# Final Project

Keith Mitchell

3/7/2020

STA135 Final Project (UC Davis)

Instructor: Professor Li

TA: Cong Xu

---

## Dataset 1: Conduct multiple linear regression;

## Dataset 2: Conduct two-sample test and LDA;

## Dataset 3: Conduct PCA.

**For each data analysis, you should write in full sentences, and have thefollowing sections for the body of your report.**

1) Introduction: Briefly summarize the goal of the analysis in your own words;
2) Summary: Summarize your data by plots or sample estimates;
3) Analysis: Implement the analysis based on what you have done in homework;
4) Conclusion: Describe and interpret your findings.

**Details:**

1) A title page including your name, the name of the class, and the name of your instructor.
2) Do not include R code in the body of your report. R code used to produce the results should all go to the appendix.(echo=FALSE, results='asis')
3) Typed.
4) Double-sided pages.

---

# DATASET 1: Multiple Linear Regression

## INTRODUCTION:

**A company is interested in considering the purchase of a computer but first must assess their future needs in order to determine the proper equipment. A computer scientist collected data from seven similar company sites so that a forecast equation of computer-hardware requirements for inventory management could be developed. (pg 380) This is an important concept for companies running web infrastructure to predict for a variety of reasons.**

1) It is preferable not to overpay for CPU time that will likely not get used for the company web platform.
2) It is important for the stability of the web platform that it has the necessary CPU resources to maintain the stability of their web platform to keep customers happy and making purchases. ### The goal of this analysis will be to introduce a model that can best predict the necessary CPU time needed given a set of their own variables ("Orders(thousands)" and "Add-delete items(thousands)") iwith a high degree of confidence.

## SUMMARIZE DATA: The first colum of our data is the Orders(in thousands), and the Second column is the Add-Delete items (in thousands)

**Lets summarize our data first and see what the numbers look like.**

| Orders(thousands) | Add-delete items(thousands) | CPU time(hours) |
|---:|---:|---:|
| 123.5 | 2.108 | 141.5 |
| 146.1 | 9.213 | 168.9 |
| 133.9 | 1.905 | 154.8 |
| 128.5 | 0.815 | 146.5 |
| 151.5 | 1.061 | 172.8 |
| 136.2 | 8.603 | 160.1 |
| 92.0 | 1.125 | 108.5 |

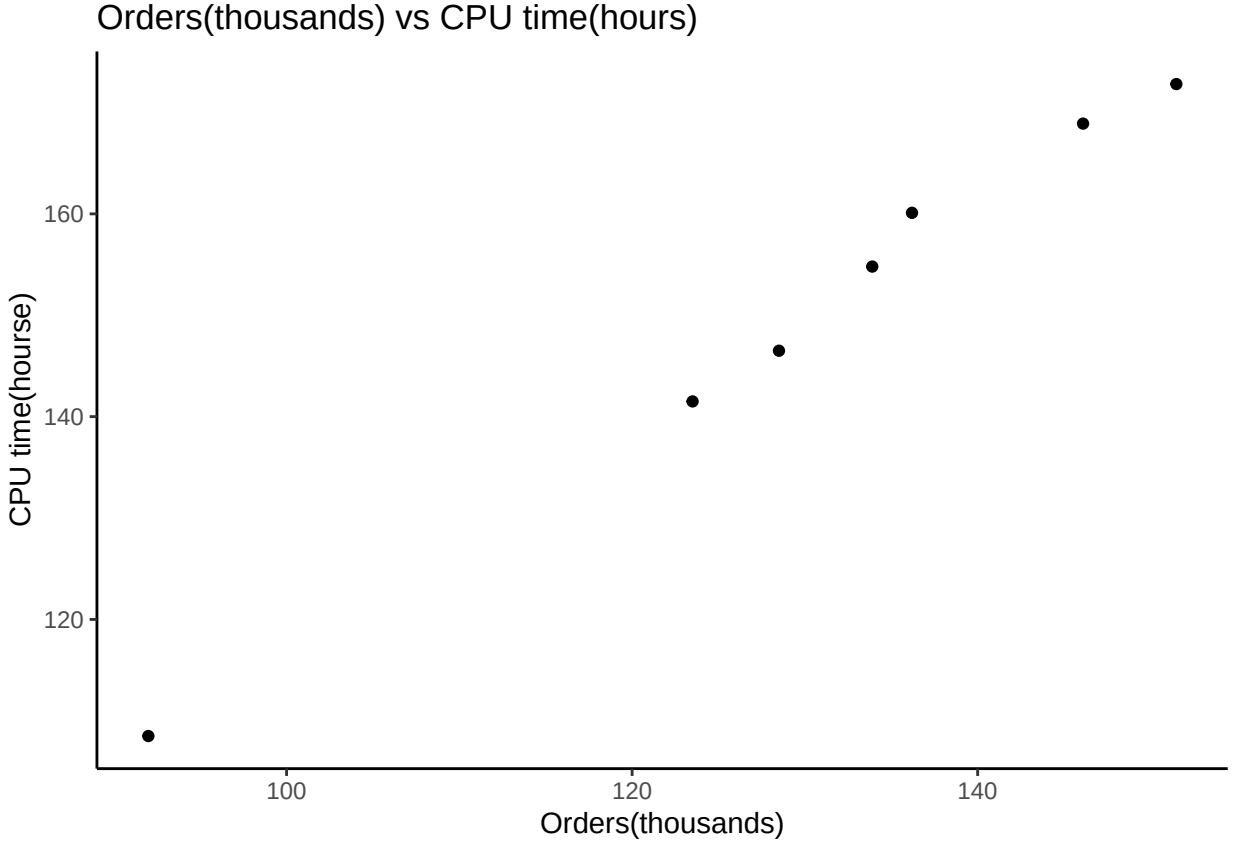
- This table is the data for Z, the predictors, and for y the variable to be forecasted from the predictors.

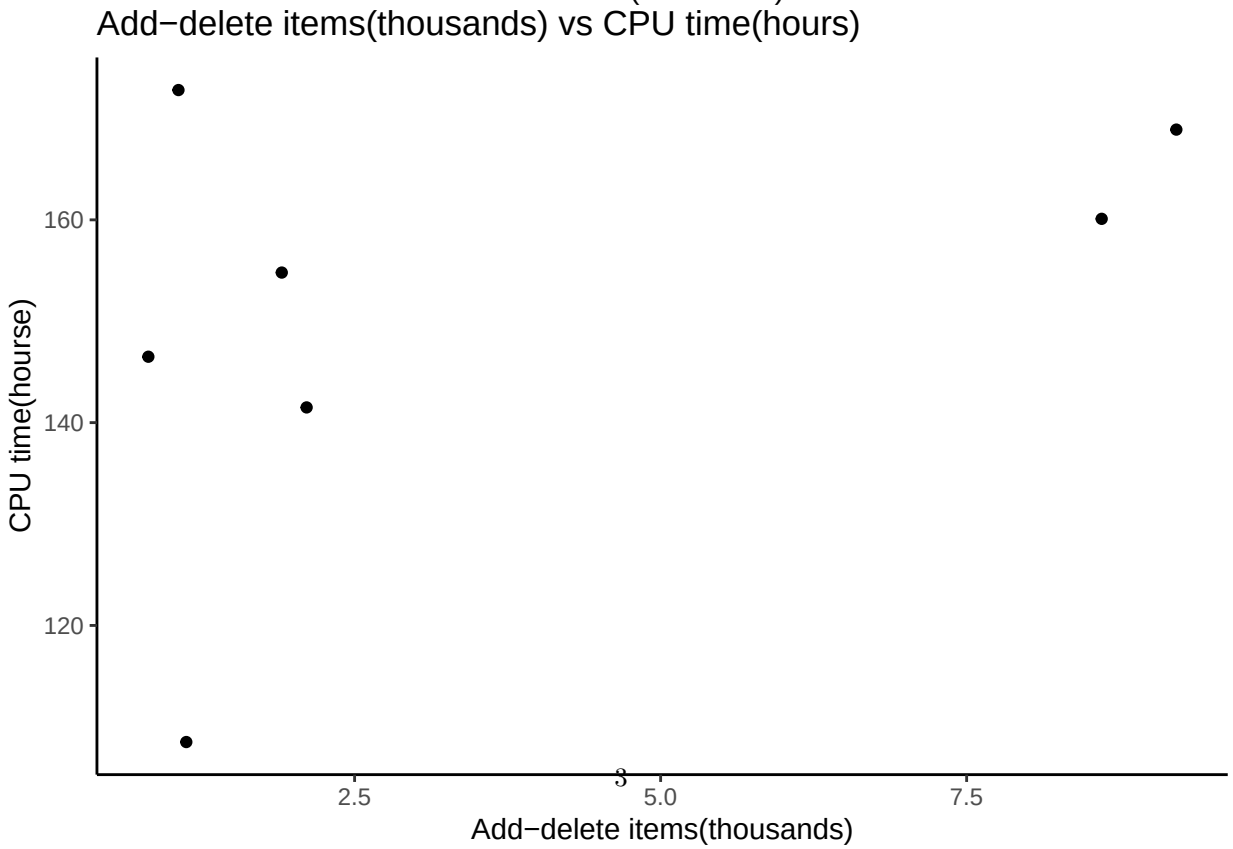
| | Orders(thousands) | Add-delete items(thousands) | CPU time(hours) |
|---|---|---|---|
| | Min. : 92.0 | Min. :0.815 | Min. :108.5 |
| | 1st Qu.:126.0 | 1st Qu.:1.093 | 1st Qu.:144.0 |
| | Median :133.9 | Median :1.905 | Median :154.8 |
| | Mean :130.2 | Mean :3.547 | Mean :150.4 |
| | 3rd Qu.:141.2 | 3rd Qu.:5.356 | 3rd Qu.:164.5 |
| | Max. :151.5 | Max. :9.213 | Max. :172.8 |

- This table is the summary data for Z, the p redictors, and for y the variable to be forecasted from the predictors.

**Now lets visualize these numbers with a box plot of each of the variables, including the one we are trying to predict.**

**Lets also just get a preliminarly look at what our variables "Orders(thousands)" andn "Add-delete items(thousands)" look like when graphed against the variable we hope to predict "CPU time"**



**We can already see some patterns in the data but lets formalize this by performing a multiple linear regression to allow the company to predict their CPU Time with confidence based on the "Orders(thousands)" andn "Add-delete items(thousands)".**

## DATA ANALYSIS: Multiple linear regression.

**This is our final predictive function for the value of CPU time given the:**

- y = 1.07898250130454*z_1 + 0.419888473166963*z_2 + 8.42368896741667

**And we get a $r^2$ value of our linear prediction:**

- 0.997934852356086

**Next we find $\hat{\sigma^2}$ and $\hat{Cov}(\vec{\beta})$**

- n is the number of observed instances (or data to learn from) for each r.
- This is equal to the number of rows in Z = 7
- r is the number of observed variables for which each n has a value.
- So this is number of columns of data in Z = 2

$\hat{\sigma^2}$

- 1.44946370175246

$\hat{Cov}(\vec{\beta})$

| | X1 | Orders.thousands. | Add.delete.items.thousands. |
|---|---|---|---|
| 1 | 11.8561719 | -0.0929277 | 0.1280071 |
| Orders(thousands) | -0.0929277 | 0.0007557 | -0.0015513 |
| Add-delete items(thousands) | 0.1280071 | -0.0015513 | 0.0208729 |

**Find a 95% confidence interval for the mean response $E(Y_0|z_0) = \beta_0 + \beta_1 z_{01} + \beta_2 z_{02}$ when $z_0$=[1,130,7.5] (for example)**

- Lower Bound (CPU time (hours): 149.807451077585
- Upper Bound (CPU time (hours): 153.873704293934

**Find a XX% prediction interval for the mean response $Y_0$ corresponding to $\bar{z_1}$, $\bar{z_2}$. This correspond to a 95% prediction interval for a new facility's CPU requirement corresponding to the same $z_0$.**

- Alpha value: 0.05 Interval: 3.91241721099891 Lower bound: 147.92816047476 Upper bound: 155.752994896758
- Alpha value: 0.01 Interval: 6.48784302704889 Lower bound: 145.35273465871 Upper bound: 158.328420712808
- Alpha value: 0.005 Interval: 7.88779247898333 Lower bound: 143.952785206776 Upper bound: 159.728370164743
- Alpha value: 0.001 Interval: 12.1331741930768 Lower bound: 139.707403492682 Upper bound: 163.973751878836

## CONCLUSION

**The company can use the following function to predict the amount of CPU time that they will need for a new set of parameters:**

- y = 1.07898250130454$z_1$ + 0.419888473166963$z_2$ + 8.42368896741667

- $z_1$ is the value of the companies number of orders (in thousands) that they need to know CPU time requirements for
- $z_2$ is the companies add-delete items (in thousands) that they need to know CPU time requirements for
- y is the predicted CPU time needed.

**Then, as generated in the last portion of the data analysis, we can see that if we want to be XX% (where an alpha value of 0.05 corresponds to 95%, alpha of 0.001 corresponds to 99.9% etc.) sure that the true value of CPU time will fall within the lower and upper bound we will add and subtract the values generated to the value of y predicted above for some value of $z_{01}$ being the orders in thousands and $z_{02}$ being the add-delete items in thousands.**

- Alpha value: 0.05 Interval: 3.91241721099891 Lower bound: 147.92816047476 Upper bound: 155.752994896758
- Alpha value: 0.01 Interval: 6.48784302704889 Lower bound: 145.35273465871 Upper bound: 158.328420712808
- Alpha value: 0.005 Interval: 7.88779247898333 Lower bound: 143.952785206776 Upper bound: 159.728370164743
- Alpha value: 0.001 Interval: 12.1331741930768 Lower bound: 139.707403492682 Upper bound: 163.973751878836

**In addtion, these values can automatically be generated with the following script by just altering the line z_0 <- as.matrix(c(1,130,7.5)) by replacing "130 and 7.5 in the following manner. The general line would be z_0 <- as.matrix(c(1,$z_{01}$,$z_{02}$)) where the value of $z_{01}$ is the orders in thousands and $z_{02}$ is the add-delete items in thousands.**

```
for (val in c(0.05,0.01,0.005,0.001))
{
  z_0 <- as.matrix(c(1,130,7.5))
  z_0_beta_hat <- t(z_0)%*%beta_hat
  statistic<-qt(1-(val/2),n-r-1)
  ans<-t(z_0)%*%(solve(t(Z)%*%Z))%*%z_0
  interval <- sqrt(sigma_sq)*statistic*sqrt(1+ans)
  int_low <- z_0_beta_hat - interval
  int_up <- z_0_beta_hat + interval
  print (sprintf("Alpha value: %s  Interval: %s  Lower bound: %s  Upper bound: %s",  val, interval, int
}
```

**As a precaution we do suggest gather more data on this situation if possible. This will increase the strength of the analysis and account for other variables in the situation. This would be achieved by increasing the value of n, or observed situaitons (company CPU time requirements), and r, the number of observed varaibles upon which to make the prediction of CPU time (orders, add-delete items, etc.). Please contact if this is the case and a more robust script will be made to factor in variations in new potential data gathering.**

---

# DATASET 2: Two Sample Test and LDA

## INTRODUCTION:

**Data has been gather on hemophilia A carriers.**

1) Predicting disease based on other characteristics is a common technique that doctors and healthcare workers use to produce a prediction for a given person.
2) This is a common method that if not used solely for diagnosing a disease is used in conjunction with other tests to increase the statistical power of the diagnoses.

**The first goal of this analysis will be to perform two sample test on the data in order to see if there is a significant difference between individuals that are carriers and noncarriers with respect to AHF activity ($log_{10}$ scale) and AHF antigen ($log_{10}$ scale). Group 1 is considered to be the non carrier group while group 2 is considered to be the obligatory carrier group.**

**The second goal of this analysis will be to perform linear discriminant analysis to try and predict the best possible distinguishable boundary between carriers and non carriers with respect to AHF activity ($log_{10}$ scale) and AHF antigen ($log_{10}$ scale).**

## SUMMARIZE DATA:

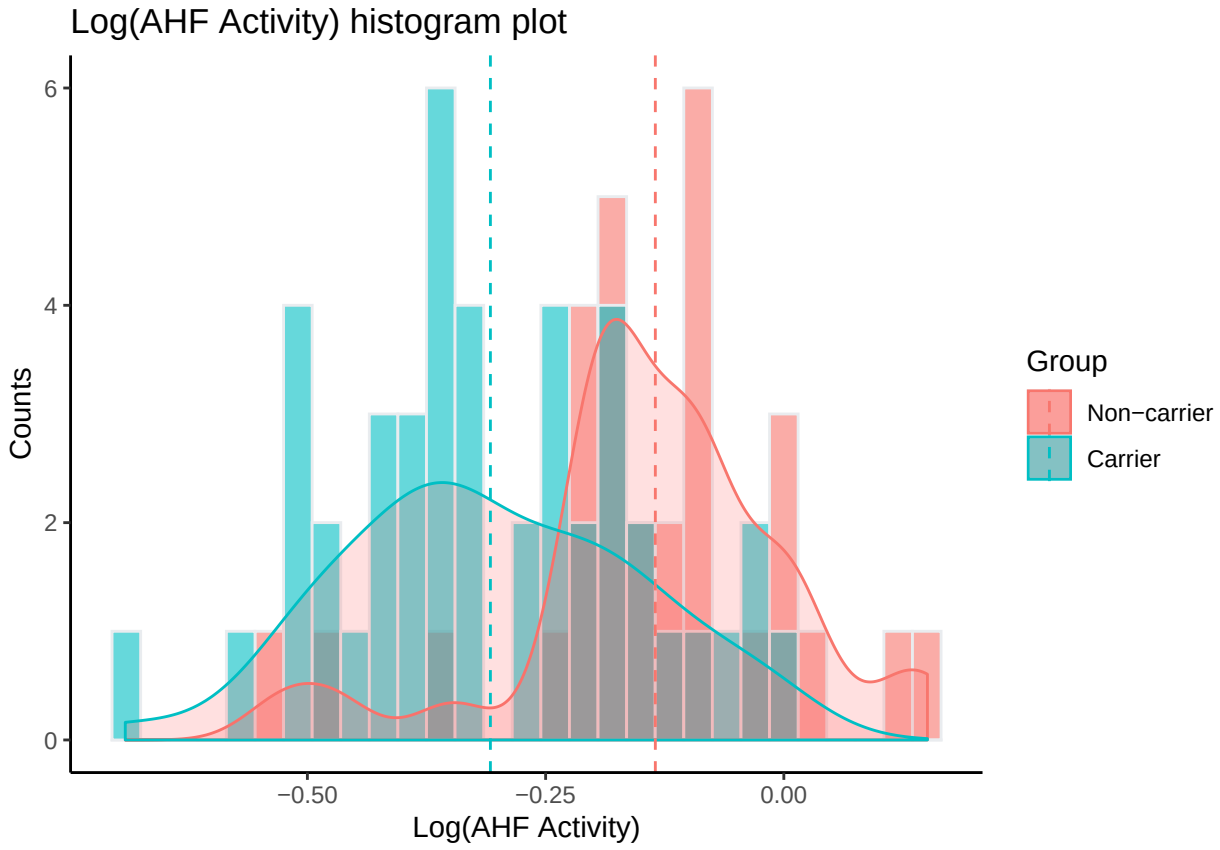
**Lets look at the data as a whole (with the two groups together (carriers and non carriers))**

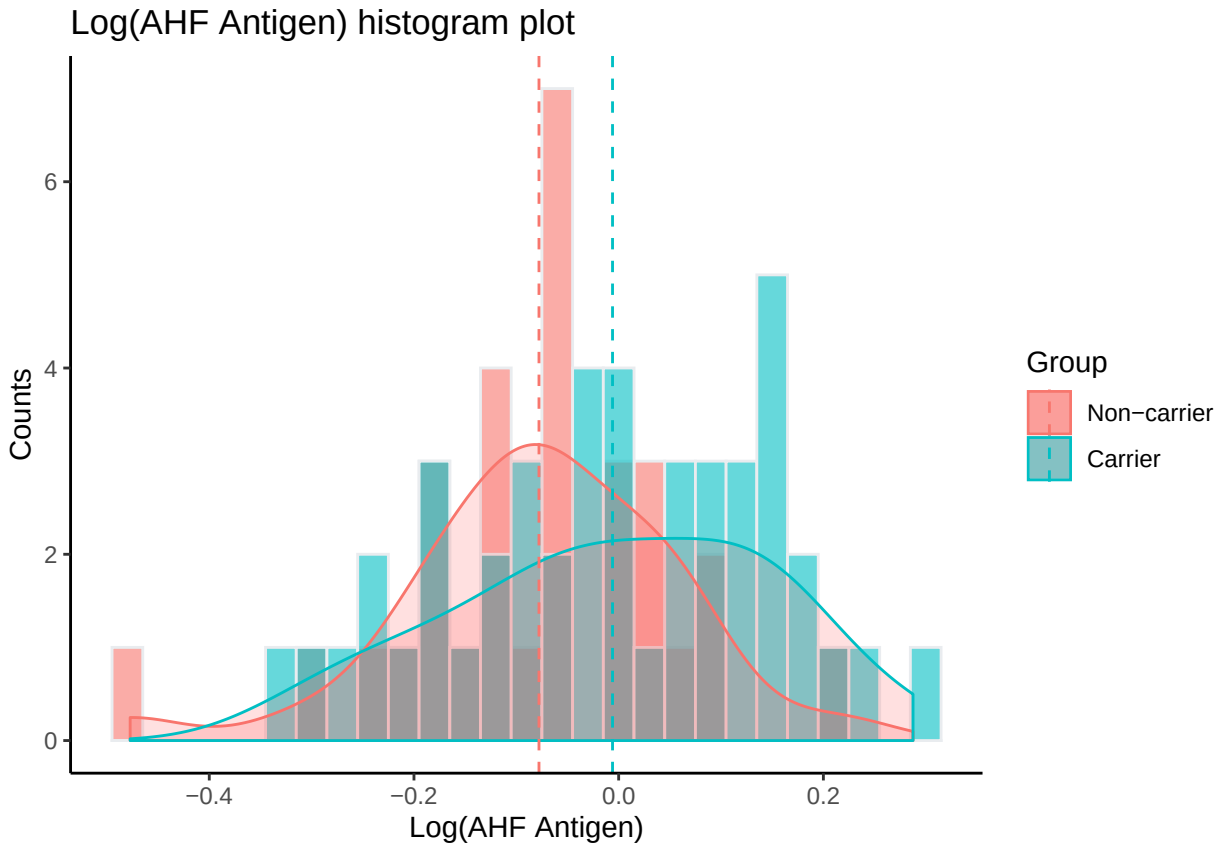
| | AHF_activity | AHF_antigen |
|---|---|---|
| | Min. :-0.52680 | Min. :-0.47730 |
| | 1st Qu.:-0.19470 | 1st Qu.:-0.14968 |
| | Median :-0.13890 | Median :-0.06775 |
| | Mean :-0.13487 | Mean :-0.07786 |
| | 3rd Qu.:-0.07805 | 3rd Qu.: 0.00955 |
| | Max. : 0.15070 | Max. : 0.21400 |

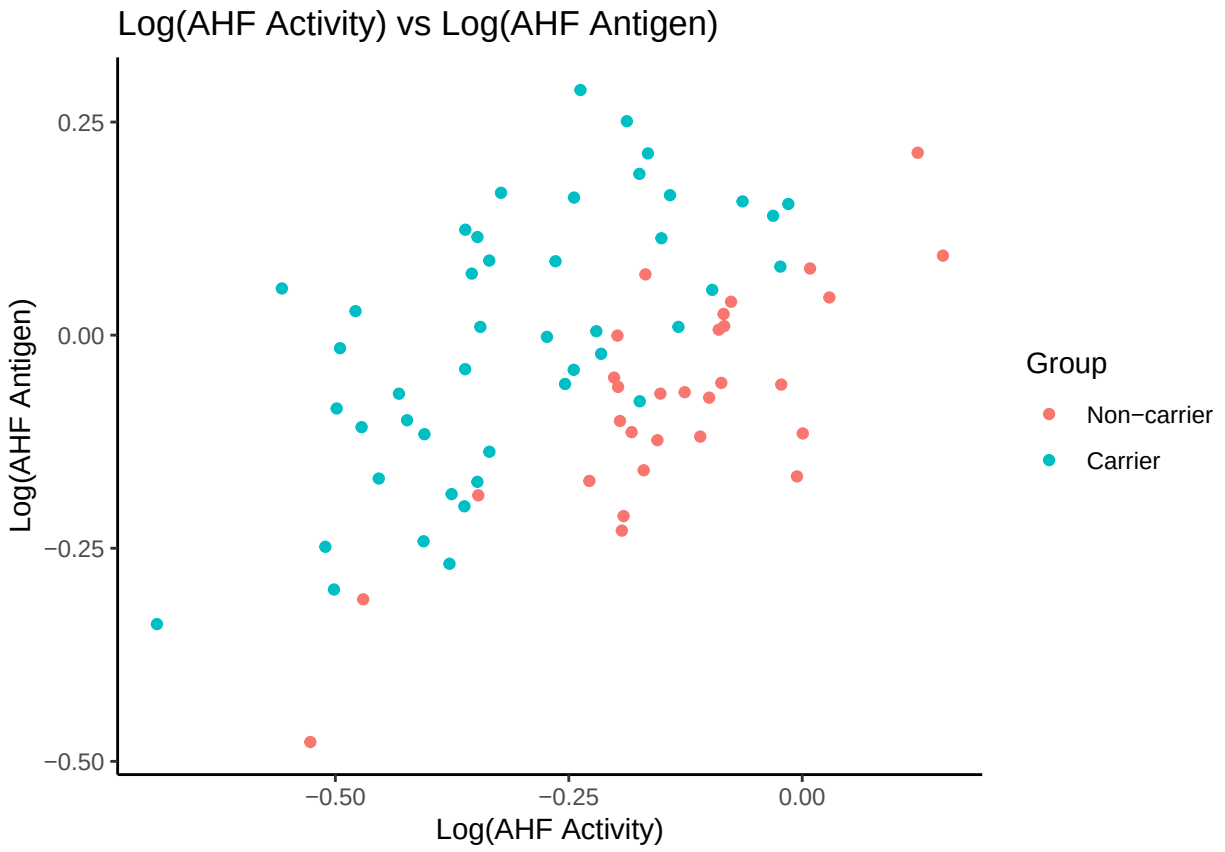
*Note:*
This table is the summary data for the Non-carrier group.

| | AHF_activity | AHF_antigen |
|---|---|---|
| | Min. :-0.6911 | Min. :-0.339000 |
| | 1st Qu.:-0.4055 | 1st Qu.:-0.107900 |
| | Median :-0.3352 | Median : 0.004600 |
| | Mean :-0.3079 | Mean :-0.005991 |
| | 3rd Qu.:-0.1878 | 3rd Qu.: 0.115100 |
| | Max. :-0.0149 | Max. : 0.287600 |

*Note:*
This table is the summary data for the Carrier group.


Log(AHF Activity) histogram plot
Counts
Log(AHF Activity)
−0.50
−0.25
0.00
0
2
4
6
Group
Non−carrier
Carrier


Log(AHF Antigen) histogram plot
Counts
Log(AHF Antigen)
−0.4
−0.2
0.0
0.2
0
2
4
6
Group
Non−carrier
Carrier




## DATA ANALYSIS:

**Partial code credit to Prof Li (UC Davis), Weiping Zhang(USTC)**

**Two sample test with:**

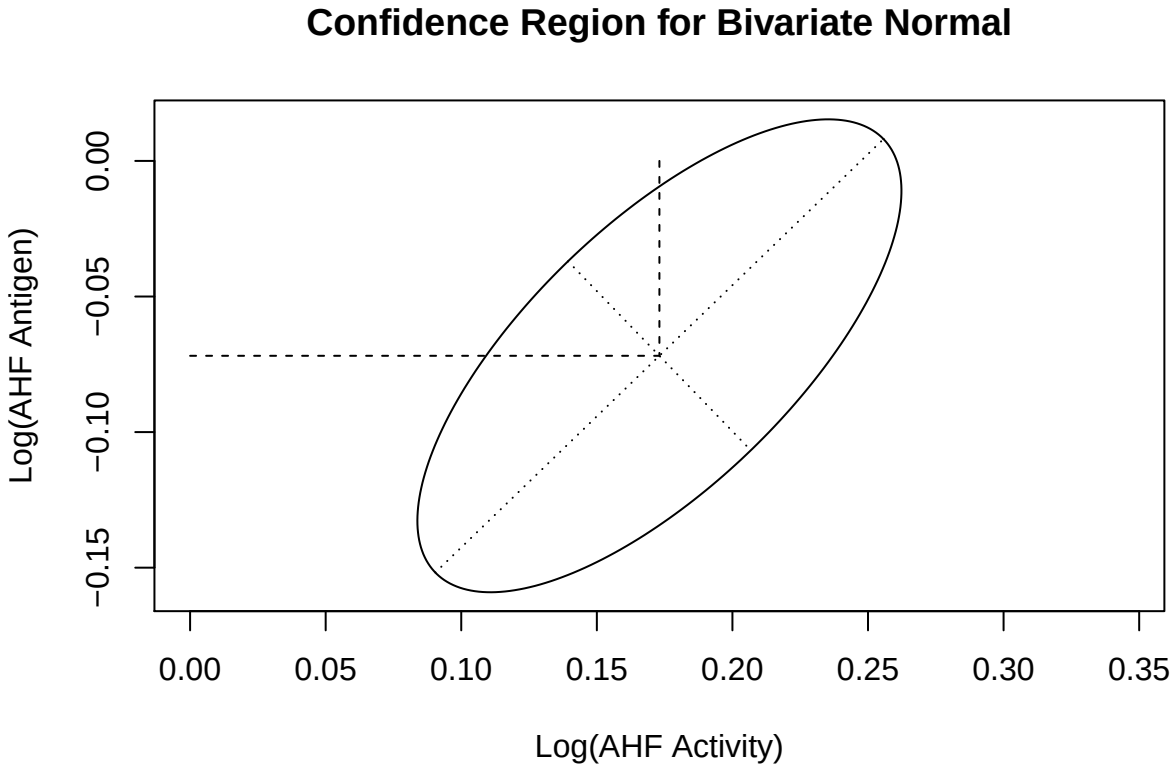
- Here we want to test the hypothesis that our two groups means (for Log(AHF activity) and Log(AHF antigen)) have a statistically significant difference from each other. The null hypothesis is that the two samples (carriers and non-carriers) are from populations with the same multivariate mean. The alternate hypothesis (if we reject the null hypothesis) is that the two samples are from different populations with different multivariate means.
- Since T_ˆ2 = 82.3375855871522 > 6.3345901039956 (the confidence value), the null hypothesis is rejected at 0.05 level of significance.

**Simultaneous confidence intervals**

This table is the confidence interval for the difference between the two group at alpha=0.05. This data shows our uncorrected siimultaneous confidence intervals for each of our values based on the difference between the two groups. This means that there is a significant difference between AHF acitivty between carriers and non-carriers since 0 is not included in the confidence interval. There is not a significant difference between AHF antigen levels betweeen carriers and non-carriers since 0 is not included in the confidence interval.

| | Lower.Bound | Upper.Bound |
|---|---|---|
| AHF_activity | 0.0838215 | 0.2623318 |
| AHF_antigen | -0.1590639 | 0.0153328 |

**We can now graph this confidence interval to better to see the confidence interval across the two axis to see which combinations of values would fall out of the confidence interval**

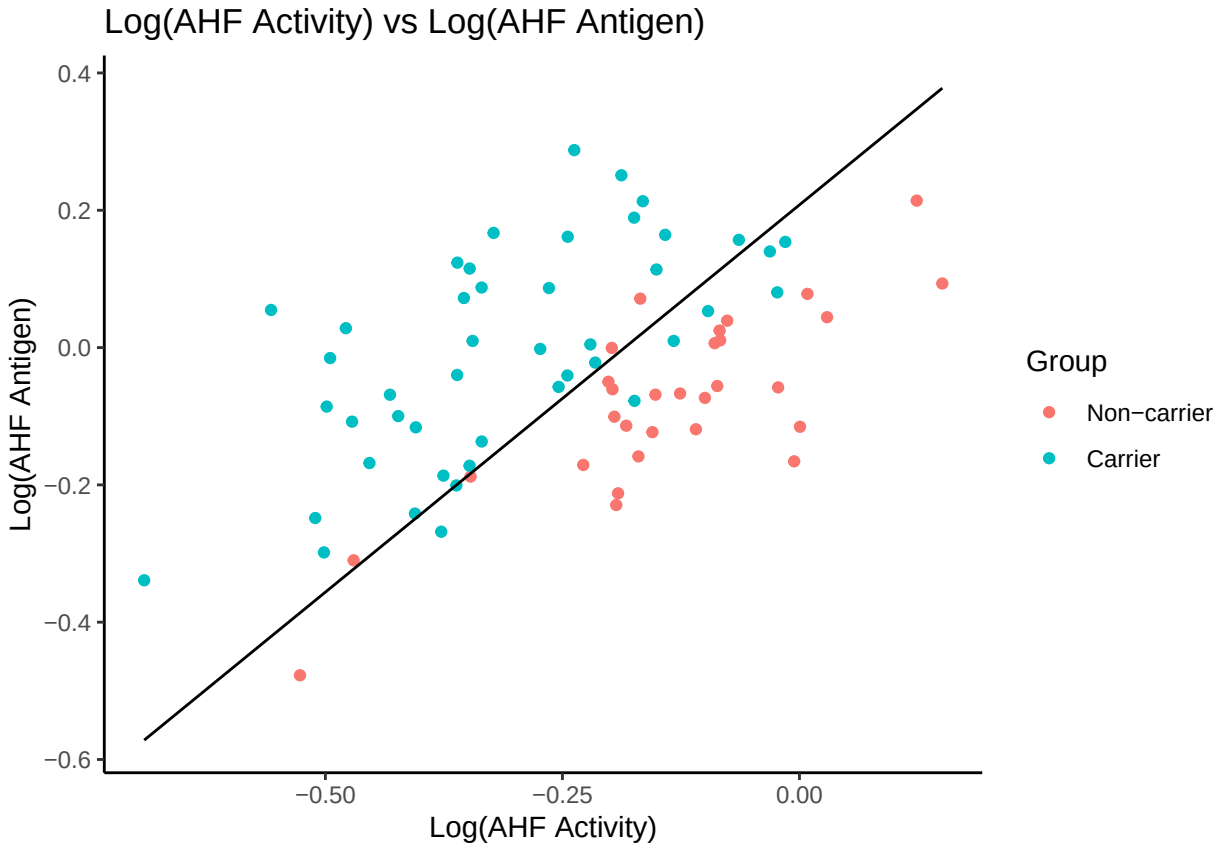



**Bonferroni simultaneous confidence intervals**

| | Lower.Bound | Upper.Bound |
|---|---|---|
| AHF_activity | 0.0919172 | 0.2542361 |
| AHF_antigen | -0.1511548 | 0.0074236 |

- This table is the confidence interval for the difference between the two group at alpha=0.05. This data shows our bonferronie corrected simultaneous confidence intervals for each of our values based on the difference between the two groups. This means that there is a significant difference between AHF acitivty between carriers and non-carriers since 0 is not included in the confidence interval. There is not a significant difference between AHF antigen levels betweeen carriers and non-carriers since 0 is not included in the confidence interval.

**Computing the LDA (linear discriminant analysis) this graphical representation will show us the boundary we will use as our cutoff when performing Fishers rule, based on the Mahalanobis distances.**




**Some general setup for fishers rule. We will construct Fisher's rule. Moreover, we will calculate theapparent error rate. This is the number of incorrectly categorized variables based on the comparison of the Mahalanobis distances.**

**Error rate for Apparent error rate:**

- 0.146666666666667

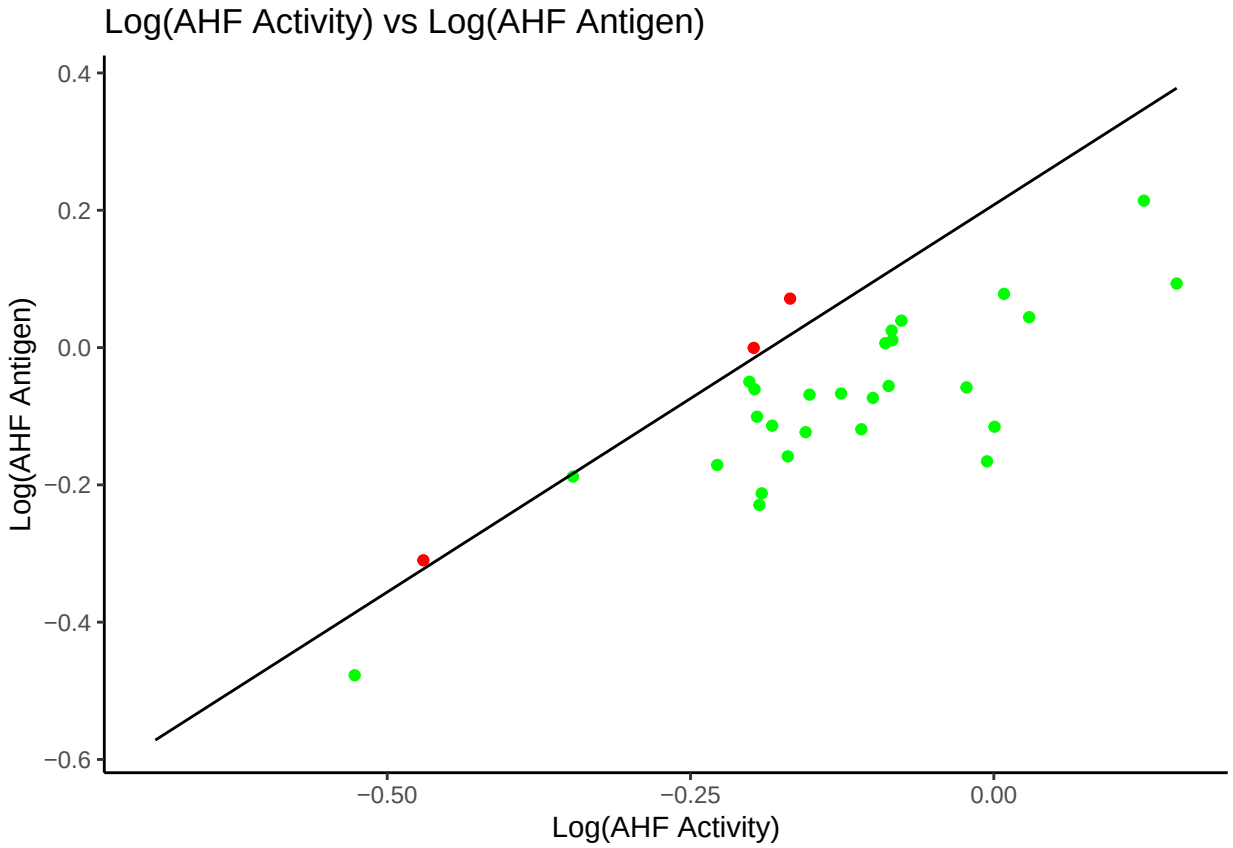
**Number incorrect non-carrier:**

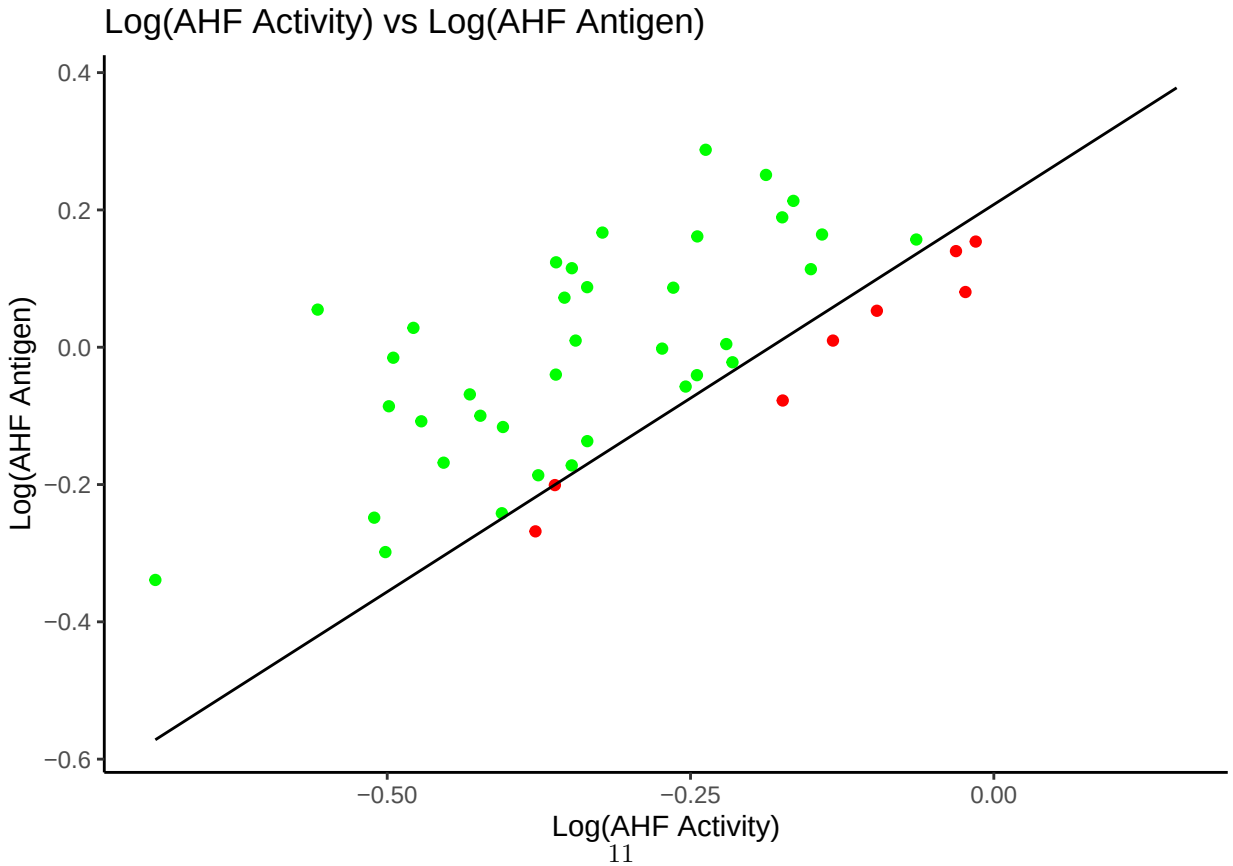
- 3

**Number incorrect carrier:**

- 8

**Lets overlay the ones that were incorrect for each group as a different color on our LDA graph just to confirm. This will allow us to compare where this error rate model compares to Lachenbruch's holdout.**



**Finally the expected error rate by Lachenbruch's holdout: this will help fix the fact that apparent error rate can often underestimate the error rate by witholding individual entries, observations in the data and then performing a LDA or Mahalanobis distance without this entry. This entry is then tested in the model as a new observation to see if it is correctly classified or not. As we can see from the graphical representation this can change our results around the border of the LDA.**

**Error rate for Apparent error rate:**

- 0.16

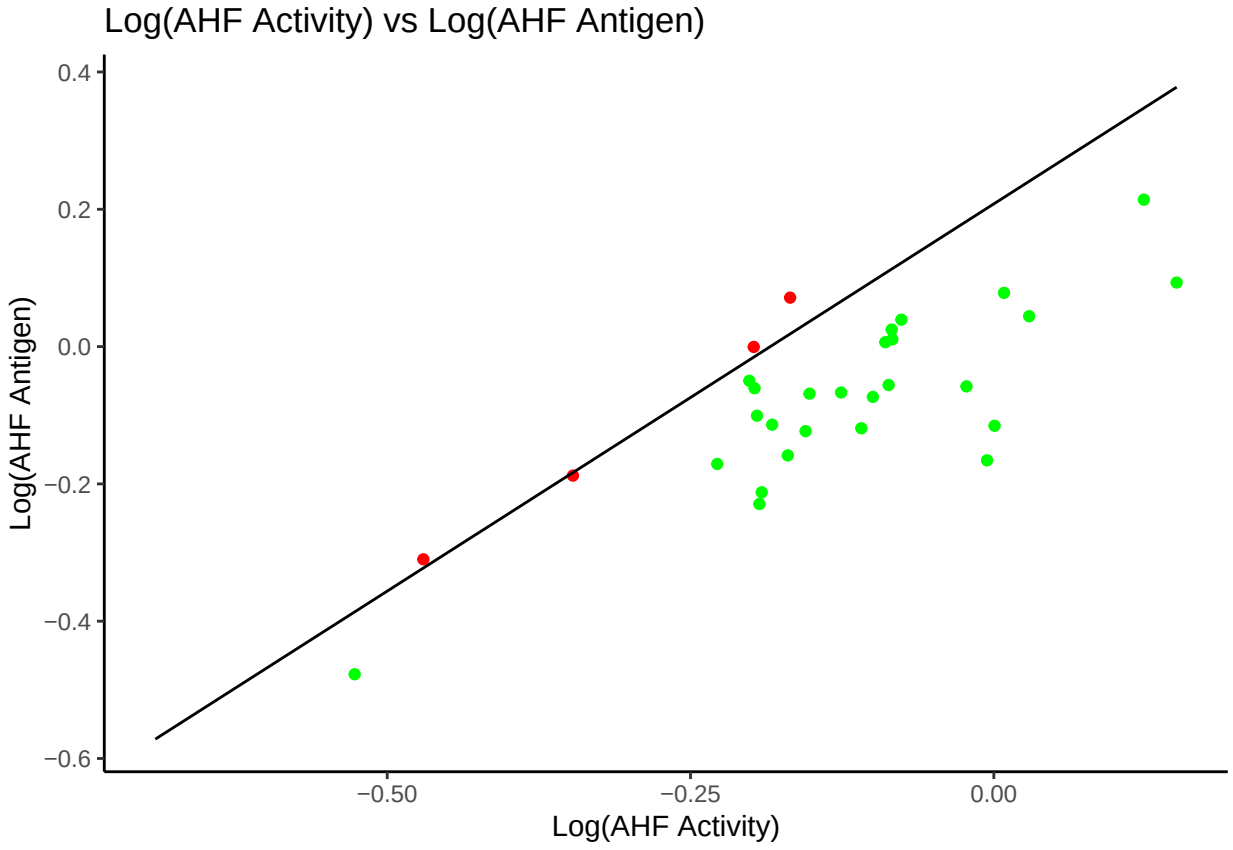
**Number incorrect non-carrier:**

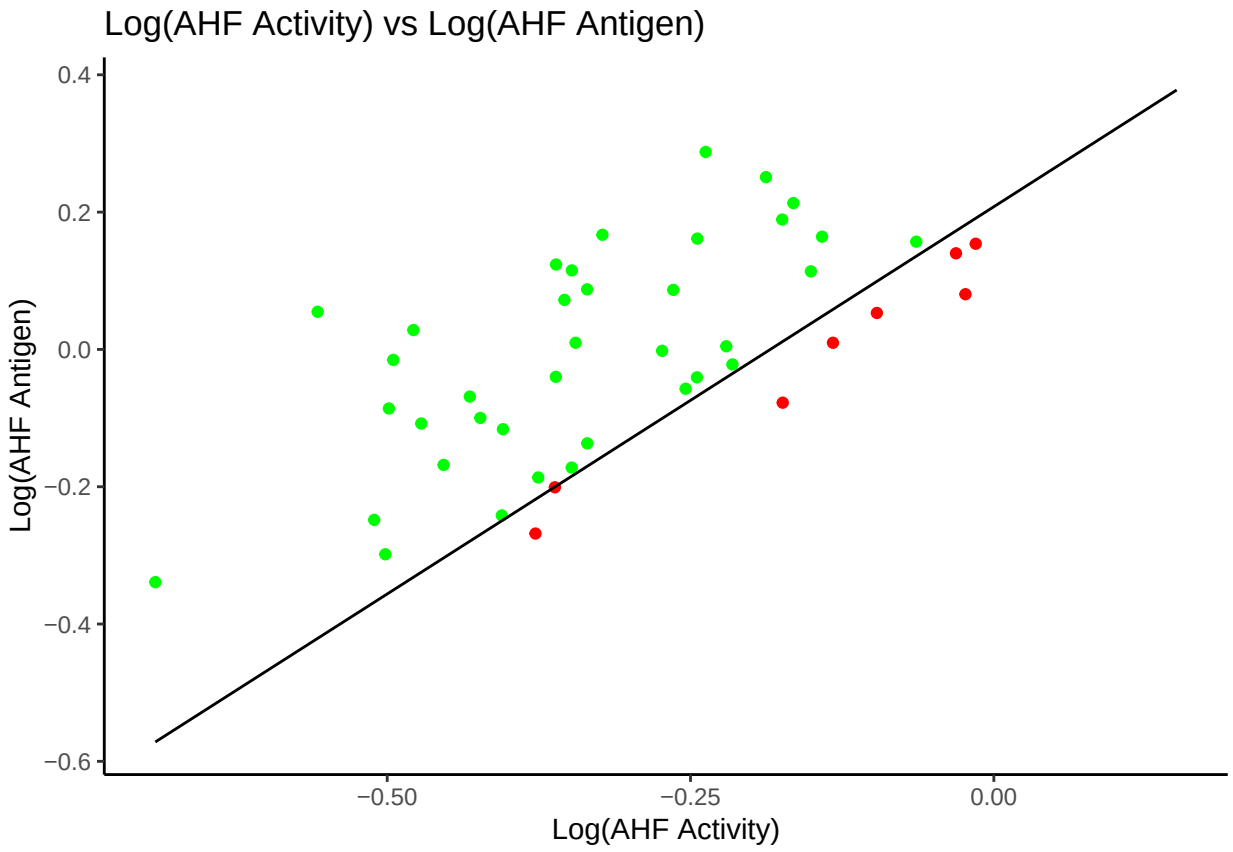
- 4

**Number incorrect carrier:**

- 8

**Lets overlay the ones that were incorrect for each group as a different color on our LDA graph just to confirm our numbers and visualize where Lachenbruch's holdout is changing the error rate in comparison to the APR.**




Log(AHF Activity) vs Log(AHF Antigen)
Log(AHF Antigen)
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.50
−0.25
0.00
Log(AHF Activity)

## CONCLUSION

**As a precaution we do suggest gather more data on this situation if possible to create a more robust model. A common method in machine learning, at least from personal experience, for assessing the error rate of a model is to build a model on a proportion of the data. This is different from the Lachenbruch's holdout in the fact that we may only use 50-70% of the data and then test the error rate of the model on the remaining 30-50%. Obviously this proportions can greatly vary but they require that a lot more data is collected.**

**The first goal of this analysis was to perform two sample test on the data in order to see if there is a significant difference between individuals that are carriers and noncarriers with respect to AHF activity ($log_{10}$ scale) and AHF antigen ($log_{10}$ scale). Group 1 is considered to be the non carrier group while group 2 is considered to be the obligatory carrier group. Using the Hotelling's $T^2$ test we see that at the 95% confidence intervale we we reject the null hypothesis, that the two samples are from the same populations with the same multivariate means.**

**The second goal of this analysis will be to perform linear discriminant analysis to try and predict the best possible distinguishable boundary between carriers and non carriers with respect to AHF activity ($log_{10}$ scale) and AHF antigen ($log_{10}$ scale). By performing Fisher's rule we were able to obtain an error rate of 0.16 for Lachenbruch's holdout. Overall the results from the LDA will allow physicians to diagnose someone with better confidence, especially if used in conjuction with other diagnostic tests. These sorts of tests are often used with Bayes to estimate the probability someone has the condition, given the prevalence in the population itself. This will increase certainty of classications, but as mentioned earlier it is always highly suggested to collect more data from a larger variety of circumstances as this data collected could not be representative of the population.**

---

# DATASET 3: PCA

## INTRODUCTION:

**Data has been gatherd on populations which is considered "Census-tract data". The variables that have been gathered are "Total Population (thousands)", "Professional Degree (%)", "Employed age over 16 (%)", "Government Employment (%)", "Median home value ($100,000s)".**

1) Census data can be very important for a variety of reasons, one of the most important/common ones is predicting voting outcomes.
2) Politicians may try to predict their popularity to certain populations by find the more common types of districts and try to gain popularity with one of those districts which would then hopefully have a similar effect on those other common areas.

**PCA (principle component analysis) is a popular way to explore datasets with multiple dimensions due to the fact that it is a diminsional reduction technique which allows exploration of the data in 2 dimensions (depending on how much of the variance can be summarized in those two dimensions). This is a great high level method to explore which groups have the most variation and potentially cluster together; therefore, this data can help those interested, such as politicians, in understanding their demographic.**

## SUMMARIZE DATA:

**Lets summarize the dataframe we are using for the problem, or the census variates for a set of observations.**

| | Total_Population(thousands) | Professional_Degree(%) | Employed_age_over_16(%) |
|---|---|---|---|
| | Min. :1.360 | Min. : 0.720 | Min. :49.50 |
| | 1st Qu.:3.120 | 1st Qu.: 1.670 | 1st Qu.:66.42 |
| | Median :4.720 | Median : 3.380 | Median :71.30 |
| | Mean :4.469 | Mean : 3.962 | Mean :71.42 |
| | 3rd Qu.:5.760 | 3rd Qu.: 4.830 | 3rd Qu.:77.33 |
| | Max. :9.210 | Max. :16.700 | Max. :86.54 |

| | Employed_age_over_16(%) | Government_Employment(%) | Median_home_value($100,000s) |
|---|---|---|---|
| | Min. :49.50 | Min. :16.30 | Min. :0.930 |
| | 1st Qu.:66.42 | 1st Qu.:20.60 | 1st Qu.:1.300 |
| | Median :71.30 | Median :24.40 | Median :1.490 |
| | Mean :71.42 | Mean :26.91 | Mean :1.636 |
| | 3rd Qu.:77.33 | 3rd Qu.:31.00 | 3rd Qu.:1.780 |
| | Max. :86.54 | Max. :68.50 | Max. :3.640 |

- This table is a general summary of the columns in our dataset. Given that the data has many columns, each with their own units, it is natrual to want to do a dimensional reduction to better understand the variance covariance structure.

## DATA ANALYSIS:

**First lets take a look at the eigen values and eigen vectors of the covariance matrix as these are important factors in the principle component analysis. In addtion, we can look at the correlation coeffecients between our principle components and our variates, which is another method of analyzing the contribution of each variate.**

| | | | | |
|---|---|---|---|---|
| 107.0153 | 39.67214 | 8.370866 | 2.867874 | 0.1546931 |

- This table is the list of eigen values, from largest to smallest, for the covariance of our dataset.

| | | | | |
|---|---|---|---|---|
| 0.0388873 | -0.0711449 | -0.1878926 | 0.9771352 | -0.0576999 |
| -0.1053220 | -0.1297524 | 0.9609958 | 0.1713518 | -0.1385541 |
| 0.4923639 | -0.8643881 | -0.0457974 | -0.0910437 | 0.0049660 |
| -0.8630699 | -0.4803318 | -0.1531854 | -0.0296858 | 0.0066918 |
| -0.0091223 | -0.0147434 | 0.1249811 | 0.0817012 | 0.9886375 |

- This table is the list of eigen vectors corresponding to the eigen values, from largest to smallest, for the covariance of our dataset.

| x |
|---|
| 0.1293466 |
| -0.3503211 |
| 0.6829211 |
| -0.9460440 |
| -0.1671804 |

- This table is the list of correlation coefecients corresponding to the first eigen value for the covariance of our dataset. Each of these values represents the amount that our variate correlates to the first principle component. The order of the values corresponds to the order of our data table columns mentioned above. This make sense because the 4th value is the largest, similar the 4th value being largest -0.8630699 for the loadings in the table above.

**Also, lets take a look at the eigen values and eigen vectors of the standardized matrix which is also sometimes a good way to summarize our data (can be preferrable if the primary eigen values are a larger proportion than that of the covariance matrix). As expected these values are the same since the loadings of the standardized matrix are the same as the correlations coeffecients of the standardized matrix.**

| | | | | |
|---|---|---|---|---|
| 1.991918 | 1.367527 | 0.8641573 | 0.535061 | 0.2413367 |

- This table is the list of the eigen values, from largest to smallest, for the covariance matrix of our standardized dataset.

| | | | | |
|---|---|---|---|---|
| 0.2625829 | 0.4629936 | -0.7839027 | 0.2169291 | 0.2347882 |
| -0.5933541 | 0.3256442 | 0.1640725 | -0.1446471 | 0.7028828 |
| 0.3256978 | 0.6051419 | 0.2248745 | -0.6628689 | -0.1943206 |
| -0.4792022 | -0.2524850 | -0.5507009 | -0.5716730 | -0.2766497 |
| -0.4932213 | 0.4996473 | 0.0688244 | 0.4072024 | -0.5801162 |

- This table is the list of eigen vectors corresponding to the eigen values, from largest to smallest, for the covariance matrix of our standardized dataset.

**Lets see if the covariance matrix or the correlation matrix summarizes the data in the first two components better.**

**Variance summarized in first two components of the covariance matrix:**

- 0.927926532382638

**Variance summarized in first two components of the covariance of standardized matrix:**

- 0.671888983734524

**So we see that using the covariance matrix is preferred here.**

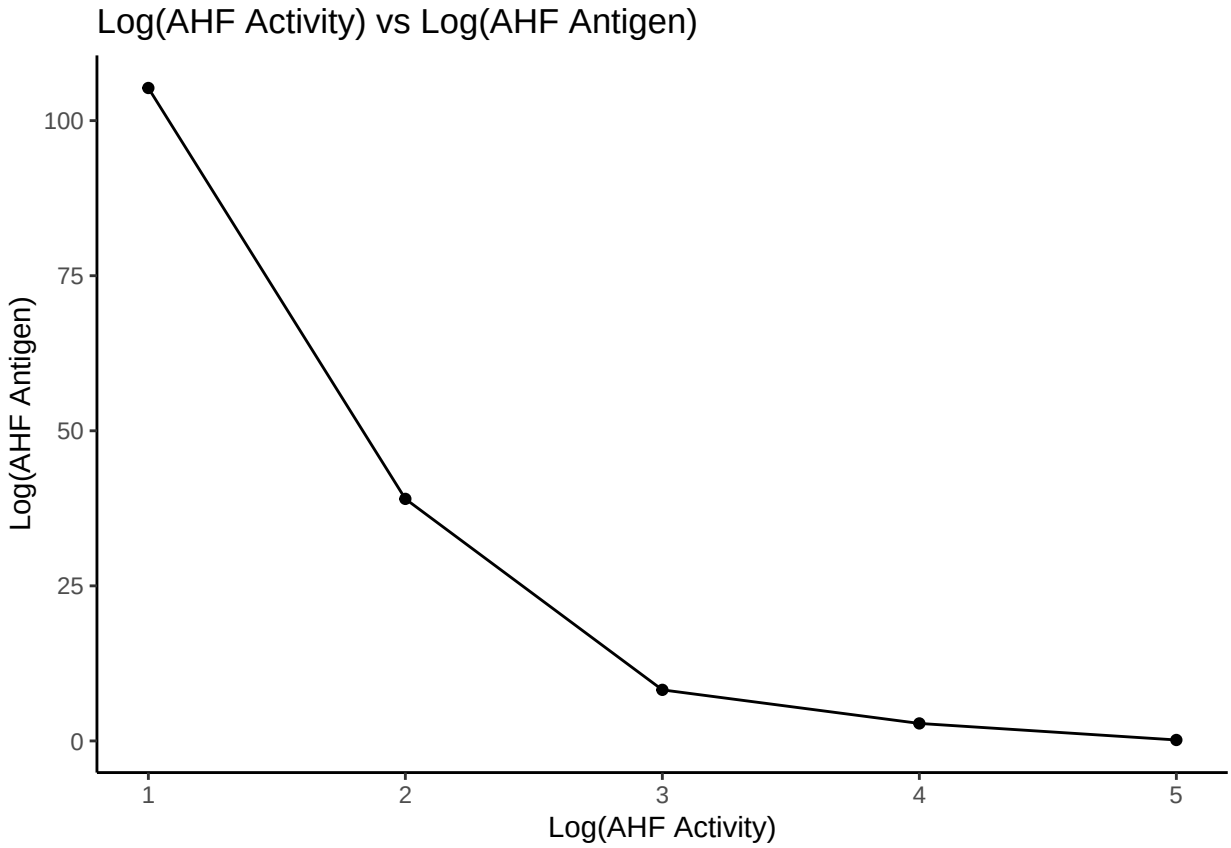
**So since we can summarize ~93% of the variance in our data using the first two principal components it would fair to graph our analysis using these two PCs. Lets also looks the the eigen value size graphed over the five components. Finally, we will overlay the PC scores for the sample data in the space of the first two principal components so we can visualize which of the sample data is most contributing to the the principal components which we see Government employment percentage and employed age over 16 (%) are the main two contributors to the**

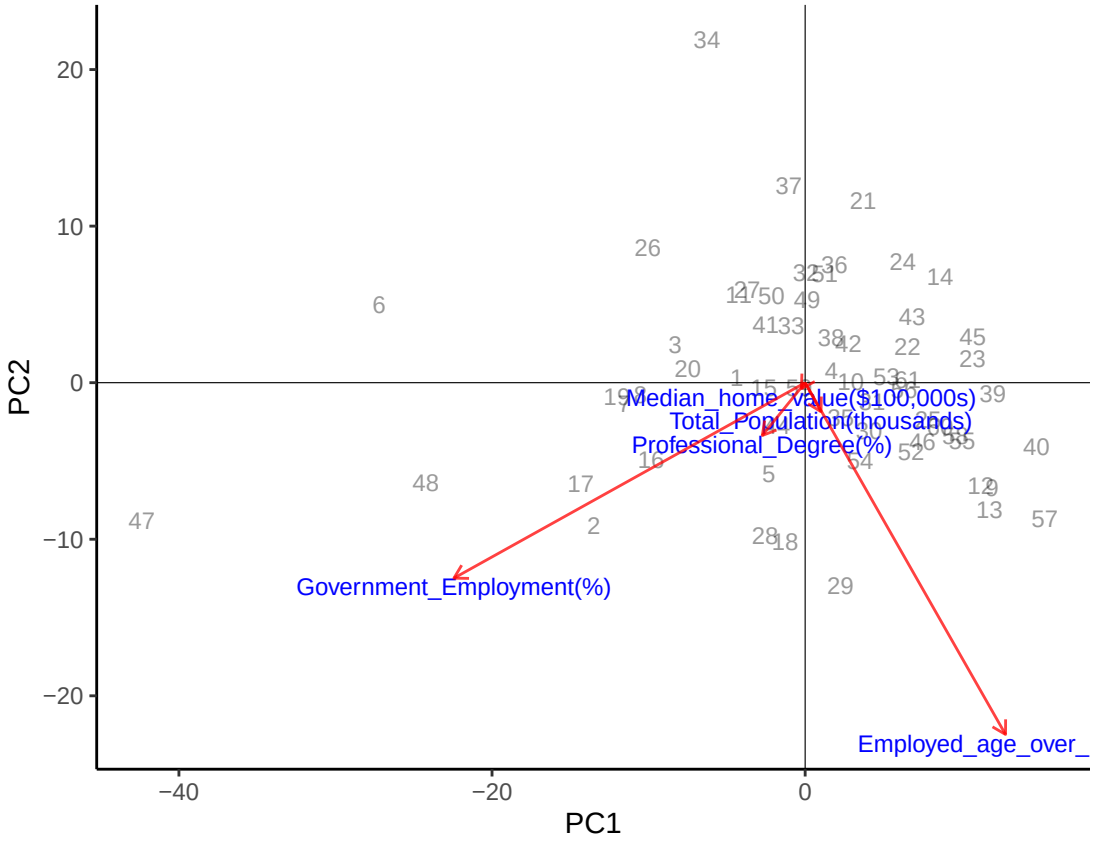
| | x |
|---|---|
| Comp.1 | 105.2609051 |
| Comp.2 | 39.0217730 |
| Comp.3 | 8.2336387 |
| Comp.4 | 2.8208596 |
| Comp.5 | 0.1521571 |

- This table is a list of the eigen values for the principle component analysis. As we can see these values match those calculated above manually for the components of the covariance matrix for the data; although here we used the princomp() funciton in R.

| | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 |
|---|---|---|---|---|---|
| Total_Population(thousands) | 0.0388873 | 0.0711449 | 0.1878926 | 0.9771352 | 0.0576999 |
| Professional_Degree(%) | -0.1053220 | 0.1297524 | -0.9609958 | 0.1713518 | 0.1385541 |
| Employed_age_over_16(%) | 0.4923639 | 0.8643881 | 0.0457974 | -0.0910437 | -0.0049660 |
| Government_Employment(%) | -0.8630699 | 0.4803318 | 0.1531854 | -0.0296858 | -0.0066918 |
| Median_home_value($100,000s) | -0.0091223 | 0.0147434 | -0.1249811 | 0.0817012 | -0.9886375 |

- This table is a list of the eigen vectors for the principle component analysis. As we can see these values match those calculated above manually for the components of the covariance matrix for the data; although here we used the princomp() funciton in R.


Log(AHF Activity) vs Log(AHF Antigen)
Log(AHF Antigen)
Log(AHF Activity)
100
75
50
25
0
1
2
3
4
5


PC2
PC1
20
10
0
−10
−20
−40
−20
0
Median_home_value($100,000s)
Total_Population(thousands)
Professional_Degree(%)
Government_Employment(%)
Employed_age_over_

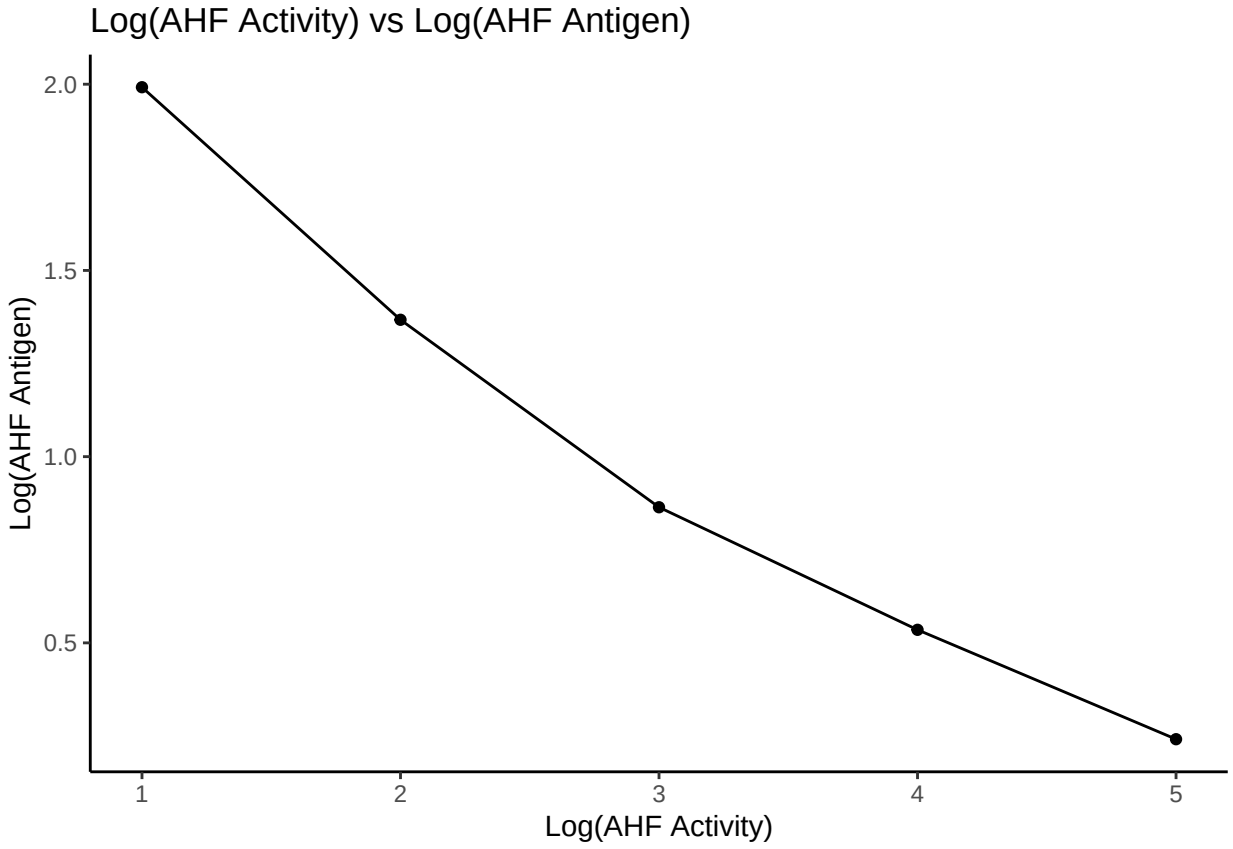
**Now the scaled or standarized dataset**

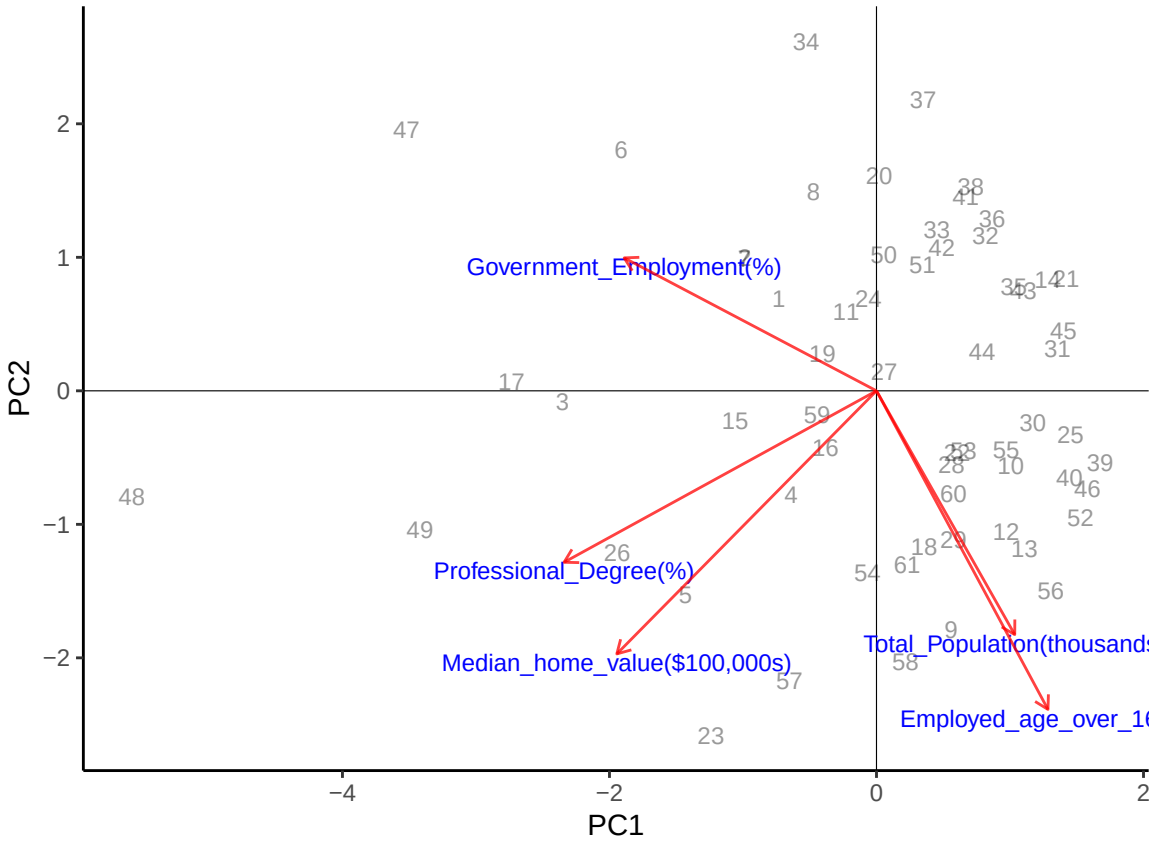
| x |
|---|
| 1.9919183 |
| 1.3675266 |
| 0.8641573 |
| 0.5350610 |
| 0.2413367 |

- This table is a list of the eigen values for the principle component analysis of the scaled or standardized dataset. As we can see these values match those calculated above manually for the components of the covariance matrix for the data; although here we used the princomp() funciton in R.

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| Total_Population(thousands) | 0.2625829 | -0.4629936 | 0.7839027 | -0.2169291 | 0.2347882 |
| Professional_Degree(%) | -0.5933541 | -0.3256442 | -0.1640725 | 0.1446471 | 0.7028828 |
| Employed_age_over_16(%) | 0.3256978 | -0.6051419 | -0.2248745 | 0.6628689 | -0.1943206 |
| Government_Employment(%) | -0.4792022 | 0.2524850 | 0.5507009 | 0.5716730 | -0.2766497 |
| Median_home_value($100,000s) | -0.4932213 | -0.4996473 | -0.0688244 | -0.4072024 | -0.5801162 |

- This table is a list of the eigen vectors for the principle component analysis of the scaled or standardized dataset. As we can see these values match those calculated above manually for the components of the covariance matrix for the data; although here we used the princomp() funciton in R.






## CONCLUSION

**As a precaution we do suggest gather more data on this situation if possible. This will increase the strength of the analysis and account for other variables in the situation. This would be achieved by increasing the value of n, or observed situaitons (census data in new areas), and r, the number of observed varaibles upon which to build PCs (such as "Total_Population(thousands)", "Professional_Degree(%)", and "Employed_age_over_16(%)" etc). Please contact if this is the case and a more robust script will be made to factor in variations in new potential data gathering.**

**From the results of the principle component analysis we are able to determine that we can summarize %92 of the variance in our dataset in the first princple component. We also are able to determine that two variates that contribute most to the variance in the population are Government Employment and Age over 16 employement rates. This type of information may be of value to a variety of people who may try to better understand the demographic. Individuals can better influence these regions based on things they have in common (home value, total population, professional degrees etc.), but they can also target other sectors more specifically based on government employment and individuals employed over the age of 16.**