MATH 579 FINAL REPORT
MODELING OF MEASURABLE MATERNAL ATTRIBUTES

DIANA AMADOR, ANDREW DEMPSEY, SAMANTHA GODFREY, AND PETER STEINHOFF

Abstract. The objective of this project was to determine if the placental weight/ratio is affected by various maternal variables such as maternal age, BMI, gestational diabetes, and race using various mathematical approaches. A study was performed on 1229 mothers who were of different age groups, five different races, various BMI, and approximately 4% had gestational diabetes. We analyzed each of these four variables separately to see if there was any relationship between each of the variables and the placental weight or placental ratio statistically. We also studied whether the relationships would affect the placenta by a combination of two or more of these variables using K-means method, Principal Component Analysis and Logistic Regression. The use of all these mathematical approaches provided that out of the four maternal characteristics, maternal age and BMI had effects on placental weight and placental ratio. Our findings indicate that the results were statistically significant, but overall too weak to be used for future predictions.

1. Introduction

The placenta plays a major role in maintaining a healthy pregnancy. The placenta works as a trading post between the mother's and the baby's blood supply, oxygen supply, and nutrients. Therefore, researching the placenta produces vital information for determining the development of the fetus. A possible placental factor that reflects the placental development is its weight. The weight of the placenta can be an indicator of any abnormalities that may be happening as the fetus is growing. Several factors can affect the weight of the placenta, causing the weight to be low, normal, or high. In this study, we decided to analyze the maternal characteristics that were provided to us to see if they had any significant effect on placental weight and hence, provide useful information that can be used by medical experts to help make informed decisions regarding factors influencing the placental health. The maternal characteristics that are analyzed in this study are age, BMI, gestational diabetes, and race.

There is research that allows us to see how these maternal characteristics relate to the placenta in various ways. For purposes of this study we decided to follow research directions that were very relevant to the connections between the placenta weight and placental ratio and our chosen maternal characteristics. Placental ratio is defined as the ratio of the placenta weight to the baby's birth weight. Research shows that mothers that were diagnosed with gestational diabetes at a certain gestational age and had a glucose treatment did have higher placental weight/ratios compared to those mothers that did not get the treatment [6]. Similarly, it is noted that African Americans had a growth restriction on placental weight due to the mothers' pre-pregnancy BMI and the weight gain during the term [3]. Furthermore, research allows us to see that the mother's age might also have an impact on the weight of the existing placenta. Mothers who are aged 25 and under tend to have normal to low placenta weights while mothers over 25 years tend to have high placental weights [2][7]. Finally, Body Mass Index has also been noted to have a more significant effect on placental ratio. Maternal obesity is due to an increase exposure to maternal glucocorticoids which determine the placental ratio [5]. In a research done by Perry, Beevers, and Bareford, the study noted that "lower gestational age at birth and higher maternal body mass index were the only significant independent predictors of placenta ratio...Gestational age

Date: May 20, 2011.

at delivery and body mass index were also both positively associated with placental weight and birth weight"[5].

In determining what maternal characteristics have an impact on the placenta and hence the development of the fetus, the goal is to help medical professionals make better decisions in a timely manner and allow mothers to better care for their newborn.

2. Research Methods

1229 mothers out of the total 2006 were studied for our specific research. This number of mothers was chosen by those having valid data in all of four maternal characteristics chosen for this study. Any data that had incorrect values or missing data were discarded. The maternal characteristics that were chosen had the following characteristics:

- Maternal Age: The age of the mothers was recorded in years. The ages ranged from 16 to 46 years.
- Mothers Race: The mothers in our data were grouped into five different races. These races were then coded with the values $1-5$. (White = 1, African American = 2, American Indian = 3, Asian or Pacific Islander = 4, and Other = 5).
- Maternal BMI: The mothers BMI was recorded in kilograms per meters squared ($\frac{kg}{m^2}$).
- Gestational Diabetes: The mothers gestational diabetes was recorded with two values, 0 and 1. 0 represented that the mother did not have gestational diabetes and 1 represented that the mother had gestational diabetes.

2.1. Statistical Analysis.

In our first approach in this research study, we decided to study our data using linear regression. Our goal for using linear regression was to be able to model a relationship between our maternal characteristics and the placental weight or ratio by looking specifically at three things:

(1) A line fitted to the data set
(2) The coefficient of determination, R^2
(3) The F-significance value

Finding a good fit would suggest that a trend exists in the data. The coefficient of determination determines how well future outcomes are likely to be predicted by the statistical model. It ranges from 0 to 1, where $R^2 = 1$ is the most desirable value. The F-significance value provides the probability that the variables that are being compared have a linear regression due to randomness. An F-significance value close to 0 is the most desirable. An example of a perfect line of fit is shown in Figure 1.

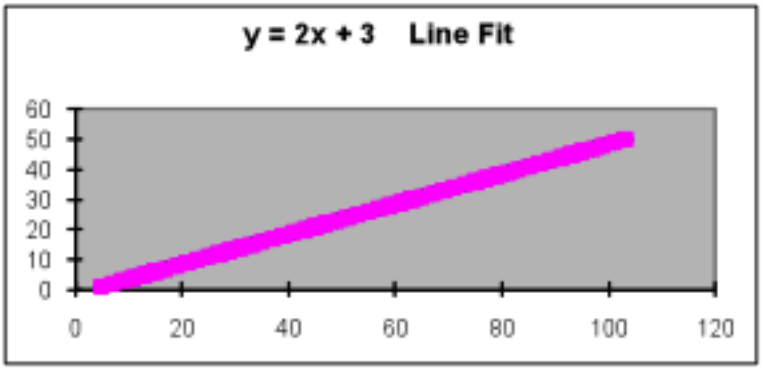

Figure 1. Example of a perfect line fit; R^2=1, F-significance=0

2.2. **Principal Component Analysis.** Principal Component Analysis (PCA) is a data analysis technique that specializes in the organization of multivariable data sets. These data sets, perhaps relating a dependent variable to several independent variables, can be difficult to visualize graphically in that they exist in high-dimensional space (i.e., more than three dimensions, with each variable representing a single dimension). With this constraint it is generally unrealistic to attempt to graph a multivariable data set composed of much more than two or three variables. The traditional method of graphing each independent variable separately with the dependent variable is always available, but correlations existing between multiple variables can be overlooked or completely missed with this method.

PCA solves this problem by translating the variables into their principal components. For each variable in the data set of interest, a single principal component is generated, with each principal component accounting for some of the variance of the original variables. However, the utility of PCA is not that each variable is now represented by yet another variable now called a principal component, but rather that the majority of variance of the original variables is now represented by just a few of the principal components, thereby reducing the amount of variables to work with and consequently making it feasible to graphically represent the original data set. In this way, PCA is sometimes referred to as a variable reduction procedure. Notice that although the variance in the original variables from the original data set is conserved in the principal components, it is not uniformly distributed through all of them, and that in fact the "first" principal component usually carries most of the variance (i.e., most of the information about the data set), the "second" principal component carries less of the variance but usually still a significant amount, and then less and less variance is carried over into the remaining principal components.

Generally, the first few principal components are enough to study the original data set and the remaining principal components are not used. The loss of information by dismissing the remaining principal components is negligible, and indeed if it were not, one could always include one more principal component if information loss was a worry. In this way, with "most" of the information from the original variables being conserved, PCA presents itself as a powerful data analysis tool for simplifying multivariable data sets.

When using PCA, however, one should be vigilant of the types of data being used. For example, suppose a data set contains temperature, time of day, time of year, and altitude. Running PCA through the data set with the temperature in Fahrenheit, and then running it again with the temperature in Celsius will *not* necessarily give the same results. Furthermore, because the data has not been normalized, there is a possibility that the principal components will not be independent causing the results of the PCA to become skewed. By normalizing the data set, the independence of the principal components is made sure, and this guarantees that no information will be represented more than once in more than one principal component. The general procedure for Principal Component Analysis is as follows:

(1) Create data set
(2) Normalize data set forming the normalized matrix
(3) Find the covariance matrix of the normalized matrix
(4) Find eigenvalues and eigenvectors of the covariance matrix
(5) Find the significant principal components by finding the eigenvectors with the highest eigenvalues
(6) Form a "feature" matrix by using the chosen eigenvectors as columns, and put in order from highest corresponding eigenvalue to lowest corresponding eigenvalue
(7) final result matrix=(feature matrix)^T(normalized matrix)^T

The final result matrix is the original data but only in respect to the eigenvectors we chose. Again, though not all of the information from the original data set is conserved, the important majority of it is carried over into the final result matrix leaving the user with a matrix closely

representing the original data and much easier to utilize due to variable reduction. This reduction of variables is the base usefulness of PCA and should be considered a potential tool whenever one is studying a multivariable data set or any data set that is difficult to represent or visualize. It is this reduction of variables which allows PCA to potentially illuminate unnoticed structures or trends within a data set.

One final ability of PCA is having reduced the dimensionality of a data set through use of its principal components, one can, with greater ease and less computation, now project this final result matrix onto a lower-dimensional space that can be visualized and represented graphically. This projection is a representation of the original data set and allows for ease of study of what originally may have been a difficult set to examine. The reduction of variables coupled with this lower-dimensional projection technique is what makes PCA both useful and popular in data analysis.

2.3. **K-means Clustering.** K-means clustering is a well-known data analysis technique that attempts to separate a set of observations into k clusters. It is an iterative process that hones in on a set of centroids that minimizes the cumulative centroid-to-observation distance. Our goal was to isolate a particular maternal characteristic that generated a well-defined partition within a larger mash of maternal and placental characteristics. Over the course of our analysis, we assigned various values to k. In each *case* considered, we designed k to match the numerical range of the "ground-truth" characteristic under examination. For example, when we looked at ethnicity we used five clusters to correspond to the five distinct ethnicities found in the data set. For other cases, we assigned the ground-truth to a different characteristic, say maternal age, or BMI. Furthermore, each ground-truth variable could itself be segmented several different ways. For example, maternal age could be separated into five-year blocks, or merely broken into two groups separated by the median age. While there is an infinite number of "ground-truth" divisions possible, we chose arrangements that mirrored our other studies.

Once we had the ground-truth characteristic pinned down, we choose a limited number of other, independent characteristics for the spatial dimensions. For each observation, we created a vector with the dimension of the vector corresponding to the number of spatial characteristics employed. The spatial characteristics were kept separate from the ground-truth characteristic, which was used only for coloring. For instance, in Figure 2 the x-axis is maternal age, the y-axis maternal BMI, and the z-axis placenta ratio. The ground-truth variable, ethnicity, is used to designate the scatter points' shape and color. While the example in Figure 2 contains three spatial dimensions, we can easily create vectors that exist in any number of higher dimensional spaces. For instance, in the case labeled *Age 1.0*, the details of which can be found in Table 1, we employed five spatial dimensions. *Age 1.0* proceeds to use maternal age for the ground-truth clustering agent, separated into seven age blocks. Clearly, we cannot visualize this data, but k-means clustering works just as well in hyperspace. In this case the seven cluster centroids would themselves be five dimensional vectors.

Once we have our spatial dimensions defined and our ground-truth pinned down, we proceed with k-means clustering. Matlab's prepackaged algorithm generates a set of centroids. These centroids mark the best-fit cluster center for each of the k clusters. For consistency, it is best to start the iterative process from the same distribution of points. Another output of Matlab's k-means algorithm pairs each spatial vector with its closest cluster. Note that the default measure of closeness is based on the Euclidean distance between the vector and each centroid in turn. The observation's vector is then nominally assigned to the cluster with the smallest distance to centroid.

Within the vector space defined by the spatial characteristics, we have k clusters, located by centroid, and the nominal cluster membership of each observation. Next we wish to give each of the k clusters a unique "identity" based on the same range of values found in the ground-truth. We developed a method to derive this identity using the vector population of each

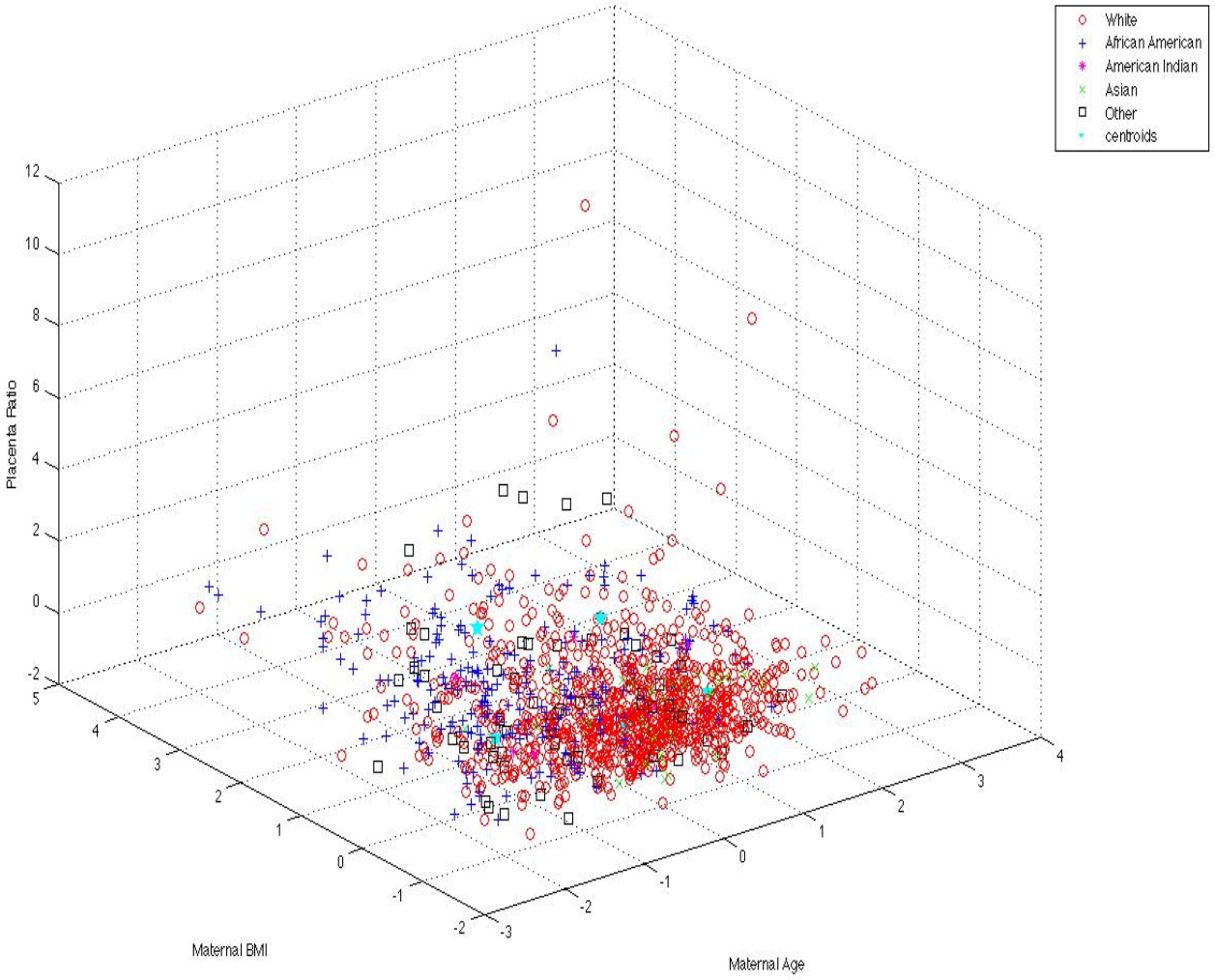

FIGURE 2. Three spatial characteristics, with ethnicity for ground truth

cluster. Specifically, the unique identity of each cluster is based on the ground-truth coloring of its members. For instance, if ethnicity is the ground-truth and the highest percentage of whites are found in cluster A, then cluster A is now identified as the "white cluster". However, a conflict may arise if the highest percentage of another ethnicity, say African Americans, is also found in cluster A. This conflict occurs often since the observation population need not be uniformly distributed among clusters. In a case like this, the first cluster takes the identity of the highest percentage of ground-truth distribution overall. That particular cluster and color is removed from consideration and the next highest distribution is found. In this way each cluster is assigned a unique identity.

The real test of k-means clustering is to examine the clustering algorithm's goodness-of-fit relative to our ground-truth selection. To review, for each spatial vector we have a ground-truth value and nominal cluster membership. Furthermore, each cluster has been assigned its own identity based on the same ground-truth scale. We wish to compare the ground-truth of each observation to the identity of its cluster of nominal membership. If the spatial characteristics in question indeed possess strong separation relative to the ground-truth coloring, it will be indicated through highly homogeneous clusters. For example, in the cases where we use ethnicity

case name	*Age 1.0*	*Age 1.5*
spatial dimensions	5	5
spatial characteristics	maternal ethnicity, maternal BMI, gestational diabetes, placenta weight, birthweight, maternal age	maternal ethnicity, maternal BMI, gestational diabetes, placenta weight, birthweight
ground truth	maternal age	maternal age
number of clusters k	7	7
g-of-f random assignment	17%	17%
g-of-f k-means R^5 (Euclidean)	28%	28%
g-of-f k-means R^5 (Mahalanobis)	27%	27%
Principal Components	3	2
g-of-f k-means PC-subspace (Euclidean)	22%	25%
g-of-f k-means PC-subspace (Mahalanobis)	21%	26%

TABLE 1. Cases with maternal age for ground-truth

as the ground-truth, we want to examine the membership of the identified white cluster. Within this cluster, we check how much of the membership is actually white. We do the same for the other four ethnicity clusters. We count a positive match if an observation's ground-truth identity matches the overall identity of its cluster, regardless of the value itself. We total the number of positive matches and divide by the overall population to get a goodness-of-fit statistic for the k-means clustering attempt. We can compare this goodness-of-fit to a situation where cluster assignments are dealt randomly to the vectors. The idea being, if the selected data set clusters nicely, most observations will match the identity of their nominal cluster. Consequently, the positive match count will far exceed random matching. On the other hand, if the spatial data is chaotic relative to the ground-truth, a cluster will more likely contain a heterogeneous mix of ground-truth values. While each cluster still takes on an identity derived from a plurality of its membership, it will also contain a large number of members with foreign ground-truth, thus decreasing overall goodness-of-fit.

There is one more wrinkle to k-means clustering, namely, the distance metric used. The distance between the individual observations and the cluster centroids, within the vector space, is relative to the method of measurement. Conceivably, an observation can change what cluster it is "closest" to, depending on the way distance is measured. By default, our k-means clustering method uses Euclidean distance, however there are other metrics to consider. The Mahalanobis distance is defined by $D_m(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}$ where x is the vector in question, μ is the centroid of the cluster in question and S is the covariance matrix of the same cluster. The motivation for using Mahalanobis over Euclidean is that Mahalanobis distance compensates for the non-spherical distribution of the points within a cluster. When data, such as ours, is derived from various maternal and placental characteristics, the different scaling employed creates distortion. A cluster of points that might ordinarily take a spherical shape will be more ellipsoidal. By incorporating the covariant matrix of the individual clusters, the Mahalanobis distance adjusts for differences in scaling among the spatial coordinates. Keeping this in mind, we tested the goodness-of-fit of our k-means clustering under both the Euclidean metric and the Mahalanobis. For each observation, we compared the ground-truth of the vector to the identity of the cluster of least distance.

2.4. **Logistic Regression.** Logistic regression uses a binomial distributed set of data to predict a success/fail outcome based on several variable combinations. It is a useful way of describing the relationship between one or more independent variables and a binary response variable expressed as a probability that has only two values.

Let $R = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}$

then $\ln(R) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$

where R=the odds ratio that a certain outcome will happen

x_i=the independent variables; $i = 1, 2, ..., p$

β_i=the logistic regression coefficients; $i = 1, 2, ..., p$

The odds ratio is the ratio of the probability that a certain event will happen to the probability that it will not happen.

$R = \frac{P}{1-P}$

where P=the probability that a certain outcome will happen

$R(1 - P) = P$

$R = P(1 + R)$

$P = \frac{R}{1+R}$

Let $z = \ln(R) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$

$P = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$

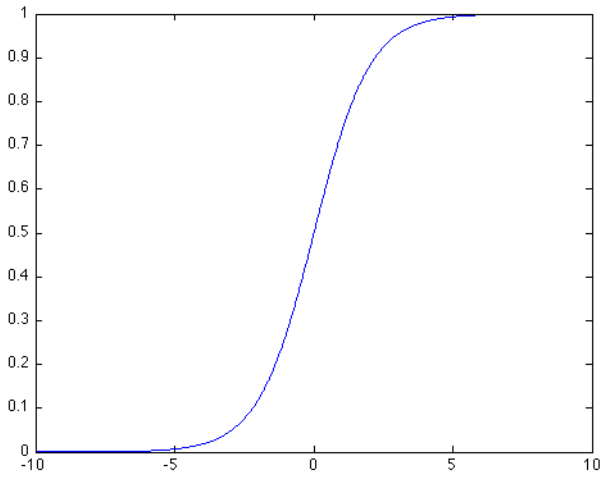

FIGURE 3. Plot of the logistic function

The logistic function $f(z) = \frac{1}{1+e^{-z}}$ is used to find the probability that a certain combination of variables will lead to a "success" situation [4]. The plot of the logistic function in Figure 3 shows an S-shaped curve that lies between 0 and 1, representing the probablity at a certain value of z. The goal is to determine the logistic regression coefficients from the collected data set. Once this is completed, one can determine the probability that a certain event will happen with different combinations of independent risk variables.

3. RESULTS

3.1. **Statistical Analysis.**

3.1.1. *Maternal Age.* Within our data set, maternal age is normal distributed (see Figure 4), as is placenta weight and placenta ratio. Figure 5 shows a best-fit least-squares linear regression between maternal age and placenta weight, which produces the coefficient of determination

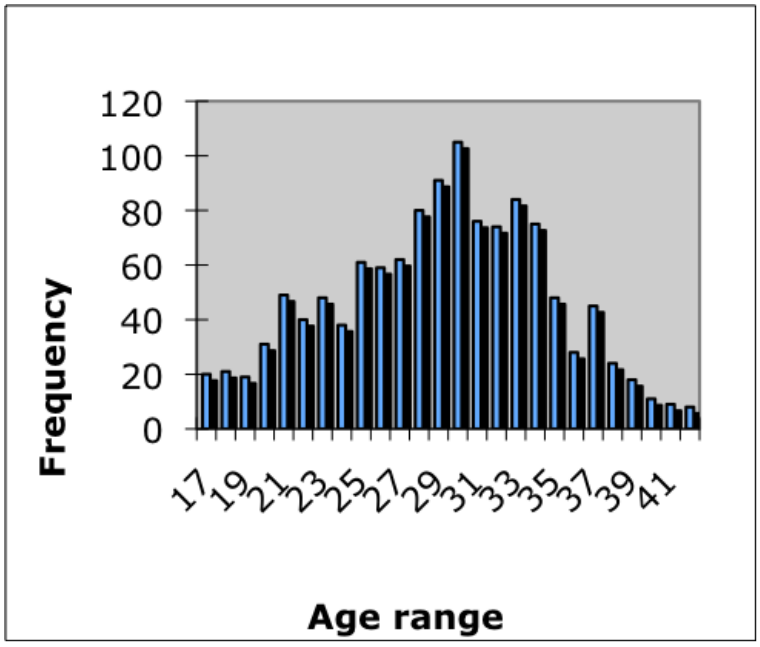

FIGURE 4. Distribution of maternal age in our data set

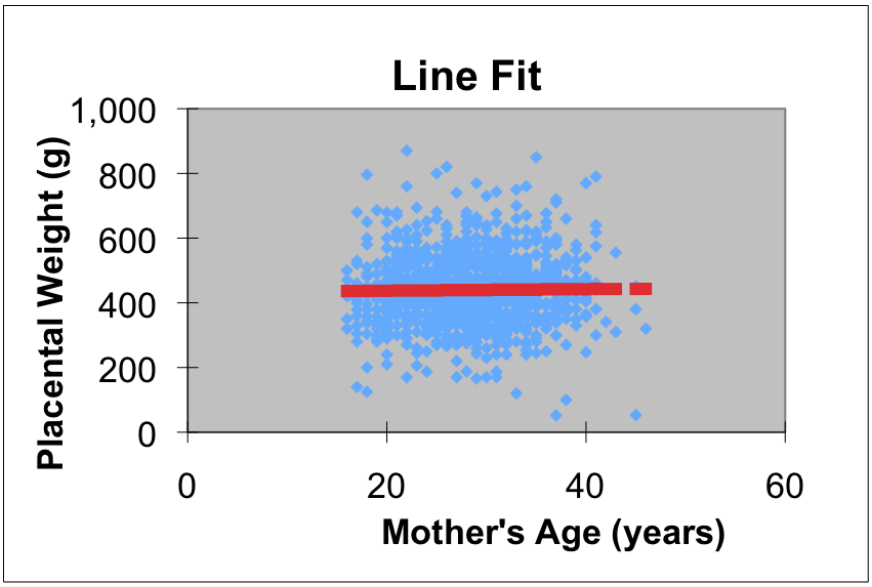

FIGURE 5. Maternal age to placenta weight line fit plot

$R^2 = 0.01$, with the associated F-stat P-value of 0.6287. This is not significant or meaningful other than to confirm that placenta weight is independent of maternal age.

However, if we compare maternal age to placenta *ratio* and run a best-fit least-squares linear regression, we find a slightly better coefficient of determination in $R^2 = 0.01$. The associated F-stat P-value is 0.0005. These results are shown in Figure 6. While this still small R^2 value indicates the difficulty in making accurate predictions of placenta ratio based on maternal age, the low P-value indicates that maternal age does indeed influence the placenta ratio. Couple this to the previously stated fact that maternal age does not influence placenta weight, and we have indirect evidence that maternal age affects birth weight. Stated another way, while maternal age may affect the weight of the offspring, those influences bypass the determination of placenta weight.

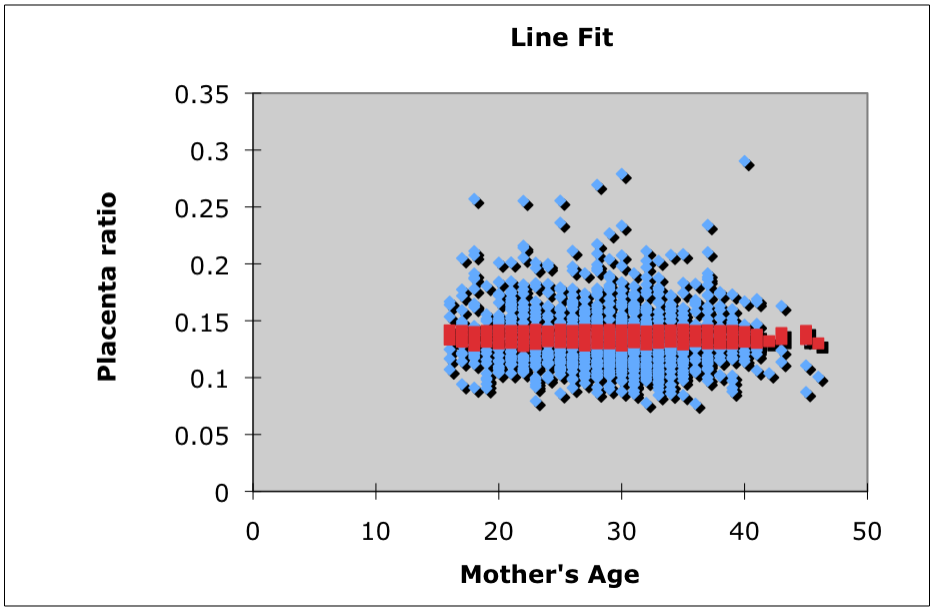

FIGURE 6. Maternal age to placental ratio line fit plot

3.1.2. *Maternal BMI.* To study the effect of the mother's BMI on the placenta, the statistical technique of linear regression was used. We found the placental ratio and the β measure of placental efficiency to be a better measure of the placenta, and specifically a better measure of the placenta-baby-mother relationship. The β value is similar to the placental ratio, but is instead defined as

$$\beta = \frac{\log(\text{placenta weight})}{\log(\text{birth weight})}$$

Figure 7 shows a best-fit least-squares linear regression between BMI and the placental ratio. For this fit, $R^2 = .020921$ and the F-stat P-value $= 3.54E-07$ suggesting a positive linear correlation between the two variables. This result is not surprising and is corroborated by recent studies [5]. A second linear regression between BMI and the β value bears similar results with $R^2 = .020158$ and the F-stat P-value $= 5.8E-07$. This result enforces the positive linear correlation between maternal BMI and the placenta-baby relationship in that both the placental ratio and the β value are measures of the ratio between the weight of the baby and the weight of the placenta. Learning of and confirming this relationship caused us to ask, in what other ways are these two variables related?

A second statistical investigation was conducted to further explore the relationship between maternal BMI and the placental ratio. For our data set, the average BMI is 25.76 and the average placenta ratio is .137; Table 2 lists the amount of individuals over or under these averages, the amount of individuals that are over or under both averages, and the amount of individuals that are mixed. 233 individuals are over both the BMI average and the placenta ratio average; where as 481 individuals are under both the BMI average and the placenta ratio average. This result indicates that the tendency to be under both averages is greater than the tendency to be over both averages. Moreover, only 216 individuals were over the BMI average but under the placenta ratio average, suggesting that an individual under the BMI average is 2.22 times more likely to be under the placenta ratio average than an individual over the BMI average.

299 individuals are under the BMI average but over the placenta ratio, suggesting that an individual under the BMI average is actually 1.28 times more likely to be over the placenta ratio than an individual who is over the BMI average. These results are less significant (and less accurate) when one considers that although the average BMI for our data set was 25.76,

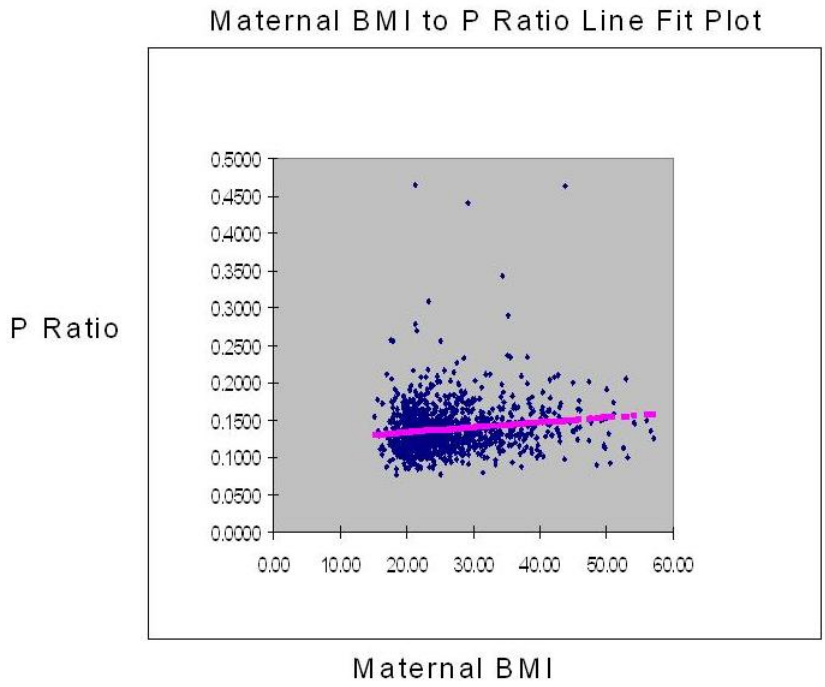

FIGURE 7. Maternal BMI to placental ratio line fit plot

the majority of the mothers in the set measured closer to a BMI of 22. This discrepancy is due to our data set containing 327 mothers over a BMI of 28, 254 mothers over a BMI of 30, and even 139 mothers over a BMI of 35. The high frequency of high BMI's makes for more individuals being under the "average" BMI in that the average BMI is elevated above the norm which is generally considered to be somewhere between 20-25 for a healthy women. Regardless of the high BMI's, a definite trend exists with significantly more individuals being under both averages than over both averages.

> BMI Average	> P Ratio Average	Individuals over both averages	> BMI average and < P Ratio Average
449	532	233	216
< BMI Average	< P Ratio Average	Individuals under both averages	< BMI average and > P Ratio Average
780	697	481	299

TABLE 2. Amount of individuals over or under the BMI and placenta ratio averages

3.1.3. *Mother's Ethnicity.* To study the effect of the mother's ethnicity on the placenta, the average placenta weight and placental ratio were calculated for each ethnicity. The data was split up by ethnicity and the mean and variance of each set was calculated for the placenta weight and placental ratio. Figure 8 shows the highest average placenta weight was with white mothers (449.49 ± 101.94) and the lowest average placenta weight was with African American mothers (418.81 ± 115.79). The highest average placental ratio was with African American mothers ($.1434 \pm .0331$), shown in Figure 9. This shows that African American babies in this study had the lowest average birth weight compared to their placenta weight. However, the average placental ratio only ranged from 13.3% to 14.4% among the ethnicities, so the placental ratio remained somewhat constant regardless of ethnicity. The fact that the African American ethnicity had the lowest average placental weight is supported by many past studies [1][3].

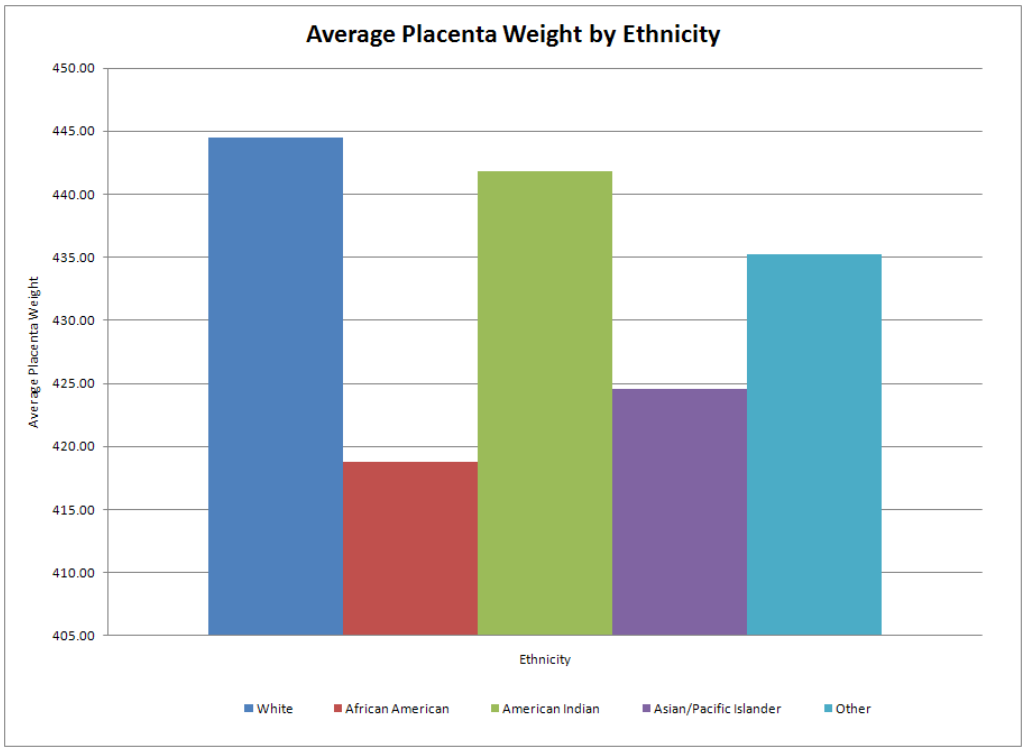

FIGURE 8. Average placenta weight by ethnicity

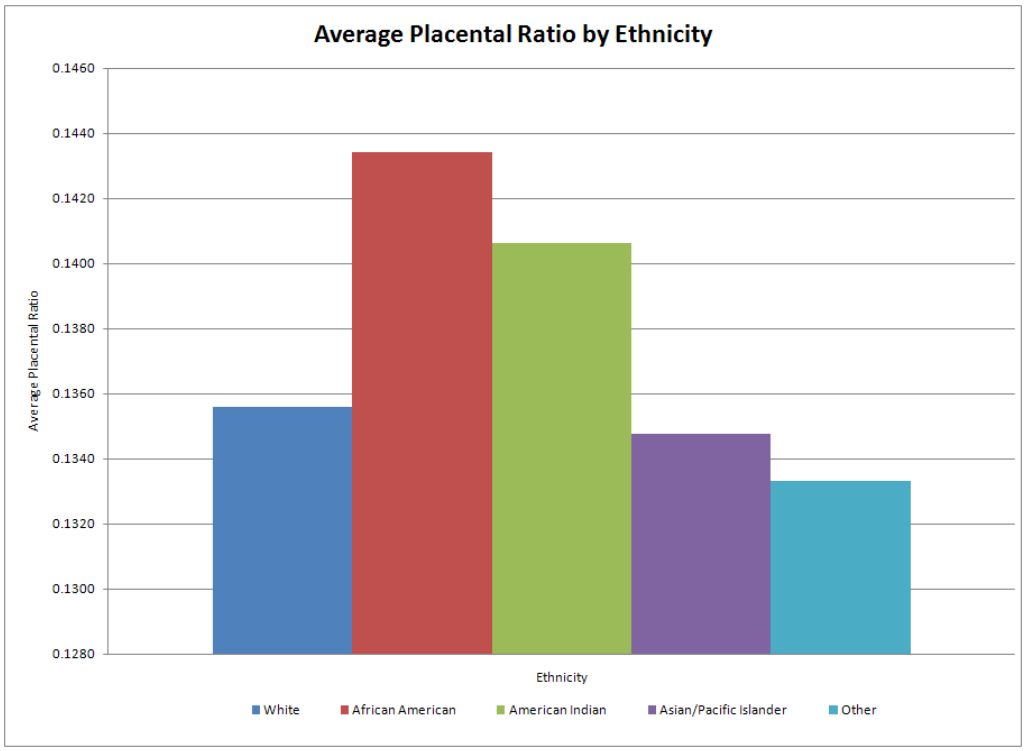

FIGURE 9. Average placenta ratio by ethnicity

3.1.4. *Gestational Diabetes.* In the gestational diabetes group, the mothers were classified with two values (0 = no gestational diabetes and 1 = gestational diabetes). Using linear regression to analyze if gestational diabetes affected placental weight and ratio provided us with poor results. Looking at the non-gestational diabetes mothers and comparing the maternal characteristics of

age and BMI of this particular group with placental weight only provided us with the results shown in Figures 10 and 11. When analyzing this data using linear regression, we decided to focus on the R^2 values and F-significance values. Hence, from Figures 10 and 11, only Figure 10 provides a slightly better F-significance value of 0.000278, indicating that BMI does influence placental weight. However, the R^2 value of 0.01 made it difficult to make any accurate predictions of placental weight based on BMI.

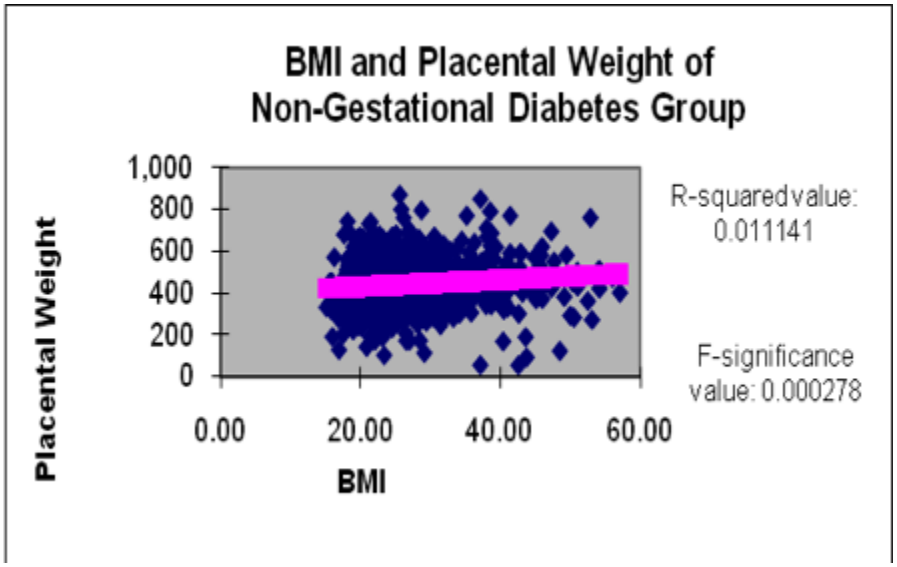

FIGURE 10. BMI and placental weight for non-gestational diabetes group

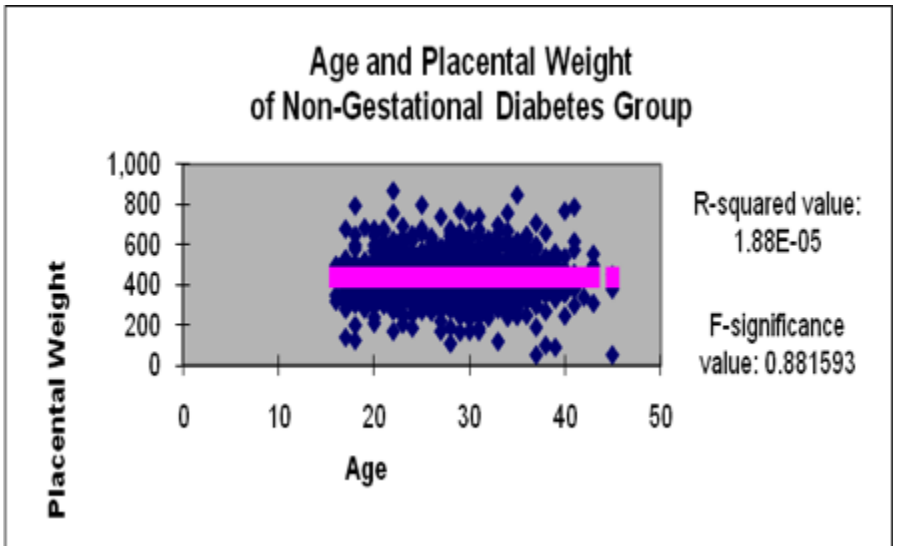

FIGURE 11. Age and placental weight for non-gestational diabetes group

We then decided to analyze the gestational diabetes group in search of better results. In doing so, we discovered that these results were worse than those obtained in the non-gestational diabetes group. Figures 12 and 13 show that placental weight is independent of gestational diabetes. The same analysis was done for both groups, but looking at the placental ratio, and the results were not meaningful. The results only provided that the placental ratio was independent of gestational diabetes.

Due to the lack of inconclusive results that were obtained using linear regression analysis for this particular group, we decided to analyze the gestational diabetes group using the Z-test. The

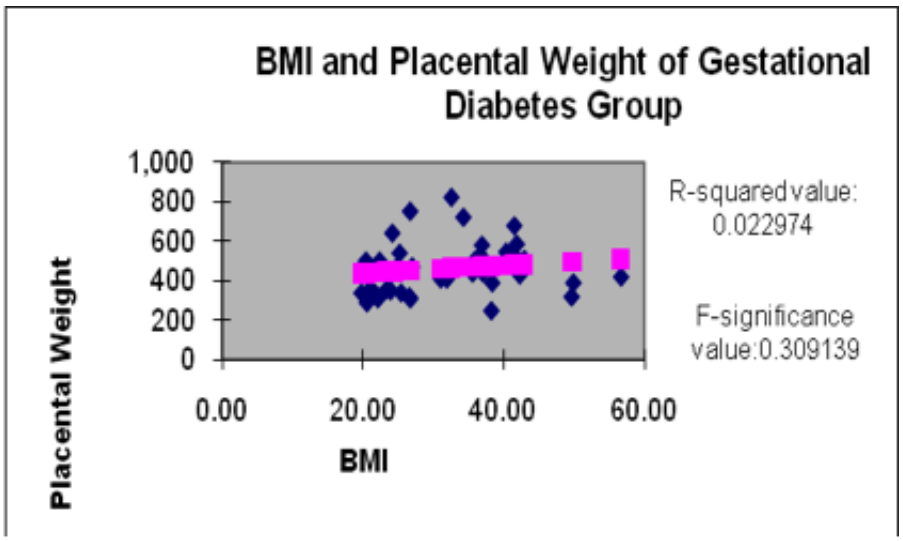

FIGURE 12. BMI and placental weight for gestational diabetes group

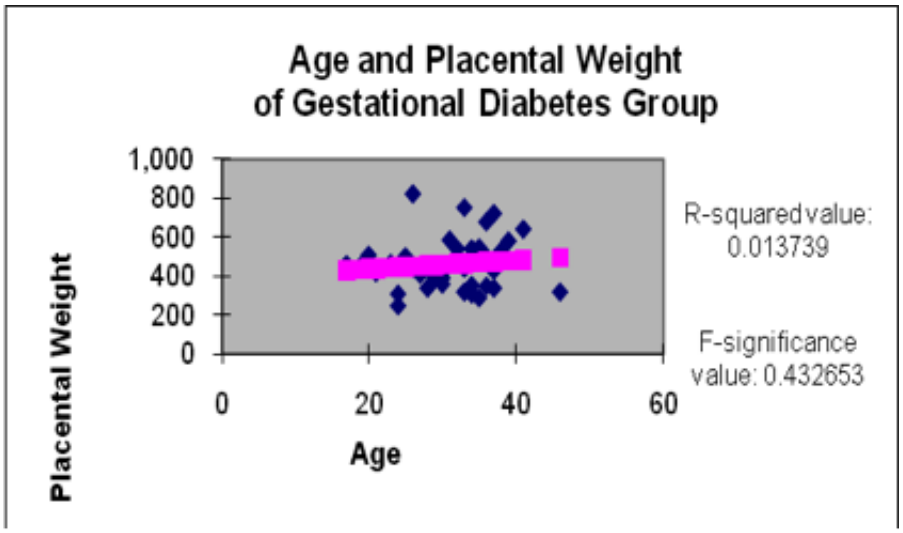

FIGURE 13. Age and placental weight for gestational diabetes group

Z-test is a test that compares sample and population means to determine if there is a significant difference. The z-score is a measure of how far away a measurement is from the mean, measured in standard deviations. The formula that is used to calculate this score is $z = \frac{x-\bar{x}}{s}$, where x is a measurable value, $\bar{x}$ is the mean of all measured values and s is the standard deviation of the measured values.

From Table 3, we can see that the average placental ratio and weight of the gestational diabetes group is higher than the mean of all the population, yet this information does not allow us to make any conclusions on how gestational diabetes affects placental weight and placental ratio. Our goal was to have a converted z-score of ± 2 because this is statistically significant. By looking at the z-scores of the placental ratio and weight, 0.33 and 1.22 are the standard deviations from the population mean, respectively. These results do not show us a significant difference in the gestational diabetes groups mean from the means population. The results simply indicate that there is a 37% probability that mothers will have a higher average than the populations placental ratio and a 11% probability that mothers will have a higher average than the populations placental weight.

	Placental Ratio	Placental Weight
Population Average	.1369	438.7063
Population Standard Deviation	.03196	104.0824
Total Cases	1229	1229
Average Ratio for Non-Diabetics	.1368	437.8503
Standard Deviation	.03205	103.34995
Total Non-Diabetic Cases	1182	1182
Average Ratio for Diabetics	.1384	460.2340
Standard Deviation	.02993	120.3212
Total Diabetic Cases	47	47
Standard Error (of Diabetes)	.004366	17.550646
Z-Score	.33695	1.226609
Converted Z-Score	.37	.11

TABLE 3. Z-test analysis for placental ratio and placental weight in relation to gestational diabetes

3.2. **K-means clustering.** By any measure, our k-means clustering produced poor numerical results. In most cases, the clustering goodness-of-fit was only a few percentage points better than random paring. This indicates minimal formation of homogeneous clusters. Examine case *Ethnicity 2.5*, found in Table 4. In this case, three-dimensional spatial vectors were formed. The characteristics used were maternal age, maternal BMI, and placenta ratio. Ethnicity was assigned to be the ground-truth. In the associated plot, seen in Figure 2 above, the three spatial characteristics position the observation while ethnicity colors and shapes it. In *Ethnicity 2.5* we divided the ground-truth into five values corresponding to the five ethnicities reported in the data set. The k-means algorithm found the best-fit centroids for five distinct clusters, represented by the large magenta stars in the scatter plot of Figure 14. Our cluster-identity mechanism assigned a ground-truth value to each of the five clusters. Each observation vector was compared to that of its nearest cluster, under both the Euclidean and Mahalanobis metric. Cycling through all 1229 observations, positive identity matches were recorded for later comparison. As seen in Table 4, the results for *Ethnicity 2.5* are rather weak, as they are for the other ethnic cases as well. In *Ethnicity 2.5* specifically, our method pairs vectors with their correct cluster only 29% of the time. If we dealt out cluster membership at random we would have achieved a 20% match rate. Note however, that by switching to the Mahalanobis metric, our success rate increases to 31%.

As displayed in the other tables, comparable results occur even when different characteristic combinations are employed. For example, in case *Age 1.5*, found in Table 1, the success rates were 28% and 27% for Euclidean and Mahalanobis distances respectively. Compare this to the 17% that would have occurred via random cluster assignment. For a last example, look at *Birthweight 1.0*, found in Table 5. Here we use five spatial dimensions and color the ground-truth by birth weight. As a change of pace, we only employed two ground-truth values; 0 for below the mean birth weight, and 1 for above the mean. This produced two cluster identities and evenly distributed the population between the two. An individual vector has a 50/50 chance of being assigned to its proper cluster. While the high number of spatial characteristics requires that the vectors exist only in hyperspace, we are still aiming for a simple two cluster separation. Unfortunately this does not happen. The goodness-of-fit tops out at 55%, indicating that there is minimal clustering relative to birth weight. The distribution of the two data types

case name	*Ethnicity 1.0*	*Ethnicity 1.5*	*Ethnicity 2.0*	*Ethnicity 2.5*
spatial dimensions n	2	2	3	3
spatial characteristics	maternal BMI, placenta ratio	maternal BMI, placenta ratio	maternal age, maternal BMI, placenta ratio	maternal age, maternal BMI, placenta ratio
ground truth	ethnicity	ethnicity	ethnicity	ethnicity
number of clusters k	5	5	5	5
g-of-f random assignment	20%	20%	20%	20%
g-of-f k-means R^n (Euclidean)	9%	9%	29%	29%
g-of-f k-means R^n (Mahalanobis)	9%	9%	31%	31%
Principal Components	2	1	3	2
g-of-f k-means PC-subspace (Euclidean)	9%	31%	29%	26%
g-of-f k means PC-subspace (Mahalanobis)	9%	31%	31%	26%

TABLE 4. Cases with ethnicity for ground-truth

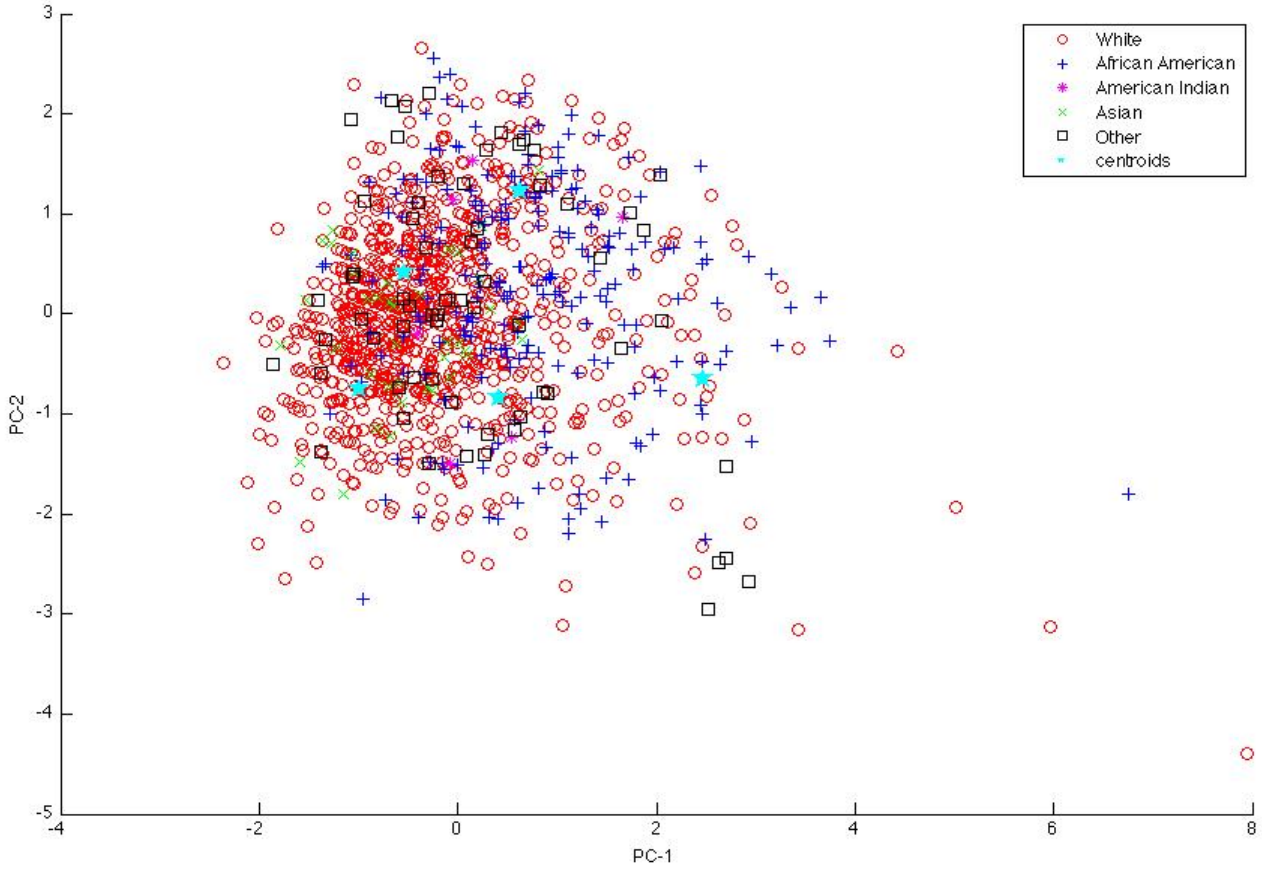

FIGURE 14. The first two principal components, with ethnicity for ground truth

is exceedingly well mixed. Conclusively, the ground-truth characteristic, namely birth weight, is almost completely independent of the five characteristics that go into the spatial dimensions.

case name	*Birthweight 1.0*	*Birthweight 1.5*
spatial dimensions	5	5
spatial characteristics	maternal ethnicity, maternal age, maternal BMI, gestational diabletes, placenta weight	maternal ethnicity, maternal age, maternal BMI, gestational diabetes, placenta weight
ground truth	birth weight	birth weight
number of clusters k	2	2
g-of-f random assignment	50%	50%
g-of-f k-means R^5 (Euclidean)	55%	55%
g-of-f k-means R^5 (Mahalanobis)	55%	55%
Principal Components	3	2
g-of-f k-means PC-subspace (Euclidean)	54%	48%
g-of-f k means PC-subspace (Mahalanobis)	55%	48%

TABLE 5. Cases with birth weight for ground-truth

3.3. **Principal Component Analysis.** In analyzing a data set along its principal components we hope to reduce clutter and clarify trends. PCA is a tool to use within a larger analytical method, which in our case is k-means clustering. For now, we forego the major weaknesses pertaining to k-means clustering detailed above and focus solely on principal component analysis. In doing so, we see that the goodness-of-fit for our method often increased by a few percentage points after a principal component projection. Other times though, the goodness-of-fit remained the same. Our efforts show that while PCA sometimes improved the overall picture, better results were not guaranteed.

Examine cases *Ethnicity 1.0* and *Ethnicity 1.5* found it Table 4 above. In *Ethnicity 1.0* the number of principal components equals the number of spatial variables. Clearly, no improvement in clustering is expected, since, in this case, PCA merely reorients the data along a different set of axes. There is no reduction in dimension and the goodness-of-fit remains at 9% across the board. In case *Ethnicity 1.5* however, we do make a reduction in dimension. The two-dimensional spatial vectors are projected onto a single axis defined by the first principal component. We observe a significant jump in clustering success, from the original 9% to 31%, but this is a rare occurrence. If we examine the neighboring case *Ethnicity 2.5*, we see that by reducing the dimensions from three spatial characteristics to the first two principal components we actual decrease the goodness-of-fit.

In the cases detailed in Tables 1 and 5, we venture into higher dimensions. Cases *Age 1.0* and *Age 1.5* of Table 1 project the same five dimensional vectors into a three-dimensional and two-dimensional subspace respectively. Unfortunately, in both cases the projections reduced the overall goodness-of-fit. Further note that the first three principal components gave slightly worse results then the first two components. Equally dismal results are demonstrated in cases *Birthweight 1.0* and *Birthweight 1.5* of Table 5. In both cases, five spatial variables are projected into subspaces defined by the first three, and first two, principal components. In *Birthweight 1.0* the projection into three-dimensional space results is no improvement, while in *Birthweight 1.5* the projection into two-dimensional space actually worsens the fit. Clearly, using principal component analysis on maternal and placental data returns mixed results. While improvements may occur in clustering, they are not consistent or substantial.

Speaking strictly in terms of visual inspection, there is only modest improvement. Examine Figures 2 and 14 again. These plots pertain to case *Ethnicity 2.5* where we project three

spatial characteristics of Figure 2 into a two-dimensional subspace defined by the first two principal components. The reduced plot, Figure 14 above does make the centroids stand out more, but the degree of improvement is rather subjective. Perhaps with less chaotic data, a spatial reduction along principal components could produce more pronounced benefits. In truth, Principal Component Analysis is still a powerful and useful tool, just not with our current data configuration.

3.4. **Logistic Regression.** The logistic function $f(z) = \frac{1}{1+e^{-z}}$ is used to find the probability that a certain combination of variables will lead to a "success" situation. z is defined as $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ... + \beta_p x_p$. $\beta_0, \beta_1, ..., \beta_p$ are the regression coefficients and $x_1, x_2, ..., x_p$ are the risk factors. The regression coefficients are based on n trials of observed data.

The Matlab built-in function, glmfit, finds the coefficients in multivariate binomial applications. The first trial to study was the binomial situation referenced to the β-value.
A "success" situation is when $\beta < .75$ and a "fail" situation is when $\beta \geq .75$ since the placenta is considered more efficient when $\beta < .75$. With this information and our set of placenta data, the regression coefficients found with the glmfit function are

$\beta_0 = -.739364, \beta_1 = .050368, \beta_2 = -.017877, \beta_3 = -.066001, \beta_4 = .048526$

With these values, we can plot the logistic function based on our data set. Figure 15 shows the logistic regression plot determined by the data set in this study. Notice that the plot only ranges from .35 to .85. This is because our data set only creates a short domain in the middle of the logistic function plot. If we extend our data set out for a larger range of values outside of our data set, we can create a plot with the same regression coefficients. Figure 16 shows the extended curve for values outside our data set. Appendix A shows how to determine the regression coefficients and plot the resulting logistic function.

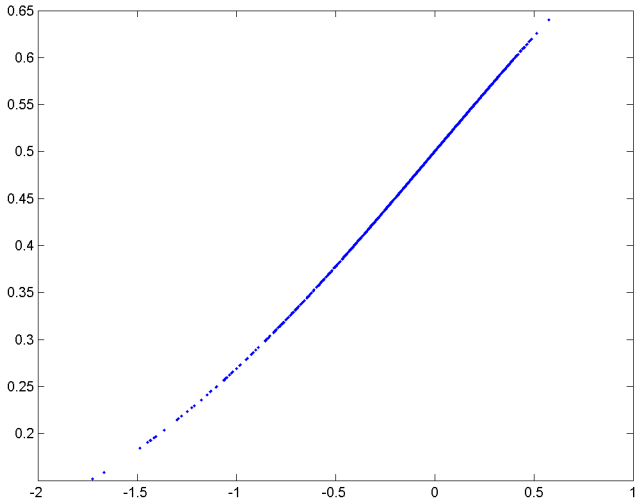

FIGURE 15. Logistic regression for β value based on our data set

Another application of logistic regression studied the binomial situation for low birth weight or not. Based on the "success" as low birth weight, the regression coefficients found with the glmfit function are

$\beta_0 = -2.118044, \beta_1 = .0150295, \beta_2 = -.0196460, \beta_3 = .3892553, \beta_4 = .0120513$

From this, we can see that the third risk factor, gestational diabetes, contributes the most to the probability of the child having low birth weight while the other factors are somewhat insignificant. Figure 17 shows the plot of the logistic curve based on our data set.

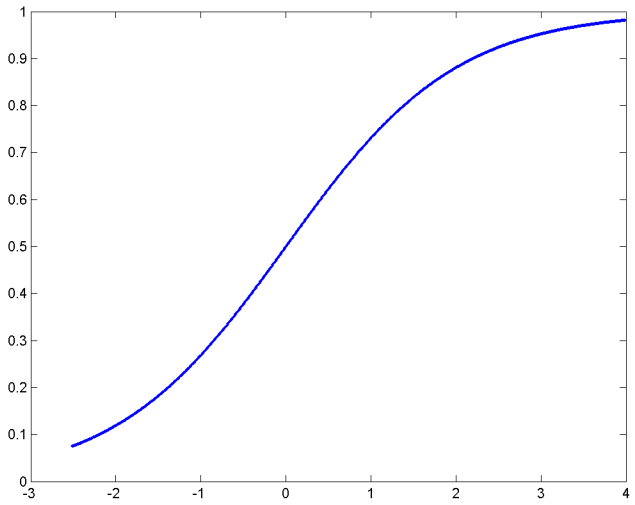

FIGURE 16. Extended logistic regression curve for β value

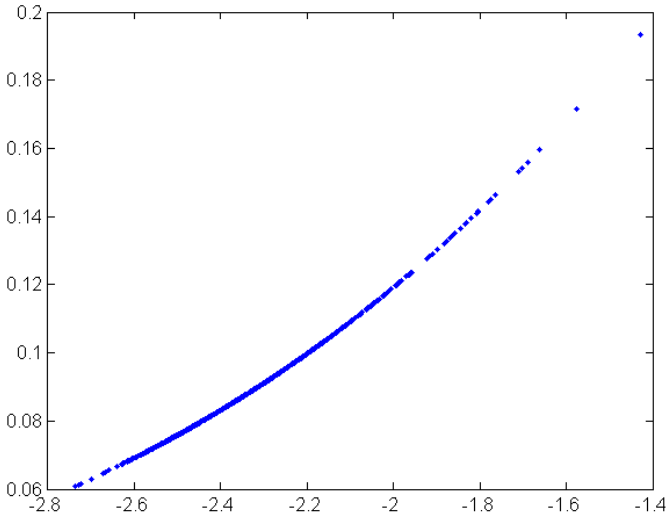

FIGURE 17. Logistic regression for low birth weight based on our data set

Since logistic regression is a method based on probabilities and not something like curve fitting, there is not a practical way of measuring error. The error is simply in the probability aspect. Thus, there is no way of determining the error in this method. Logistic regression may be helpful to determine whether a fetus will have a certain condition based on maternal characteristics and data collected in previous studies. It has a wide application to be used in many situations and thus, can be a useful tool in early detection of pregnancy risks.

4. Conclusions

In this study, we searched for direct connections between placenta weight, or ratio, and maternal characteristics, or combinations of characteristics. We used several independent mathematical approaches, namely statistical analysis, k-means clustering, principal component analysis, and logistical regression. While we were able to deduce several meaningful insights about maternal factors that influence the placenta, our methods were far from satisfactory. The results were often statistically significant, but too weak to be used for prediction. We continue to hold fast to the belief that maternal traits strongly influence placenta features. However, little to no estimation of placenta weight can be deduced from the following characteristics: maternal ethnicity, age, BMI, or diabetic status.

5. Future Work

We found the examination of maternal influences on placenta characteristics to be interesting, challenging work. There were plenty of undeveloped ideas worthy of further exploration. For instance, a covariance matrix between maternal characteristics and principal components might pinpoint the most influential principal components. Perhaps the data under consideration could be expanded to include additional maternal and placental characteristics. Non-linear regression could also be applied for possible improvements. It is clear that the data set offers a variety of information for those with the energy and patience to run down the leads.

Acknowledgments

Thank you to Dr. Carolyn Salafia for all of the placenta data and information provided for this class. Thank you to Dr. Jen-Mei Chang for providing academic assistence and guidelines to structure our project. Thank you to Amy Mulgrew who is writing her thesis, *Mathematical Features of Placenta Images Investigated with Techniques of Pattern Recognition*, and allowed her working data to be used in aid of the project.

Appendix A. Logistic Regression

```
function [B P]=logistic_regression(X,Y,x)
% Computes the regression coefficients in the equation z=b0+b1x1+b2x2+...
% +bpxp for the logistic regression equation f(z)=1/(1+exp(-z)) and plots
% the results based on the data set
% B : regression coefficients b0,b1,b2,...,bp
% X : n by p matrix representing n trials of p risk factors
% Y : binomial response of desired response. 0=fail, 1=success
% Optional input x : value at which to calculate the probability of a
% success, 1 by p vector which represents the values of the p risk factors
% P : probability, P, determined by input x
% Written by Samantha Godfrey 4/14/11
B=glmfit(X,Y,'binomial','link','logit');
n=size(X,2);
z=B(1);
for j=1:n
    z=z+B(j+1)*X(:,j);
end
Z=1./(1+exp(-z));
plot(z,Z,'.')
if nargin==3
    P=glmval(B,x,'logit');
end
```

Appendix B. The Matlab script for K-means and PCA

```
clear
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% the following four variables are user defined
%load in the workspace bucket containing the prepared data set
load % enter pre-processed workspace here

% set display desire to one to show the scatter plot, zero to suppress.
display_desire = 1;
max_colors_to_plot = 7;

% set the maximum number of dimensions for our PCA subspace
set_max_PCA_dim = 3;
%set the maximum number of clusters we want
set_max_number_of_clusters = 6;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Split the atrributes matrix from the color logic column
[r,c] = size(data_set);
X = (data_set(:,1:c-1))';
color_logic_index = data_set(:,c);

% sets the dimension of our PCA subspace
if set_max_PCA_dim <= c-1
    PCA_dim = set_max_PCA_dim;
else
    PCA_dim = c-1;
```

```
end

% finalize the number of clusters we will have.
number_of_colors = max(color_logic_index)-min(color_logic_index)+1;
if number_of_colors <= set_max_number_of_clusters
    number_of_clusters = number_of_colors;
else
    number_of_clusters = set_max_number_of_clusters;
end

% finalize the number of color logic values get plotted.
if max_colors_to_plot >= number_of_colors
    colors_to_plot = number_of_colors;
else
    colors_to_plot = max_colors_to_plot;
end

% we will run through the loop two times. The first time working with our
% as is attribute matrix, the second time projecting the atrributes to a
% lower dimension using PCA
for cordinate_method = 1:2
    if cordinate_method == 1
        Y = X;
    else
        % This is the heart of PCA, Y is the vector attribues in the new subspace
        [W, sigma, V] = svd(X);
        W_L = W(:,1:PCA_dim);
        Y = (W_L')*X;
    end

    %we pick the first few scatter points to be the start of the kmeans iteration
    centroid_starter = Y(:,1:number_of_colors)';

    % This is Matlab's k-mean function
    [cluster_index,centroids] = kmeans(Y',number_of_clusters,'start',centroid_starter);

    %here we determine how many of each color are in each cluster
    cluster_vs_color = zeros(number_of_clusters,number_of_colors);
    for i = 1:number_of_clusters
        for j = 1:number_of_colors
            counter = 0;
            for k = 1:1229
                if cluster_index(k) == i
                    if color_logic_index(k)==j
                        counter = counter +1;
                    end
                end
                cluster_vs_color(i,j) = counter;
            end
        end
    end
    color_count_sums = sum(cluster_vs_color,1);

    % we determine how much of each color is in each cluster as a percentage
```

```
cluster_v_color_percent = zeros(number_of_clusters,number_of_colors);
for i = 1:number_of_clusters
    for j = 1:number_of_colors
        cluster_v_color_percent(i,j) = cluster_vs_color(i,j)/color_count_sums(j);
    end
end

%we determine the dominant color for each cluster
permission_matrix = ones(number_of_clusters,number_of_colors);
winner_matrix = zeros(number_of_clusters,number_of_colors);
cluster_identity_color = zeros(number_of_clusters,1);
for h = 1:number_of_clusters
    held_item = 0;
    max_row = 1;
    max_col = 1;
    for i = 1:number_of_clusters
        for j = 1:number_of_colors
            if permission_matrix(i,j) == 1
                if cluster_v_color_percent(i,j) > held_item
                    held_item = cluster_v_color_percent(i,j);
                    max_row = i;
                    max_col = j;
                end
            end
        end
    end
    winner_matrix(max_row,max_col) = held_item;
    cluster_identity_color(max_row) = max_col;
    permission_matrix(max_row,:) =0;
    permission_matrix(:,max_col) = 0;
end

% we determine the goodness of fit of the cluster dominant color
fit_positive_count_1 = 0;
fit_negative_count_1 = 0;
kmeans_cluster_membership = 0;
for k = 1:1229
    kmeans_cluster_membership = cluster_index(k);
    if color_logic_index(k) == cluster_identity_color(kmeans_cluster_membership)
        fit_positive_count_1 = fit_positive_count_1 + 1;
    else
        fit_negative_count_1 = fit_negative_count_1 + 1;
    end
end

% we turn the goodness of fit into a percentage
if cordinate_method ==1
    straight_fit_positive_percent_1 = fit_positive_count_1/1229
else
    PCA_fit_positive_percent_1 = fit_positive_count_1/1229
end

% we divide up the cluster for later processing
for h = 1:number_of_clusters
```

```
    clear cluster_subsets
    counter = 1;
    for k = 1:1229
        if cluster_index(k) == h
            cluster_subsets(:,counter) = Y(:,k);
            counter = counter +1;
        end
    end
    array_of_cluster_subsets{h} = cluster_subsets;
end

% We determine the Mahalanobis distance between each observation and
% each cluster centroid
Mahal_distances = zeros(number_of_clusters,1229);
for h = 1:number_of_clusters
    cluster_subsets = array_of_cluster_subsets{h};
    mu = mean(cluster_subsets,2);
    [r,c] = size(cluster_subsets);
    subset_minus_mean = zeros(size(cluster_subsets));
    for j = 1:c
        subset_minus_mean(:,j) = cluster_subsets(:,j) - mu;
    end
    cluster_sigma2 = subset_minus_mean*subset_minus_mean';
    cluster_inverse_sigma2 = pinv(cluster_sigma2);
    for j = 1:1229
        Dm = sqrt((Y(:,j)-mu)'*cluster_inverse_sigma2*(Y(:,j)-mu));
        Mahal_distances(h,j) = Dm;
    end
end

% for each observation we determine what centroid is closest under the
% Mahalanobis metric
closest_cluster_Mahal = zeros(1,1229);
for j = 1:1229
    held_item = 100;
    for i = 1:number_of_clusters
        if Mahal_distances(i,j) < held_item
            held_item = Mahal_distances(i,j);
            closest_cluster_Mahal(j) = i;
        end
    end
end

% we now conduct a goodness of fit count under the Mahalanobis distance
fit_positive_count_2 = 0;
fit_negative_count_2 = 0;
closest_cluster_membership = 0;
for k = 1:1229
    closest_cluster_membership = closest_cluster_Mahal(k);
    if color_logic_index(k) == cluster_identity_color(closest_cluster_membership)
        fit_positive_count_2 = fit_positive_count_2 + 1;
    else
        fit_negative_count_2 = fit_negative_count_2 + 1;
    end
```

```
end

% we convert the goodness of fit count to a percentage
if cordinate_method ==1
    straight_fit_positive_percent_2 = fit_positive_count_2/1229
else
    PCA_fit_positive_percent_2 = fit_positive_count_2/1229
    vs_random_positive_odds = 1/number_of_colors
end

% now we proceed to plot the scatter plots
if display_desire ==1
    % we do so by first splitting the observations into subsets arranged by
    % logic value
    if min(color_logic_index) ==0
        for h = 0:(number_of_clusters-1)
            clear logic_subset
            counter = 1;
            for k = 1:1229
                if color_logic_index(k) == h
                    logic_subset(:,counter) = Y(:,k);
                    counter = counter +1;
                end
            end
            array_of_logic_subsets{h+1} = logic_subset;
        end
    end

    if min(color_logic_index)==1
        for h = 1:number_of_clusters
            clear logic_subset
            counter = 1;
            for k = 1:1229
                if color_logic_index(k) == h
                    logic_subset(:,counter) = Y(:,k);
                    counter = counter +1;
                end
            end
            array_of_logic_subsets{h} = logic_subset;
        end
    end

    if min(color_logic_index)==2
        for h = 1:number_of_clusters
            clear logic_subset
            logic_subset = zeros(size(Y,1),1);
            counter = 1;
            for k = 1:1229
                if color_logic_index(k) == (h)
                    logic_subset(:,counter) = Y(:,k);
                    counter = counter +1;
                end
            end
            array_of_logic_subsets{h} = logic_subset;
```

```
            end
        end

        %now that we have the observations broken up by logic "ground
        %truth" value we proceed to plot them one at a time with a
        %different color and marker shape.
        color_options = {'r','b','m','g','k','y'};
        marker_options = {'o','+','*','x','s','^'};
        marker_size = 50;
        centroid_size = 200;

        % if the PCA dimension is 2 or less we only get two plots, one for
        % euclidean distance, one for Mahalanobis distance
        if PCA_dim <=2
            figure(cordinate_method)
            for h = 1:colors_to_plot
                clear logic_subset
                logic_subset = array_of_logic_subsets{h};
                scatter(logic_subset(1,:),logic_subset(2,:),marker_size,color_options{h},marker_
                hold on
            end
            scatter(centroids(:,1),centroids(:,2),centroid_size,'pc','filled')
            hold off
        else
            % if we more then 2 principal components we create four plots.
            % Two are three dimensional, two are 2 dimensional, for the
            % Euclidean distanc and Mahalanobis distance respectively
            figure(cordinate_method*2 -1)
            for h = 1:colors_to_plot
                clear logic_subset
                logic_subset = array_of_logic_subsets{h};
                scatter3(logic_subset(1,:),logic_subset(2,:),logic_subset(3,:),marker_size,color
                hold on
            end
            scatter3(centroids(:,1),centroids(:,2),centroids(:,3),centroid_size,'pc','filled')
            hold off

            figure(cordinate_method*2)
            for h = 1:colors_to_plot
                clear logic_subset
                logic_subset = array_of_logic_subsets{h};
                scatter(logic_subset(1,:),logic_subset(2,:),marker_size,color_options{h},marker_
                hold on
            end
            scatter(centroids(:,1),centroids(:,2),centroid_size,'pc','filled')
            hold off
        end
    end
end
```

References

[1] Greg R. Alexander, Michael Kogan, Deren Bader, Wally Carlo, Marilee Allen, and Joanne More. Us birth weight/gestational age-specific neonatal mortality: 1995-1997 rates for whites, hispanics, and blacks. *Pediatrics*, 111(1):61–66, 2003.

[2] M. Asgharnia, N.Esmailpour, M. Poorghorban, and Z. Atrkar-Roshan. Placental weight and its association wiht maternal and neonatal characteristics. *Acta Medica Iranica*, 46(6):467–472, 2008.

[3] Kesha Baptiste-Roberts, Carolyn M. Salafia, Wanda K. Nicholson, Anne Duggan, Nae-Yuh Wang, and Frederick L. Brancati. Maternal risk factors for abnormal placental growth: The national collaborative perinatal project. *British Medical Journal Pregnancy and Childbirth*, 8(44), 2008.

[4] James R. Frederick. Logistic regression. http://www.uncp.edu/home/frederick/DSC510/LogisticReg.htm, March 2003.

[5] Ivan J Perry, D.G. Beevers, P.H. Whincup, and D Bareford. Predictors of ratio of placental weight to fetal weight in multiethnic community. *BMJ*, 310(6977):436, 1995.

[6] C.P Lee T.T Lao and W.M Wong. Placental weight to birthweight ratio is increased in mild gestational glucose intolerance. *Placenta*, 18:227–230, 1997.

[7] D Warburton and A F Naylor. The effect of parity on placental weight and birth weight: an immunological phenomenon? a report of the collaborative study of cerebral palsy. *The American Journal of Human Genetics*, 23(1):41–54, 1971.

Department of Mathematics and Statistics, California State University, Long Beach, 1250 Bellflower Boulevard, Long Beach, CA 90840-1001

E-mail address: dianavama86@yahoo.com

E-mail address: adempc@gmail.com

E-mail address: Samantha.Godfrey@student.csulb.edu

E-mail address: Steinhoff111@yahoo.com