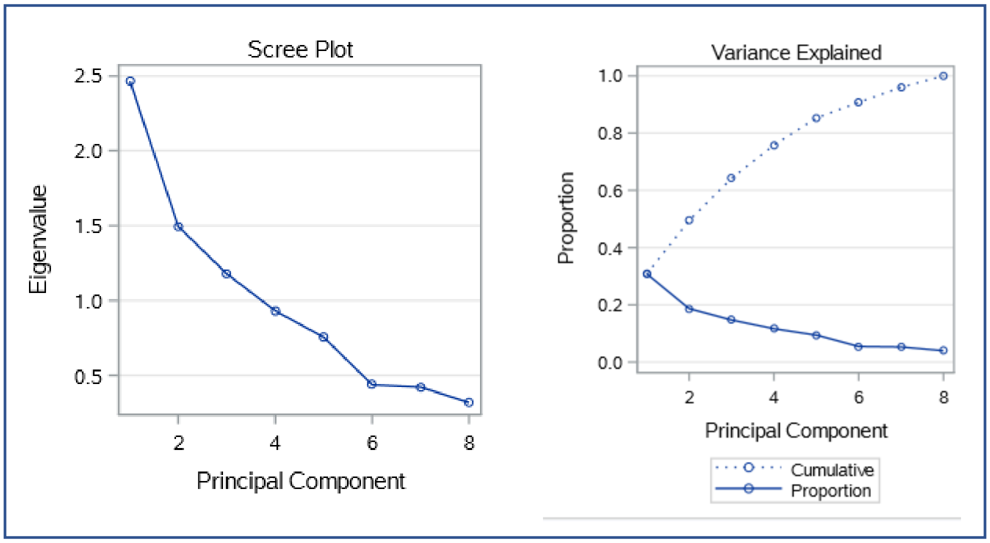


Figure 18 – PCA Scree and Variance Plots

We hypothesized those five principal components could thus be inputs to a discriminant analysis. We determined our hypothesis was false: this approach, while meeting the assumptions of discriminant analysis, did not gain us any predictive power. As such, for ease of interpretation, we chose the discriminant analysis path as our final solution.

Conclusion / Interpretability

While there are significant limitations to this data set, we found it promising we could achieve approximately 77% classification rate with this population. Given that treatment of diabetes is best begun early, any sub-segment of the population that can be identified and given interventions early on can reduce the impact and the cost of treatment of diabetes. Classifying patients using tools like this as described holds great promise for better use of limited treatment and education resources. Correctly targeting 3 out of 4 diabetics before there are clinical symptoms of the disease that are diagnosable will lead to better outcomes and lower costs. Lifestyle changes, closer monitoring, and educational offerings to patients who are classified as being at a high risk for developing diabetes can be very impactful. There are some significant limitations to our research here, however. The results we achieved should not be generalized to a larger population as there are some unique features of the Pima Indian subjects that don't apply to a larger, more ethnically diverse population. Pima Indians have some of the highest rates of diabetes in North America(6). The predictor variables used here may have less or more predictive value on other populations with less incidence of diabetes. Since this was an observational study, we make no claims of causation here, but rather point to some interesting correlations in the data. We would also note that our Model 2, with 19200 observations may slightly overstate the predictive ability of the model, due to an artificially low variance in the imputed data. This is a trade-off we chose to make rather than to reduce the size of the data set by removing those subjects who had missing observations of one or more of the components.

Additional limitations include the age of the data set, a lack of new technological testing (Hga1c) available to patients today and the high number of missing values. While our imputation was methodologically rigorous, it's preferable to have actual observations as frequently as possible.

Appendix – References and Code

References:

1. The American Diabetes Association – The Cost of Diabetes – Retrieved 07/09/2017 from http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html
2. The Centers for Disease Control – Diabetes Latest – Retrieved 07/09/2017 from https://www.cdc.gov/features/diabetesfactsheet/
3. The University of California, Irvine – Pima Indians Data Set – Retrieved 07/01/2017 from https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes
4. Ibid
5. SAS MI procedure document https://support.sas.com/documentation/onlinedoc/stat/141/mi.pdf
6. Trends in Diabetes Prevalence Among American Indian and Alaska Native Children, Adolescents, and Young Adults, Acton, Burrows, et al. Am J Public Health, 2002 September, 92(9):1485-1490 Retrieved on 07/05/2017 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1447266/

Code:

```
%web_drop_table(WORK.raw);
%web_drop_table(WORK.train);
%web_drop_table(WORK.test);
%web_drop_table(WORK.logs_train);
%web_drop_table(WORK.logs_test);
%web_drop_table(WORK.std_log_test)
;
%web_drop_table(WORK.std_log_train
);
%web_drop_table(WORK.Counts);

*read in raw data;
data raw;
infile
'/home/harisanadhya0/sasuser.v94/M
SDS 6372/Project 2/Pima.csv' DLM=','
FIRSTOBS=2;
input pregnancies gtt bp skinFold
insulin bmi pedigree age diabetic;
run;

/* Raw Data Analysis : measurement of
statistical parameters */
proc means data=raw n nmiss mean
min max stddev stderr;
run;

*Create new dataset to change 0 to
NA, keep the raw dataset untouched;
data raw_missing;
set raw;
run;

*Change 0 to missing values, required
for imputation only for vars where it
makes sense;
*from
https://stackoverflow.com/questions/
24012797/how-do-i-change-0s-to-
missing-value?rq=1;
data raw_missing;
modify raw_missing;
array vars{*} bp skinfold insulin bmi
gtt;
do i = 1 to dim(vars);
   if vars{i}=0 then call missing(vars{i});
end;
run;

/* Raw data with zero replaced by .  :
measurement of statistical parameters
*/
proc means data = raw_missing N
nmiss mean std stderr;
 var _numeric_ ;
run;
quit;

/* Retain a copy of the raw data set
having NA instead of zero */
data raw_missing_copy;
set raw_missing;
run;

/* Missing Raw Data Analysis */
proc mi data=raw_missing nimpute=0;
ods select misspattern ;
run;

/* Determine Correlation between
Raw Data Variables */
proc corr data=raw nosimple;
var pregnancies gtt bp skinFold insulin
bmi pedigree age;
run;

/* Generate Scatterplot matrix to view
distribution of the Raw Data*/
proc sgscatter data=raw;
matrix pregnancies gtt bp skinFold
insulin bmi pedigree
age/diagonal=(histogram);
run;

* Begin with raw data stats;
proc glm data = raw;
model  diabetic=pregnancies age
pedigree bmi insulin skinfold bp gtt ;
run;

*log transform data;
*Skewness in data-log transform to fix
right skewness;
data raw_missing (drop=pedigree age
insulin gtt skinFold bmi bp);
set raw_missing;
logped = log(pedigree);
logage = log(age);
loginsulin = log(insulin);
loggtt = log(gtt);
logskin = log(skinFold);
logbmi = log(bmi);
logbp = log(bp);
run;

*EDA Post log xform;
/* Scatterplot matrix for logged Raw
Data with zero changed to missing
values */
proc sgscatter data = raw_missing;
matrix logped logage loginsulin loggtt
logskin logbmi
logbp/diagonal=(histogram) ;
run;

* Begin with raw data stats;
proc glm data = raw_missing plots=all;
model  diabetic=logped loginsulin
loggtt logskin logbmi logbp ;
run;

TITLE " LISTWISE REGRESSION";
proc glm data = raw_missing;
model diabetic =  logped logage
loginsulin loggtt logskin logbmi
logbp/solution ss3;
```

```
run;
quit;

*Imputation Phase - single imputation;
*From the UCLA Paper;
*Trace plots here show good
convergence;
* Imputation Phase - single imputation
of the raw missing data (Model 1) ;
proc mi data= Raw_missing_copy
nimpute=1
out=mi_mvn_imputation_then_log
seed=54321 round=1 minimum=0;
var diabetic pedigree age insulin gtt
skinfold bmi bp;
run;

* log conversion of the imputed data
set;
data mi_mvn_imputation_then_log
(drop=pedigree age insulin gtt skinFold
bmi bp);
set mi_mvn_imputation_then_log;
logped = log(pedigree);
logage = log(age);
loginsulin = log(insulin);
loggtt = log(gtt);
logskin = log(skinFold);
logbmi = log(bmi);
logbp = log(bp);
run;

/* Scatterplot matrix for Data logged
after imputation */
proc sgscatter data =
mi_mvn_imputation_then_log;
matrix logped loginsulin loggtt logskin
logbmi logbp/diagonal=(histogram) ;
run;

*Analysis Phase - estimate model for
each data set
* Data set used is the one with 25
imputations on raw missing data
logged after the imputation step;
TITLE " MULTIPLE IMPUTATION
REGRESSION - MVN";
proc glm data =
mi_mvn_imputation_then_log ;
model diabetic =  logped loginsulin
loggtt logskin logbmi logbp;
ods output
ParameterEstimates=a_mvn;
run;
quit;

*Pooling Phase - combining parameter
estimates across datasets;
TITLE " MULTIPLE IMPUTATION
REGRESSION - MVN";
proc mianalyze parms=a_mvn;
modeleffects intercept  bmi insulin
skinfold bp gtt;
run;

*Check out the priors and see how
balanced data is;
proc freq
data=mi_mvn_imputation_then_log;
tables diabetic;
run;

*Trying to split dataset;
DATA train test;
SET mi_mvn_imputation_then_log;
Random1 = RANUNI(14380132);
IF Random1 < 0.7 THEN output train;
          ELSE output test;
Run;

/* Print the Train dataset */
proc print data=train;run;

* View the number of diabetic and
non-diabetic records;
proc freq
data=mi_mvn_imputation_then_log;
tables diabetic;
run;

/* Prediction of diabetic patients in the
test data based on the model built
using  normal data */
proc discrim data=train pool=no
testdata=test crossvalidate
method=normal anova manova wcov
pcov listerr crosslist
testout=diabetic_out
outstat=diabetic_stat;
class diabetic;
var logped loginsulin loggtt logskin
logbmi logbp;
priors "0"=.6510 "1"=.3490;
run;

* imputation of logged raw data;
proc mi data= Raw_missing nimpute=1
out=mi_mvn_logged_raw seed=54321
round=1 minimum=0;
var diabetic logped logage loginsulin
loggtt logskin logbmi logbp;
run;

/* Scatterplot matrix for Data imputed
after log transformation */
proc sgscatter data =
mi_mvn_logged_raw;
matrix logped loginsulin loggtt logskin
logbmi logbp/diagonal=(histogram) ;
run;

*Imputation Phase - 25 imputation of
the logged data;
proc mi data= Raw_missing
nimpute=25 out=mi_mvn_logged
seed=54321 round=1 minimum=0;
var diabetic logped logage loginsulin
loggtt logskin logbmi logbp;
run;

/* Scatterplot matrix for Data with 25
imputations on the log transformed
data */
proc sgscatter data = mi_mvn_logged;
matrix logped loginsulin loggtt logskin
logbmi logbp/diagonal=(histogram) ;
run;

* Imputation Phase - 25 imputation of
the raw missing data (Model 2);
proc mi data= Raw_missing_copy
nimpute=25 out=mi_mvn seed=54321
round=1 minimum=0;
var diabetic pedigree age insulin gtt
skinfold bmi bp;
run;

/* Log transformation of the imputed
data set */
data mi_mvn(drop=pedigree age
insulin gtt skinFold bmi bp);
set mi_mvn;
logped = log(pedigree);
loginsulin = log(insulin);
loggtt = log(gtt);
logskin = log(skinFold);
logbmi = log(bmi);
logbp = log(bp);
logage = log(age);
run;

/* Scatterplot matrix for Data imputed
before log transformation - 25
Imputations */
proc sgscatter data = mi_mvn;
matrix logped loginsulin loggtt logskin
logbmi logbp/diagonal=(histogram) ;
run;

*Analysis Phase - estimate model for
each data set
* Data set used is the one with 25
imputations on raw missing data
logged after the imputation step;
TITLE " MULTIPLE IMPUTATION
REGRESSION - MVN";
proc glm data = mi_mvn ;
model diabetic =  logped loginsulin
loggtt logskin logbmi logbp;
ods output
ParameterEstimates=a_mvn;
run;
quit;

*Pooling Phase - combining parameter
estimates across datasets;
TITLE " MULTIPLE IMPUTATION
REGRESSION - MVN";
proc mianalyze parms=a_mvn;
modeleffects intercept  bmi insulin
skinfold bp gtt;
run;
```

```
*Check out the priors and see how
balanced data is;
proc freq data=mi_mvn;
tables diabetic;
run;

*Trying to split dataset;
DATA train test;
SET mi_mvn;
Random1 = RANUNI(14380132);
IF Random1 < 0.7 THEN output train;
        ELSE output test;
Run;

/* Print the Train dataset */
proc print data=train;run;

* View the number of diabetic and
non-diabetic records;
proc freq data=mi_mvn;
tables diabetic;
run;

/* Prediction of diabetic patients in the
test data based on the model built
using the normal data */
proc discrim data=train pool=no
testdata=test crossvalidate
method=normal anova manova wcov
pcov listerr crosslist
testout=diabetic_out
outstat=diabetic_stat;
class diabetic;
var logped loginsulin loggtt logskin
logbmi logbp;
priors "0"=.6510 "1"=.3490;
run;

*Try PCA as an alternative;
proc mi data= Raw_missing_copy
nimpute=25 out=mi_mvn seed=54321
round=1 minimum=0;
var diabetic pedigree age insulin gtt
skinfold bmi bp;
run;

* Split the dataset as train and test (70-
30 ratio);
DATA train test;
SET mi_mvn;
Random1 = RANUNI(14380132);
IF Random1 < 0.7 THEN output train;
        ELSE output test;
Run;

* Use proc princomp to generate the
principle components;
proc princomp plots=all data=mi_mvn
out=pca;
var pregnancies gtt bp skinFold insulin
bmi pedigree age;
run;

* print the principle components;
proc print data=pca;
run;

* Remove the fields that are not
required;
data pca_2;
set pca (drop=pedigree age insulin
pregnancies gtt skinFold bmi bp);
run;

*LDA trial run on master data using
principle components;
proc discrim data=pca_2 pool=test
crossvalidate;
class diabetic;
var Prin1 Prin2 Prin3 Prin4 Prin5;
run;
```