Using Real Figures to Invest in Real Estate: A Multivariate Statistical Analysis of the US Housing Market

Angela Buck - Saint Mary's University, Winona, MN 55987
Takisha Harrison - Alabama A&M, Normal, AL 35762
Linden Johnson - Albany State, Albany, GA 30344
Holly Sontag - Butler University, Indianapolis, IN 46385

July 2008

Abstract

For many Americans, investing in property is a quick and easy way to make money. The real estate game has become a very popular phenomenon, even for those without a millionaire's wallet. However, due to recent struggles in the economy, for some, investing has become more of a burden than a success story. Using multivariate statistical analysis techniques such as Principal Component Analysis, Factor Analysis, and Discriminant Analysis, we determine the factors having the most effect on housing markets. We also discover which of the 50 US states' housing markets are likely to provide a stable or risky investment for those wishing to dabble in real estate.

> *"A seemingly minor dislocation originating in the housing sector, such as a higher rate of foreclosures, might cascade through the rest of the economy in unforeseen ways, for example, in a collapse in bank earnings or a hiccup in the huge market for securities that back residential mortgages."*
>
> *Jeffrey Knight*
> *Chief Investment Officer of Global Asset Allocation at Putnam Investments*
> *Business Week October 30, 2006*

Introduction

In recent years, the US housing market has shown an incredible increase in the number of foreclosures. Many suspect this is due to a decline in the economy. According to the RealtyTrac report, one in every 483 U.S.households received a foreclosure filing in May of 2008, the highest number since RealtyTrac started the report in 2005. This year nationwide, over 260,000 homeowners received at least one foreclosure-related filing in May alone. The report also states that foreclosure filings increased from a year earlier in all but 10 states. The highest statewide foreclosure rates were found in Nevada, California, Arizona, Florida and Michigan (Zibel, Associated Press).

In this paper, we statistically analyze data on all fifty states to deduce the best states in which to invest in property. To do this, we consider thirteen variables that could affect the housing market. To analyze the data set, we use three different multivariate statistical techniques to discover underlying patterns. The first method, Principal Component Analysis (PCA), is be applied to the original data to create a reduced data set with fewer variables. This smaller data set is easier to work with in further analysis. Then we apply Factor Analysis (FA) to the original data to find underlying factors affecting the variables. Finally, we apply Discriminant Analysis (DA), a classification method for observations, to develop a model to predict whether a state's housing market should be considered a stable or risky investment. Our thirteen variables are listed in Table 1.

Table 1:List of Variables

Unemployment Rate
Population Denisty per Square Mile
Property Crime per 100,000 people
Foreclosure Rate per 1,000 people
Average Mortgage Rate on 30 year Fixed Loans
Average Credit Score
Percentage Increase in Foreclosures
Homeowners Insurance
Percentage of Population Below the Poverty Level
Percentage Increase in Value of Home
Population Growth 2000-2006
Median House Price
Percentage Increase in Bachelor's Degrees 2000-2004

I. Normality

Assessing Normality Our first concern is to test the normality of our data. This is done because both Factor Analysis and Discriminant Analysis assume that the data comes from a multivariate normal distribution. The data on a given variable is said to have come from a normal distribution if it has a probability distribution function (pdf) shaped like a bell curve. One way of assessing the normality of our data is by the Quantile - Quantile plot, commonly called the Q-Q plot. The Q-Q plot gives the relationship between observed variables and theoretical normal quantiles. It plots the sample quantile versus the quantile we would expect to observe if the observations were actually normally distributed. If the plotted pairs display a positive linear relation, the data is said to be normally distributed. However, since the Q-Q plot is a visual assessment and sometimes data sets do not always give a positive linear relation, we apply a statistical test for linearity as well. In theory, normality requires the observed variables to have a linear relationship with the theoretical quantiles, so for a more formal investigation, we check to determine if the correlation coefficient between the two is close to 1.

We calculate the sample correlation coefficient r, given by the equation

$$r = \frac{\sum_{j=1}^{n}(X_{(j)}-\bar{X})(Z_j-\bar{Z})}{\sqrt{\sum_{j=1}^{n}(X_{(j)}-\bar{X})^2}\sqrt{\sum_{j=1}^{n}(Z_j-\bar{Z})^2}} = \frac{C\hat{o}v(Z,X)}{\sqrt{V\hat{a}r(Z)V\hat{a}r(X)}}$$

This formula uses a ranked list for the values of variable X, with $X_{(1)}$, $X_{(2)}$,..., $X_{(n)}$ being in ascending order. $Z_{(1)}$ represents the first quantile of the standard normal distribution. Thus $P(Z \leq Z_j) = (j - 0.05)/n$, where n represents the sample size. $\bar{X}$ in this equation represents the sample mean of $X_{(i)}$'s and $\bar{Z}$ is the sample mean of the $Z_{(j)}$'s.

To signify a strong linear relationship we should obtain a value for r that is close to 1. In hypothesis testing, the null hypothesis H_0 is $\rho = 1$, and the alternative hypothesis H_1 is $0 < \rho < 1$, where ρ is the population correlation coefficient between X and Z. The null hypothesis is accepted if $r \geq c$, which implies normality. The sample size and desired α-level of significant determines the critical value c.

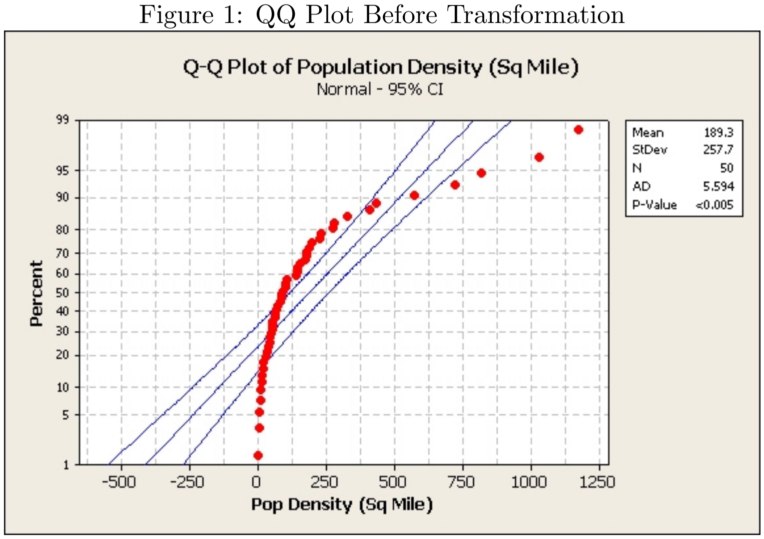
Transformations

For our particular data set we have a sample size of $n = 50$ states, 13 variables, a significance level of $\alpha = 0.05$, and a critical value of $c = .9768$ for the correlation coefficient of each variable. If the $r - value$ is above our fixed critical value, then we accept H_0 (hypothesis of normality) for that variable.

Unfortunately for 7 out of our 13 variables, this test rejected H_0. In order to normalize these rejected variables we use transformations such as a square-root, cube-root, or natural log transformation. For example, for the variable **Population Growth 2000-2006**, we use the square-root of the values in order to get an r-value that accepts H_0, and for the variable **Population Density** we take the natural log in order to get an acceptable r-value.

The QQ plots in Figure 1 and Figure 2 demonstrate the use of a natural log transformation on the variable **Population Density**. The first QQ plot produced (Figure 1) is not linear by observation, which indicates a non-normal distribution. However, in order to support that conclusion, we find the r-value. For the original data the r-value=.891 which is less than the critical value c=.9768. Both of these indicate a non-normal distribution. Hence, we apply a natural log transformation. Figure 2 shows the data after transformation. Since it appears linear, we once again find the $r - value$ to see if the variable comes from a normal distribution. For the transformed data the r-value=.987 $> c = .9768$. Therefore, we can conclude that this data comes from a normal distribution.

Figure 1: QQ Plot Before Transformation



However for variables such as **Percent Increase in Foreclosures**, there are negative values for certain states, so taking the square-root or natural log of these values will result in imaginary numbers. Therefore, to transform these variables, we add to or "shift" the values to get a positive value first, and then take the square-root of the new shifted value. For the variable **Population Growth 2000-2006**, we do not obtain an r-value greater than our critical value c=.9768. Instead we have a value of r=.9750. Even though the r-value of this variable rejects, it is very close to our critical value, and we observe the scree plot which is also an indicator of normality. The graph is relatively linear allowing us to accept H_0 for the variable. The r-values, and transformations (if needed) for all 13 variables are displayed in Table 2.

Figure 2: QQ Plot After Transformation

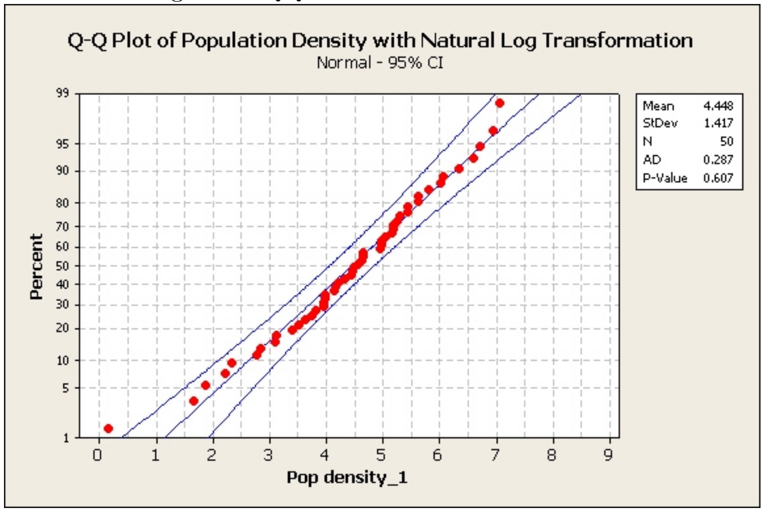


Table 2 : List of Normalized Variables

Variable	r-value	Transformation	r-value after Transformation
% Unemployment Rate	.994		
Population Density per square mile	.819	LN(x)	.987
Property Crime per 100,000 people	.985		
Foreclosure Rate per 1,000 people	.905	SQRT(x)	.984
Mortgage Rate on 30 Year Fixed Loan	.995		
Average Credit Score	.993		
% Increase in Foreclosures	.967	SQRT($x+75$)	.995
Homeowners Insurance	.949	LOG(x)	.983
% Poverty Level	.989		
% Increase in Value of Home	.915	SQRT$(x+11)^4$	.983
Population Growth 2000-2006	.938	LN($x+7$)	.966*
Median House Price 2006	.916	LN(x)	.979
2000-2004 Percent of Bachelors Degrees	.981		

II. Principal Component Analysis

Theoretical Background

Principal Component Analysis (PCA) is a method used to reduce the dimensionality of the variables, since a large data set can be difficult to manage. The application of PCA transforms p variables into *principal components* which are linear combinations of $X_1, X_2, ..., X_p$. By grouping together similar variables, we can reduce the number of variables analyzed from our original selection to just a few principal components. The PCA will also reveal patterns in the data and help link the variables. The goal is to find as few principal components as possible, while accounting for the largest portion of the total sample variance $V\hat{a}r$.

Principal components rely upon the covariance matrix $\sum$ (or the correlation matrix R) of $X_1, X_2, ..., X_p$. Let the random vector $X^T = [X_1, X_2, ..., X_p]$ have the covariance matrix $\sum$ with eigenvalues $\lambda_1 \geq \lambda_2, \geq ... \geq \lambda_p \geq 0$.

Consider the linear combinations

$$Y_1 = a_1^T X = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$
$$Y_2 = a_2^T X = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$$
$$\vdots$$
$$Y_p = a_p^T X = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p$$

We obtain:

$$Var(Y_i) = a_i^T \sum a_i \text{ for } i = 1, 2, ..., p$$
$$Cov(Y_i, Y_k) = a_i^T \sum a_k \text{ for } i \neq k = 1, 2, ..p$$

The first principal component is the linear combination $a_1^T X$ that maximizes $Var(a_1^T X)$ subject to $a_1^T a_1$=1. The second principal component is the linear combination $a_2^T X$ that maximizes $Var(a_2^T X)$ subject to $a_2^T a_2$=1 and $Cov(a_1^T X a_2^T X)$=0. The ith principal component is the linear combination of $a_i^T X$ that maximizes the $Var(a_i^T X)$ subject to $a_i^T a_i$=1 and $Cov(a_i^T X, a_k^T X$=0 for $i \neq k$.

We will extract m principal components where $m < p$. To calculate the proportion of total variance explained by m principal components we use the following expression:

$$\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

The number of principal components can also be obtained by a visual analysis of the *scree plot*. The scree plot is constructed by placing the number of principal components m on the x-axis and the corresponding eigenvalues on the y-axis. The points on the graph represent the amount of total sample variance by each eigenvalue. Think of the scree plot as an arm, the number of components to be extracted is determined by where the elbow occurs.

Analysis

Let $x_1, x_2, ..., x_n$ be a random sample on $\mathbf{X}$. Let S and R denote the sample covariance and sample correlation matrix respectively. Using the Principal Component Analysis feature of MINITAB we obtain the linear combination and the variance contributed by each component. PCA can be done by using either the covariance or correlation matrix. To use the covariance matrix each variable must be in equivalent units. If the variables are not in equivalent units then the covariance matrix can give results that are greatly skewed and are not effective in the further analysis of the project. Since it is not practical to transform all of our variables into equivalent units we use the correlation matrix for our data set.

A result of running Principal Component Analysis in MINITAB, Figure 3 is the Scree Plot produced. The elbow seems to occur between 4 and 6 components. Since there is no distinct or clear elbow on our scree plot, we also look at the eigenvalues and take components with eigenvalue greater than 1. This keeps only 5 out of the 13 principal components. These 5 principal components account for 76.9% of the cumulative variance, shown in Table 3.

Figure 3: PCA Scree Plot

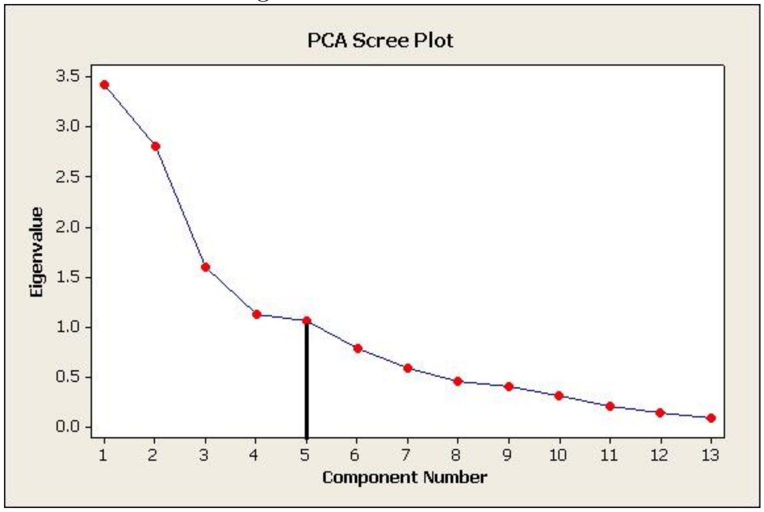


Table 3

Eigenvalue	3.4191	2.8024	1.5942	1.1198	1.0615
Proportions	0.263	0.216	0.123	0.086	0.082
Cumulative	0.263	0.479	0.601	0.687	0.769

After obtaining the 5 principal components the next step is to label these components. We label the first principal component **Economy** since it is heavily influenced by Foreclosure Rate per 1,000, Average Credit Score, Percent Increase in Value of Home, and Population Growth 2000-2006. We named PC2 **Social Environment** since it is heavily influenced by Population Density per Square Mile, Percent in Poverty Level, Median Home Price in 2006, and Percent Increase in Value of Home. We use this same technique to name the other principal components: PC3=**Population Growth,** PC4=**Employment/Education** and PC5=**Housing Market**.

Using Principal Component Analysis, we have reduced the dimensionality of our data to five principal components from the 13 variables. Next, we further analyze our data through Factor Analysis, where we will find underlying factors that influence the variables.

III. Factor Analysis

Theoretical Background

The purpose of Factor Analysis (FA) is to reduce the dimensionality of the variables by identifying underlying relationships among two or more variables. In doing so, we are grouping together highly correlated variables under a single factor. We hope to develop an m-factor model with m much less than p, the number of original variables.

We account for the variation of the variables using both common factors and unique factors. The $common factors$ $F_1, F_2, ..., F_m$ affect every variable, whereas $unique factors$ $\epsilon_1, \epsilon_2, ..., \epsilon_p$, also called errors, are specific to only the individual variable. Each variable is expressed as a linear combination of common factors with coefficients l_{ij}, called the loading of the jth factor on the ith variable, plus the unique factor. For example,

$$X_1 - \mu_1 = \ell_{11}F_1 + \ell_{12}F_2 + ... + \ell_{lm}F_m + \epsilon_1$$

where F_i's are the common factors and ϵ_1 is the unique factor.

In Factor Analysis, we assume $F_1, F_2, ..., F_m$ and $\epsilon_1, \epsilon_2, ..., \epsilon_p$ are independent with multivariate normal distribution. We also assume $E(F) = 0$, $Cov(F) = E(FF^T) = I$, $E(\epsilon)$=0, and $Cov(\epsilon) = E(\epsilon\epsilon^T) = \Psi$. Let $\mathbf{X} \sim \mathbf{MN}(\mu,\mathbf{\Sigma})$ with mean μ and covariance matrix Σ. We can show the covariance matrix can be factored as:

$$\Sigma = E(\mathbf{X}-\mu)(\mathbf{X}-\mu)^{\mathbf{T}} = \mathbf{LL^T} + \mathbf{\Psi} \text{ (m-factor model)}$$

where $\mathbf{\Psi}$ is a diagonal matrix composed of the variances (Ψ_i) of the unique factors ϵ_i, and L is a $p \times m$ matrix composed of the factor loadings l_{ij}.

Suppose $X_1, X_2...X_n$ is a random sample on $\mathbf{X}.$ To assess the adequacy of our m-factor model, we test the null hypothesis H_0: $\Sigma = LL^T + \Psi$ with a given m with H_1: Σ is any positive definite matrix. We then test the goodness of fit using the chi-squared test statistic. If χ^2 is greater than $\chi^2_{\upsilon,\alpha}$, we reject the $m-factor$ model and test H_o with higher value of m. We use the test statistic:

$$\chi^2 = [n - 1 - \frac{2p+5}{6} - \frac{2}{3}m] \ln \frac{\left|\widehat{\Psi} + \widehat{L}\widehat{L}^T\right|}{|R|}$$

where $\widehat{\Psi}$ and $\widehat{L}$ are the Maximum Likelihood Estimates of Ψ and L, and R is the sample correlation matrix. We compare the χ^2 to $\chi^2_{\upsilon,\alpha}$ at the α-level with:

$$\upsilon = \frac{(p-m)^2 - p - m}{2}$$

degrees of freedom to determine if m factors describe an adequate model for the analysis.

Analysis

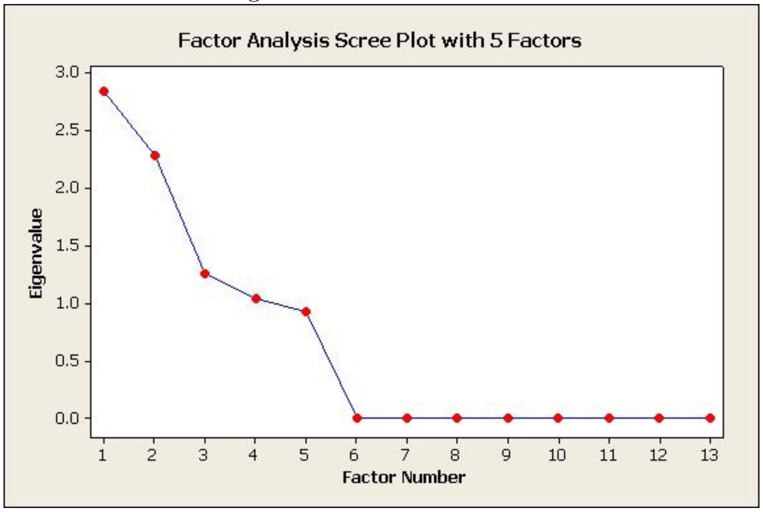
Our goal in performing Factor Analysis is to obtain m factors that model the variation of all 13 variables. Observing the scree plot in FA using MINITAB, we are able to deduce a reasonable place to begin our m-factor extraction.

In Figure 4, the elbow seems to appear at six components. We start our analysis using six factors which extracts 76% of the total variance. In addition, the factor model needs to pass the chi-squared test of adequacy with a p-value greater than α=0.05. Therefore, we start the test of adequacy using $m = 6$. Using $m = 6$, we compare the test statistic against the χ^2 with $\alpha = .05$ and degrees of freedom $\upsilon = 15$. The test statistic is 7.28 which is less than the χ^2 value of 25, and has a p-value of 0.95 which is greater than α=.05. This result shows that $m = 6$ passes the test of adequacy. Further, to find the lowest possible value of m that still passes the test of adequacy, we repeat this test with $m = 5, 4, 3,$ and 2 until we reject at $m = 2$ since the p-value of 0.01 is less than α=.05, and the test statistic is 90.78 which is greater than the χ^2 value of 67.5 with degrees of freedom $\upsilon = 53$. Table 4 shows these results.

Table 4:Test for Adequacy

m-Factor	Test Statistic	χ^2	Degrees of Freedom	P-Value	% Variance
6	7.28	25	15	.95	0.677
5	15.14	35.17	23	.88	0.632
4	26.31	43.7	32	.75	0.556
3	45.53	55.7	42	.33	0.454
2	90.78	67.5	53	.01	0.330

Figure 4: Factor Scree Plot



We label each factor by considering the variables with the greatest influence on the factor. Since the first factor places emphasis on Foreclosure Rate per 1,000 people, Mortgage Rate on a 30 year fixed loan, Average Credit Score, and Population Growth between 2000-2006, we labeled it **Economy.** In a similar manner, we named the second through sixth factor, **Social Environment**, **Housing Market**, **Population**, **Affordability**, and **Education/Home Value**.

IV. Discriminant Analysis

Theoretical Background

Discriminant Analysis (DA) is a statistical technique we use to classify a state into one of two mutually exclusive groups, π_1 and π_2, on the basis of a set of independent variables. We use a linear combination of all the variables to distinguish between the two groups.

We define the two multivariate normal subgroups as π_1 and π_2 where $\pi_1\sim$ MN(μ_1, Σ_1) and $\pi_2\sim$MN(μ_2, Σ_2). It is assumed π_1 is the superior group while π_2 the inferior group. It is also assumed at the beginning of this analysis that $\Sigma_1 = \Sigma_2$, and therefore, Linear Discriminant Analysis is the first analysis used in DA.

For instance, suppose we have two training samples of sizes n_1 and n_2 from π_1 and π_2 respectively. Each with sample covariance of S_1 and S_2 respectively. The test of the null hypothesis $\Sigma_1 = \Sigma_2$ is performed by first calculating the pooled unbiased estimate of the common covariance matrix under H_0 Σ, S_p where

$$S_p = \frac{1}{n_1+n_2-2}(\sum_{i=1}^{2}(n_i - 1)S_i).$$

S_p is then used to find the test statistic. The test statistic is stated as $\dfrac{M}{c}$ where,

$$M = \sum_{i=1}^{2}(n_i - 1)\ln(det(S_p)) - \sum_{i=1}^{2}[(n_i - 1)\ln(det(S_i))] \text{ and}$$

$$\frac{1}{c} = 1 - \frac{2p^2 - 3p - 1}{6(p-1)}(\frac{1}{n_1} + \frac{1}{n_2} - \frac{1}{n_1+n_2-2})$$

The test statistic has χ^2 distribution with $\upsilon = \frac{1}{2}p(p+1)$ degrees of freedom under H_0. If the test statistic is less than $\chi^2_{\alpha,\upsilon}$, then the null hypothesis is accepted. Otherwise, the null hypothesis is rejected, thus $\Sigma_1 \neq \Sigma_2$. We then must apply the Quadratic Classification in a similar way.

After determining the classification rule, we calculate the Apparent Error Rate (APER). The APER measures how well our model is classifying the training sample and gives us the percentage of the observations that are misclassified; the objective is to obtain a small APER. MINITAB will output the proportion correct which we then subtract from one, to find the percentage of misclassified data. In addition, the Total Probability of Misclassification (TPM) is calculated to further analyze the accuracy of the method. To find TPM, we must first find the estimated Mahalanobis distance (M-distance) between the two populations. The M-distance squared is,

$$(\hat{\Delta}_p)^2 = (\overline{x_1} - \overline{x_2})^T(S_p^{-1})(\overline{x_1} - \overline{x_2})$$

where $\overline{x_1}$ and $\overline{x_2}$ are the sample mean vectors for the two training samples.

The last step in Discriminant Analysis is to classify the remaining twenty states from our test sample into π_1 or π_2. We use the DA function of MINITAB to yield a prediction for each state based on the classification rule. It will classify each state as a stable or risky investment based on the M-distance and tell us the probability of correct classification. We can then use this to rank these twenty states from most stable to most risky.

Analysis

The first step in DA is to separate the sample population into two mutually exclusive groups. We divide the states into π_1, a stable investment, or π_2, a risky investment. Our criteria for separation is a linear combination of two variables, Percentage Increase in Foreclosures and Percentage Increase in the Value of a Home. We used the following equation to separate into π_1 and π_2, while being sure to account for the difference in units.

$$(.5)(\%\ Increase\ of\ Home\ Value) - (.5)(\%\ Increase\ In\ Foreclosures)\left(\frac{Med\ Increase\ of\ Home\ Value}{Med\ Increase\ in\ Foreclosures}\right)$$

Taking the median value of the linear combination stated above, -3.38, we classify the 25 states with a value less than -3.38 into π_2, risky, and the 25 states with a value greater than -3.38 into π_1, stable, shown in Table 5. We then took a random sample of $n_1 = n_2 =$15 from both groups, π_1 and π_2, to use as our training sample in DA. The testing sample are the remaining 20 states.

Table 5: Training Sample Classifications

Risky	Stable
Arizona	Arkansas
California	Colorado
Georgia	Idaho
Iowa	Indiana
Massachusetts	Kansas
Minnesota	Kentucky
Missouri	Maine
Nebraska	Montana
Nevada	North Carolina
New Hampshire	Oklahoma
New Jersey	Oregon
Ohio	Rhode Island
Pennsylvania	Utah
Vermont	Washington
West Virginia	Wisconsin

Next, we determine whether linear or quadratic discriminant analysis is appropriate for our given data set. We do this by carrying out a hypothesis test with the null hypothesis: $H_0 : \Sigma_1 = \Sigma_2$ against $H_1 : \Sigma_1 \neq \Sigma_2$. We find the test statistic, $\frac{M}{c}$=137.85 with a $p-value$=.0011175. The corresponding chi-squared value for υ= 91 degrees of freedom, is 113.15. Our test statistic is greater than the chi-squared value, so we are unable to accept $\Sigma_1 = \Sigma_2$ at α=.05.

Even though we reject $\Sigma_1 = \Sigma_2$, we run both Linear and Quadratic discriminant analysis using MINITAB on the training sample data to find the Apparent Error Rate (APER). With Linear DA, we find that all fifteen in each group were placed correctly, giving us an APER of 0%, which can be seen in Table 5.

Table 6 : Training Sample Results with Linear Discriminant Analysis

	Classified: Stable	Classified: Risky	Total
True: Stable	15	0	15
True: Risky	0	15	15
Correct	15	15	

N=30 N Correct=30 Proportion Correct=1.00

Since this is an extremely low error rate, we proceeded to calculate the TPM. Using SAS, another Statistical Computing program, we find TPM = .0796 $\approx$ 8%. Lastly, we run Linear DA for the twenty states in the training sample, where 10 were classified as risky and 10 were classified as stable. This correctly predicted the placement of 17 states into their respective grouping. We follow up our analysis by running a Quadratic Discriminant Analysis on our data as well. This time we still have a 0% APER, but the Quadratic DA only classified 13 out of the 20 states in our test sample correctly. Despite rejecting $\Sigma_1 = \Sigma_2$, we stay with the Linear DA because it is a more accurate classification of the test sample with the same APER as Quadratic DA. Table 6 shows the 20 states in our test sample, their true group, the squared distance, probability of correct classification, predicted group and ranking from the Linear Discriminant Analysis from MINITAB. The Linear DA correctly classified all 10 from the stable test sample, but only 7 from the risky test sample. The states are ranked from 1 to 20 with 1 being the most stable state and 20 being the most risky state based on their M-distance. The smaller the M-distance to stable, the more stable the state is. Similarly, the smaller the M-distance to risky, the more risky a state is. Thus, amongst the test sample states, Illinois is the best state to invest in and Connecticut is the worst state to invest in based on the Linear DA.

Table 7:Linear Discriminant Analysis Test Sample Results

State	True Group	Squared Distance	Probability	Predicted Group	Rank
Illinois	R	7.558	0.997	S	1
Tennessee	S	18.017	0.944	S	2
South Dakota	S	20.655	0.999	S	3
Mississippi	S	22.070	0.999	S	4
South Carolina	S	24.944	0.917	S	5
Wyoming	S	28.517	1.000	S	6
Alabama	S	30.506	1.000	S	7
New Mexico	R	33.827	0.596	S	8
Delaware	S	42.965	1.000	S	9
Hawaii	R	53.397	1.000	S	10
North Dakota	S	53.802	1.000	S	11
Texas	S	55.544	0.999	S	12
Louisiana	S	177.526	1.000	S	13
Alaska	R	116.679	0.560	R	14
Florida	R	38.750	0.914	R	15
Michigan	R	32.274	1.000	R	16
New York	R	21.439	0.997	R	17
Maryland	R	20.582	1.000	R	18
Virginia	R	18.335	1.000	R	19
Connecticut	R	16.037	1.000	R	20

V. Conclusion

We started our statistical analysis with 15 variables. Two variables, Cost of Living and Median Income, were highly correlated with other variables, so we removed them prior to our analysis. Next, we assessed the normality of our data using QQ plots and a test for linearity. After evaluating normality for the remaining 13 variables, we applied transformations to seven of our variables to achieve normality. Our next step was to perform Principal Component Analysis to reduce the dimensionality from 13 variables to five principal components which accounted for 76.9% of the variance. We labeled these five principal components, Economy, Social Environment, Population Growth, Employment/Education and Housing Market. This allowed us to better evaluate the variables that were having the greatest impact on our data.

After PCA we applied a second technique called Factor Analysis to reduce the dimensionality by finding the underlying structure of our data set. We were able to accept a 6-factor model since it accounted for 67.7% of the variance and passed a Test of Adequacy with a 95% confidence level. We labeled these factors as Economy, Social Environment, Housing Market, Population, Affordability, and Education/Home Value.

Finally we used Discriminant Analysis to classify the states as a risky or stable investment. We used the linear combination of Percentage Increase in Home Value and Percentage Increase in Foreclosures to separate the 50 states into π_1, (Stable) and π_2 (Risky) . Then we took a random sample of 15 states from π_1 and π_2 to use as our training sample. We first ran Linear Discriminant Analysis on these thirty states, and calculated the Apparent Error Rate (APER) and Total Probability of Misclassification (TPM). The APER was 0%, so we tested the null hypothesis, $\Sigma_1 = \Sigma_2$ for the training sample. Since we rejected $\Sigma_1 = \Sigma_2$, we ran the Quadratic Discriminant Analysis as well which also had an APER of 0%. Thus we used MINITAB to predict the classification for the remaining 20 states, called the test sample as either a stable or risky investment. Out of the 20 states in the test sample, 17 were classified correctly using Linear Discriminant Analysis, and only 13 were classified correctly using Quadratic Discriminant Analysis. Despite rejecting $\Sigma_1 = \Sigma_2$ we kept the Linear Discriminant Analysis because more accurately classified the twenty states in the test sample. Using the Mahalanobis distance we ranked the twenty states in the test sample from 1, being the most stable state, to 20, being the most risky state to invest in. Amongst the test sample states, Illinois is the best state to invest in and Connecticut is the worst state to invest in based on the Linear Discriminant Analysis.

In testing $\Sigma_1 = \Sigma_2$, we encountered problems with our data set. We decided to investigate our data to see if it was unstable. Looking through the covariance matrix we noticed one of our variables, Mortgage Rate on a 30-year Fixed Loan, had a variance close to zero since the values ranged from 6.19 to 6.44. Due to time constraints, we were unable to run PCA, FA and DA again without this variable. One way of continuing this research would be to look at the same data set while suppressing the variable, Mortgage Rate on a 30-year Fixed Loan, to test the null hypothesis $\Sigma_1 = \Sigma_2$. Another way to expand this research would be to change the focus of the analysis. Instead of looking at the data from an investment standpoint, one could also account for variables that would affect personal preference. This research would unveil a more detailed analysis of the US Housing Market.

Acknowledgments

We would like to thank Dr. Vasant Waikar for his time teaching and facilitating this research project. We would also like to show our gratitude to Dr. Tom Farmer for his supportive assistance during the revision and editing of our report. We thank Dr. Dennis Davenport, co-director of SUMSRI, and Mrs. Bonita Porter, program coordinator for their time and effort. We also would like to acknowledge our graduate assistant, Kevin Tolliver, and all of the remaining SUMSRI participants, graduate assistants, and professors for their help, support, and camaraderie throughout this program. Finally, we would like to show gratitude to the National Security Agency (NSA), the National Science Foundation (NSF), and Miami University for giving us this opportunity.

Bibliography

[1] CBSNEWS, *2007 Foreclosure Rates*: Retrieved 23 June 2008 from http://www.cbsnews.com/elements/2008/02/12/business/map3823166.shtml.

[2] Dillon, William R., and Matthew Goldstein. Multivariate Analysis Methods and Applications. New York: John Wiley and Sons, 1984.

[3] Federal Bureau of Investigation, *Crime in the United States by State, 2006*: Retrieved 24 June 2008 from $http://www.fbi.gov/ucr/cius2006/data/table_05.html$.

[4] Home Improvement Web, *State By State Average Consumer Credit Score, Ranking, and Ratings(2006)*: Retrieved 23 June 2008 from http://www.homeimprovementweb.com/information/finance-money/state-credit-scores.htm.

[5] Insurance Information Institute, *Average Premiums for Homeowners and Renters Insurance, by State, 2005*: Retrieved 23 June 2008 from http://www.iii.org/media/facts/statsbyissue/homeowners/.

[6] Johnson, Richard A., and Dean W. Wichern. Applied Multivariate Statistical Analysis. 6th ed. Upper Saddle River: Pearson Prentice Hall, 2007.

[7] Office of Federal Housing Enterprise Oversight, *Percent Change in House Prices Period Ended March 31, 2008*: Retrieved 23 June 2008 from http://www.ofheo.gov/media/pdf/1q08hpi.pdf.

[8] RealtyTrac, *National Foreclosures Increase In Every Quarter Of 2005 According TO RealtyTrac U.S. Foreclosure Market Report*: Retrieved 23 June 2008 from http://www.realtytrac.com/news/press/pressRelease.asp?PressReleaseID=86.

[9] StateMaster, *Bachelors degree or higher by percentage (most recent) by state*: Retrieved 24 June 2008 from $http://www.statemaster.com/graph/edu\_bac\_deg\_or\_hig\_by\_per - bachelor - s - degree - higher - percentage$.

[10] Top 50 State, *Causes of Poverty in the United States*: Retrieved 23 June 2008 from http://www.top50states.com/causes-poverty-united-states.html.

[11] Top 50 State, *Cost of Living By State*: Retrieved 23 June 2008 from http://www.top50states.com/cost-of-living-by-state.html.

[12] Top 50 States, *Average Job Salaries(2007)*: Retrieved 23 June 2008 from http://www.top50states.com/average-job-salaries.html.

[13] Top 50 States, *Florida Unemployment Statistics: Unemployment Rates By State*: Retrieved 23 June 2008 from http://www.top50states.com/florida-unemployment.html.

[14] World Atlas, *Population Density of All U.S. States Smallest to Largest*: Retrieved 23 June 2008 from http://www.worldatlas.com/aatlas/populations/usadensityh.htm.

Appendix 1: Data Set

State	% Unemployment Rate	Pop Density per Square Mile	Property Crime per 100,000 people
Alabama	4.1	4.49781	3936.1
Alaska	6.7	0.14842	3604.9
Arizona	4.0	3.95642	4627.9
Arkansas	4.9	3.97744	3967.5
California	6.2	5.44536	3170.9
Colorado	4.4	3.80622	3451.3
Connecticut	5.3	6.58551	2504.1
Delaware	3.8	6.06771	3417.9
Florida	4.9	5.79876	3986.1
Georgia	5.3	5.05421	3889.2
Hawaii	3.1	5.29099	4230.4
Idaho	3.0	2.84897	2418.8
Illinois	5.5	5.43643	3019.6
Indiana	5.1	5.16404	3502.4
Iowa	3.5	3.97199	2802.7
Kansas	4.1	3.51304	3750.2
Kentucky	5.7	4.65444	2544.5
Louisiana	4.5	4.64285	3993.7
Maine	5.0	3.75701	2518.7
Maryland	3.6	6.35087	3480.9
Massachusetts	4.4	6.70462	2391.0
Michigan	7.2	5.18274	3212.8
Minnesota	4.7	4.16620	3079.5
Mississippi	5.9	4.13148	3208.8
Missouri	5.7	4.43319	3826.5
Montana	3.6	1.86097	2687.5
Nebraska	2.9	3.13026	3340.7
Nevada	5.8	3.09059	4088.8
New Hampshire	3.9	4.98409	1874.1
New Jersey	4.8	7.06936	2291.9
New Mexico	3.7	2.76569	3937.2
New York	4.8	6.01083	2052.7
North Carolina	5.2	5.18324	4120.8
North Dakota	3.1	2.22246	2000.3
Ohio	5.7	5.63468	3678.6
Oklahoma	3.1	3.94488	3604.2
Oregon	5.7	3.63574	3672.1
Pennsylvania	4.9	5.62524	2443.5
Rhode Island	6.1	6.93717	2586.9
South Carolina	6.0	4.95103	4242.3
South Dakota	2.5	2.32532	1619.6
Tennessee	5.6	4.97446	4128.3
Texas	4.3	4.46958	4081.5
Utah	3.3	3.40320	3516.4
Vermont	4.6	4.21005	2304.7
Virginia	3.7	5.25295	2478.2
Washington	4.9	4.54849	4480.0
West Virginia	4.7	4.32360	2621.5
Wisconsin	4.8	4.62438	2817.8
Wyoming	3.1	1.65823	2980.6

Appendix 1 cont.: Data Set

State	Foreclosure Rate per 1,000 Homes	Mortgage Rate on 30 year Fixed Home	Average Credit Score
AL	1.64317	6.26	670
AK	2.21359	6.31	674
AZ	3.89872	6.28	659
AR	2.25832	6.30	666
CA	4.38178	6.25	672
CO	4.38178	6.22	671
CT	2.88097	6.25	695
DE	1.64317	6.40	679
FL	4.47214	6.19	672
GA	3.96232	6.21	662
HI	1.41421	6.41	690
ID	2.46982	6.30	684
IL	3.53553	6.29	682
IN	3.20936	6.25	672
IA	1.76068	6.34	697
KS	1.41421	6.38	679
KY	1.64317	6.30	675
LA	1.41421	6.44	662
ME	0.63246	6.32	699
MD	2.88097	6.21	684
MA	2.56905	6.25	701
MI	4.41588	6.27	677
MN	2.25832	6.27	706
MS	1.04881	6.29	666
MO	3.01662	6.29	678
MT	1.64317	6.36	703
NE	2.16795	6.33	694
NV	5.81378	6.36	654
NH	1.44914	6.30	704
NJ	3.00000	6.33	691
NM	1.89737	6.28	661
NY	2.21359	6.36	685
NC	2.72029	6.24	663
ND	0.89443	6.37	705
OH	4.24264	6.33	681
OK	2.28035	6.31	664
OR	2.32379	6.28	686
PA	1.73205	6.30	692
RI	2.02485	6.34	693
SC	1.48324	6.24	663
SD	0.31623	6.34	708
TN	3.13050	6.22	674
TX	3.06594	6.28	647
UT	2.91548	6.30	684
VT	0.31623	6.41	707
VA	2.25832	6.25	689
WA	2.38747	6.27	685
WV	0.70711	6.39	675
WI	2.21359	6.37	695
WY	1.22474	6.31	689

Appendix 1 cont.: Data Set

State	% increase in Foreclosures	Homeowners Insurance	% Poverty Level	% Increase in value of home
AL	4.4788	6.74170	16.0	239.012
AK	9.9624	6.72982	9.3	152.276
AZ	9.1088	6.45362	14.7	30.250
AR	9.6778	6.65286	15.6	171.872
CA	9.5310	6.79682	12.9	0.176
CO	8.4362	6.69332	10.4	176.624
CT	16.2358	6.71296	9.1	111.092
DE	0.4583	6.21060	9.2	144.480
FL	6.7713	6.98749	11.4	8.122
GA	10.2333	6.51026	13.3	175.033
HI	7.4498	6.72982	8.8	114.490
ID	5.8652	6.12468	9.8	183.602
IL	7.7936	6.49224	11.5	135.956
IN	9.4594	6.45990	11.6	175.298
IA	11.5209	6.38688	10.8	190.440
KS	9.7816	6.72863	12.2	187.690
KY	5.9758	6.44254	16.5	190.716
LA	4.2356	7.04229	17.4	204.490
ME	5.1962	6.31536	11.5	177.422
MD	13.8597	6.54535	9.3	94.284
MA	16.5587	6.71780	10.5	84.824
MI	15.6723	6.59851	12.9	63.044
MN	10.6818	6.67203	7.7	99.800
MS	7.5895	6.84482	19.8	197.684
MO	11.1355	6.53379	11.7	160.276
MT	11.5853	6.49677	13.8	253.128
NE	13.1875	6.64249	9.7	169.260
NV	9.7232	6.50877	10.4	0.490
NH	8.6603	6.45520	5.5	102.617
NJ	13.5303	6.52503	7.9	97.022
NM	12.4173	6.40523	17.1	206.784
NY	14.6823	6.73578	14.5	125.888
NC	8.5434	6.46770	13.8	225.901
ND	3.2078	6.60123	10.8	231.344
OH	12.9341	6.27476	12.0	135.490
OK	9.0515	6.90375	13.9	225.000
OR	6.5704	6.19644	11.9	174.240
PA	12.0499	6.43615	11.3	179.560
RI	2.8862	6.74406	11.3	63.362
SC	7.9498	6.70564	13.7	219.632
SD	7.0385	6.43294	12.0	231.344
TN	10.1789	6.53959	15.2	222.010
TX	11.3389	7.22402	16.4	246.176
UT	6.8964	6.16752	9.5	274.896
VT	9.7468	6.47080	7.7	163.073
VA	15.0526	6.46303	9.1	119.684
WA	6.7772	6.37843	9.9	193.766
WV	10.6104	6.46147	15.0	181.441
WI	8.5387	6.20456	10.9	163.840
WY	8.5393	6.47543	10.2	300.676

Appendix 1 cont.: Data Set

State	Population Growth between 2000-2006	Median House Price 2006	2000-2004 % Increase Bachelors Degrees
AL	2.34181	11.5806	17.3684
AK	2.63189	12.2700	3.2389
AZ	3.30322	12.3737	19.1489
AR	2.49321	11.4500	12.5749
CA	2.68102	13.1913	19.1729
CO	2.86220	12.3584	8.5627
CT	2.29253	12.6079	9.8726
DE	2.76632	12.3331	7.6000
FL	3.00568	12.3484	16.5919
GA	3.06339	11.9627	13.5802
HI	2.57261	13.1801	1.5267
ID	3.01062	12.0070	9.6774
IL	2.33214	12.2071	4.9808
IN	2.37955	11.7011	8.7629
IA	2.18605	11.6316	14.6226
KS	2.28238	11.6475	16.2791
KY	2.40695	11.6173	22.8070
LA	1.06471	11.6501	19.7861
ME	2.37024	12.0465	5.6769
MD	2.56495	12.7210	12.1019
MA	2.12823	12.8223	10.5422
MI	2.12823	12.8223	10.5422
MN	2.48491	12.2463	18.6131
MS	2.23001	11.3919	18.9349
MO	2.43361	11.7898	30.0926
MT	2.45959	11.9544	4.5082
NE	2.33214	11.6886	4.6414
NV	3.46261	12.6610	34.6154
NH	2.59525	12.4419	23.3449
NJ	2.37024	12.8120	16.1074
NM	2.67415	11.8579	6.8085
NY	2.16332	12.6228	11.6788
NC	2.83908	11.8292	4.0000
ND	1.79176	11.5099	14.5455
OH	2.09186	11.8145	16.5877
OK	2.37024	11.4564	12.8079
OR	2.72130	12.3741	3.1873
PA	2.11626	11.8859	12.9464
RI	2.17475	12.5971	6.2500
SC	2.68785	11.7150	22.0588
SD	2.36085	11.6316	18.6047
TN	2.57261	11.7208	23.9796
TX	2.98062	11.6440	5.6034
UT	3.05400	12.1469	18.0077
VT	2.25129	12.1704	16.3265
VA	2.74084	12.4972	7.9422
WA	2.70805	12.4057	12.2034
WV	2.02815	11.4042	3.3784
WI	2.36085	12.0046	14.2857
WY	2.42480	11.9110	2.7397