Justin Beer

3/15/2017

MSDS 6372

**Interferon Flu Data: ANOVA, PCA, and MLR Analysis**

Even in today’s modern world we are surrounded with biological contaminants that pose a significant danger to the human species. Chief among these is the influenza virus which is not only a discomforting infection but can also be a deadly one. The CDC estimates that anywhere between 3,300 and 49,000 people every year die from influenza (Appendix Item 1). The virus is considered highly contagious and infects the respiratory passages. The symptoms of infections are fever, sever aching, and catarrh (discharge).

For this study we are analyzing the influence the IFITM3 gene may have on combating the influenza virus. We are using data that was collected on 17 subjects after they were injected with the virus. We recorded whether or not the subject is symptomatic or asymptomatic, showing signs of infection or not showing signs, respectively. The subjects underwent repeated measurements during their illness. We also have the gender and age for each subject.

The IFITM3 gene is thought to encode a specific type of protein that provides immunity to the influenza virus. Our broader population are all humans that have the IFITM3 gene. We will set out to identify whether or not this is the case. This data set was obtained from Professor Turner at Southern Methodist University.

**Problem Statement**

We need to identify whether or not there is a difference in the symptomatic and asymptomatic groups when compared to their IFITM3 gene values. Then we will search for a multiple linear regression model to describe the IFITM3 gene response.

**Limitations and Assumptions**

We are assuming the following about this data set:

1. The data set is a large enough sample size with 14 repeated samples per subject and 252 total observations.
2. Residuals will be normally distributed.
3. The standard deviation is constant.

This data set is observational. Every subject received an influenza virus injection. The subjects were not randomized for who would or would not receive the injections. Due to this we cannot draw cause and effect conclusions that fit to the broader population of humans. However, we can draw inferences that apply within this observed group of subjects. This will assist us in guiding further research questions such as “Is it worthwhile to continue to study the IFITM3 gene for possible influenza immunities?”

Also, since this data set is observational there may be confounding factors that contribute to the results. One such confounding factor is the status of the subject’s health at the beginning of the study. Did the subject just get over an infection thereby causing a better immune system response to the influenza virus? Does the subject possess some other gene or mechanism that provides immunity?

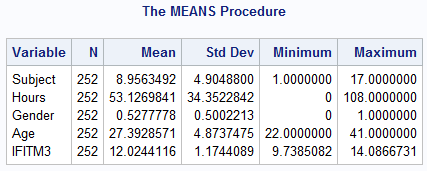
**Data Set Description**

Our response variable is the IFITM3 variable. The age, gender, and time series are all explanatory variables.

* **Age** – the age of the subject.
* **Gender** – 0 or 1 for each subject (0 = Male, 1 = Female).
* **Subject** – a subject ID number.
* **Response** – Either Symp (showing signs of symptoms) or Asymp (no sign of symptoms).
* **Hours** – The time series aspect. Represents the hour after influenza injection. Can be 0, 5, 12, 21, 29, 36, 45, 53, 60, 69, 77, 84, 93, 101, or 108.
* **IFITM3** – A measure of the gene’s response, creation of its protein, to the influenza virus.

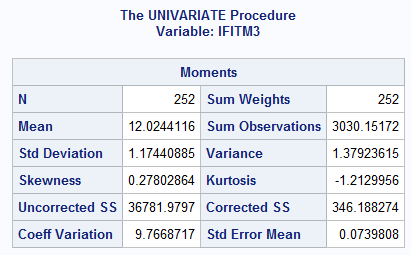
**Descriptive Statistics**

The below Figure 1 shows descriptive statistics for our variables, excluding our categorical variable Response.



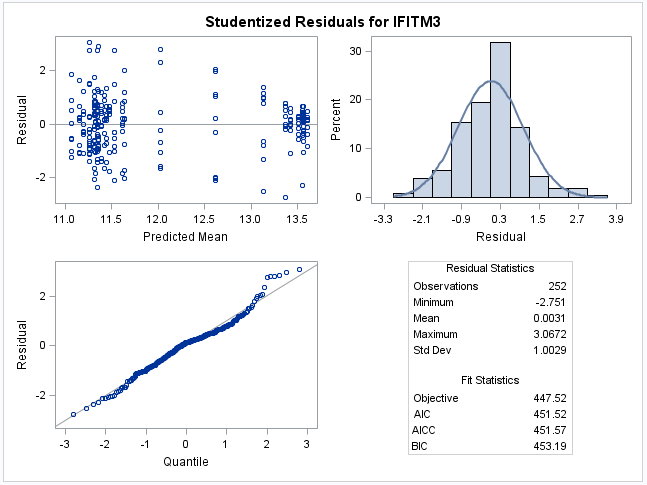
**Figure 1: Descriptive Statistics all Variables**

Figure 2 below shows us the descriptive statistics for just our response variable IFITM3.



**Figure 2: Descriptive Statistics for IFITM3**

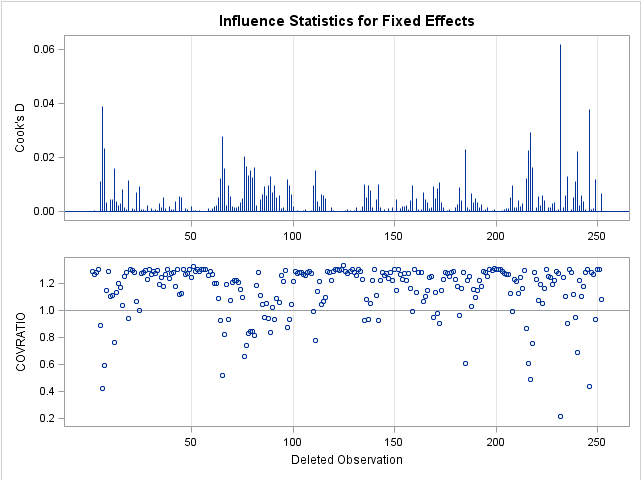
We can see that if there is going to be a significant interaction between IFITM3 and the Response then the Response categories will diverge from each other around IFITM3’s mean of 12. Let’s take a look at the residuals for IFITM3 to ensure they are normally distributed and meeting our assumptions.



**Figure 3: Residuals for IFITM3**

From Figure 3, the IFITM3 residuals appear normally distributed. In the upper left residual plot, there appears to be two clusters. These clusters will most likely correspond with the two groups that we are interested in, Response = Symptomatic or Asymptomatic. Seeing that they are clustered here, I will guess that we will see a difference in the Response categories means.

Figure 4 below shows us the Cook’s D and COVRATIO for IFITM3. We can see that all Cook’s D values are below 0.06. Also, the COVRATIO shows us that there is significant covariance, any values above the 1.03 line. Due to the flexibility of proc mixed to model the variance and covariance this will not be an issue.



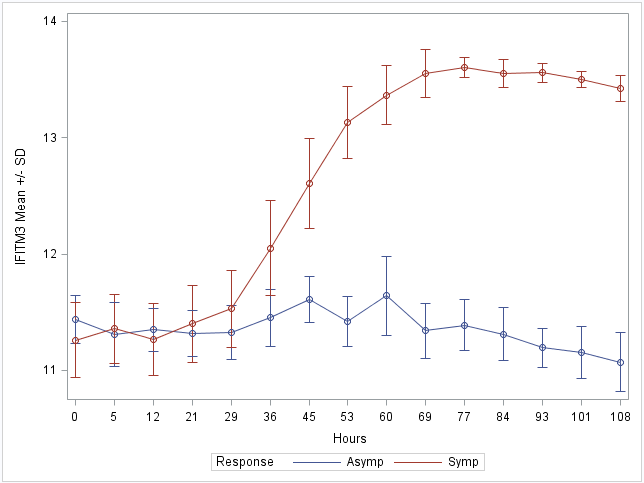
**Figure 4: Cook’s D and COVRATIO for IFITM3**

**Initial Model Analysis**

For this study we already know the initial model we want to investigate. Firstly, we will investigate the initial model. The linear regression model of interest is:

Our hypothesis testing for this model is that the null hypothesis is that the Symptomatic and Asymptomatic means will be equal for all time points. Our alternative hypothesis is that their means will be different for at least one time point.

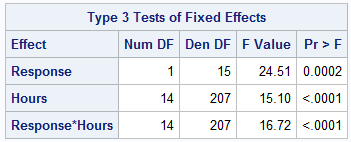
Figure 5 is a plot of the mean for each hour for all subjects broken down into the two Response groups, Symptomatic or Asymptomatic. We want to know whether the subject showing signs of the infection and therefore fighting it is related to higher IFITM3 values. A higher IFITM3 value means more of the protein this gene creates is being used to fight the virus.



**Figure 5: Mean +/- SD bars per Hour by Response**

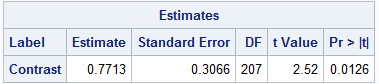
We can see from Figure 5 that the two groups, Symptomatic and Asymptomatic, do indeed have different means and this starts at hour 45 (where their SD’s no longer touch). Interestingly, the IFITM3 mean = 12 is just about the point where the two group means are separated.

Figure 6 below, shows us the results of our hypothesis tests with α = 0.05. For the Response, p-value = 0.0002. For Hours, p-value < 0.0001. For Response\*Hours, p-value < 0.0001. We interpret these results to be that the effects we included in our initial model (Response, Hours, Response\*Hours) are all statistically significant and not equal to zero. Therefore, we should keep all of them in our model.



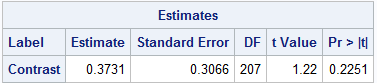
**Figure 6: Response, Hours, and Interaction Term Tests**

Now we will move onto investigating the interaction term Response\*Hours. It may be that not all combinations of Response\*Hours will be statistically significant in explaining IFITM3. Also, it does not make sense to compare current hours to future hours such as Asymptomatic Hour 5 to Symptomatic Hour 60. We will start off by contrasting the two groups between Hours = 36 and 45.



**Figure 7: Contrast Hours 36 and 45 by Response**

We see from Figure 7, which shows our contrast results, that there is a statistically significant difference between these points, with p-value = 0.0126. Now let us look at Hours 29 and 36.



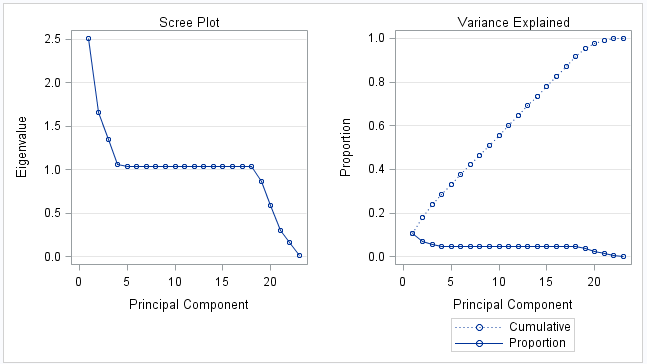
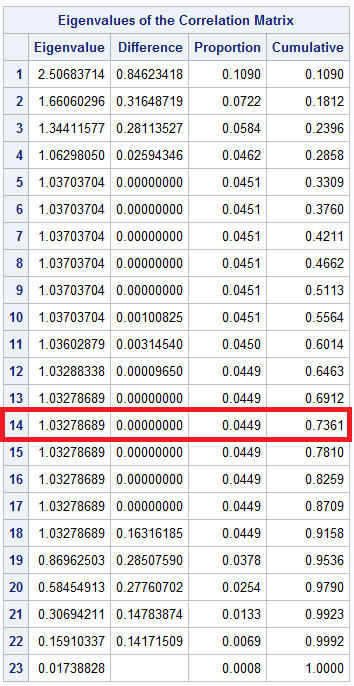
**Figure 8: Contrast Hours 29 and 36 by Response**

Figure 8 is the resulting contrast which unsurprisingly shows that there is not a statistically significant difference between the group means at Hours 29 and 36, p-value = 0.2251. If we first looked at Figure 5, we would expect to encounter the results that we did for these contrasts.

Thus far, we have established that there is a difference in the means of the Asymptomatic and Symptomatic groups starting at Hour 45. This leaves us with 18 interaction terms to account for (9 hours \* 2 groupings) along with the Gender, Age, Hours, and Response variables.

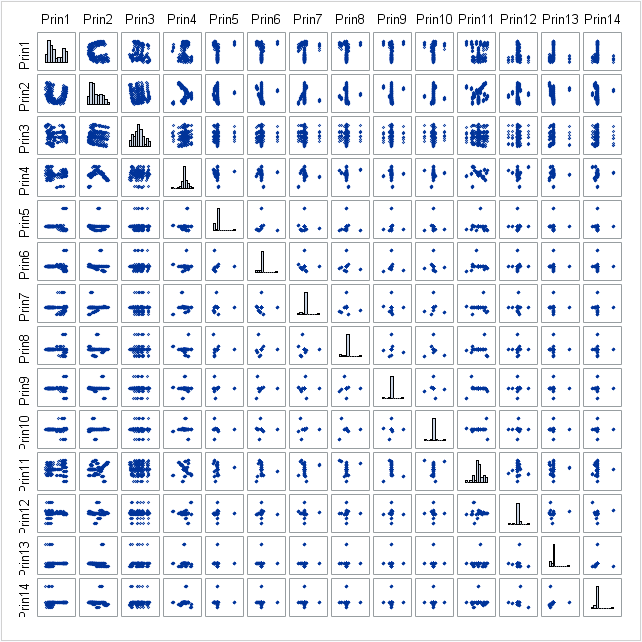
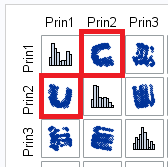
**Principal Component Analysis**

Thus far we have established the validity of our initial model and that there is a difference in means of the two groups after hour 36. We will now employ Principal Component Analysis to simplify our model. The target for the variance explained will be 70%. Before PCA, we needed to change the Response to a value variable. The value of 1 was given to the Symptomatic group and the value of 0 to the Asymptomatic group. Figure 9 shows that we will need 14 principal components in order to describe 73.61% of our data set.

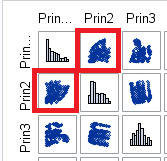
 

**Figure 9: Scree Plot, Variance Explained, Eigenvalue Correlation Matrix**

Now we will take the 14 principal components and use them as our variables to select from with PROC GLMSELECT. Let’s take a look at the scatter plot matrix shown in Figure 10 for the 14 principal components. The only component to note is Prin1 needing to be a squared term to account for its shape.

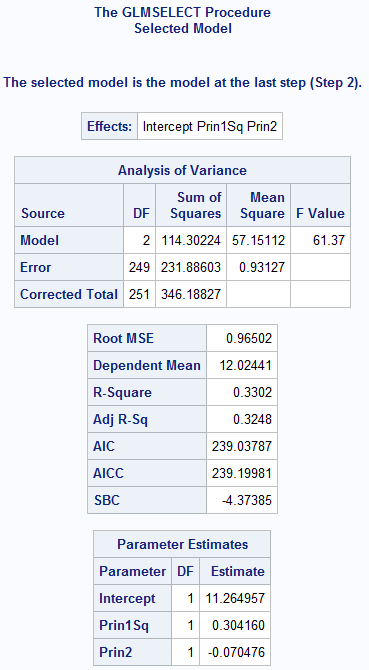
 

**Figure 10: Principal component Scatter Plot**



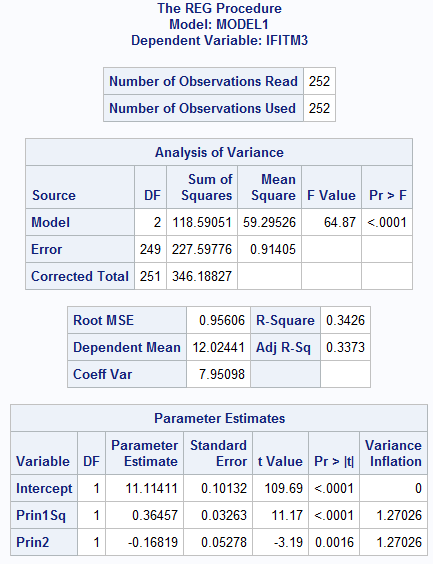
**Figure 11: Squaring Prin1**

After squaring Prin1 we now arrive at our final model for predicting our response IFITM3. The model below has 3 terms instead of the 24 terms.

**Figure 12: Principal Components in Final Model and Final Model**

Running our model through PROC REG provides us with the models p-value < 0.0001 (for α = 0.05), Adjusted R2 of 0.3373, and VIF’s = 1.27.



**Figure 13: Final Model Statistics**

**Conclusion**

We have arrived at a Multiple Linear Regression model that is statistically significant, p-value < 0.0001 (F-Value = 64.87), and relatively simple with only 2 variables and the intercept. Our model explains 33.73% (Adj. R2) of the variation of our response IFITM3.

We found that there is a statistically significant difference between the Symptomatic and Asymptomatic groups after Hour 36. We also were able to get our final model containing three components instead of twenty four.

The Symptomatic group seems to be creating a greater amount of the protein that the gene IFITM3 generates. However, this being an observational study we cannot draw any cause-effect conclusions. There seems to be potential in the IFITM3 gene in creating a greater immune system response to the influenza virus but too many confounding factors exist. Therefore, better experiments are needed to isolate cause-effect relationships between the influenza virus and the IFITM3 gene.

**Appendix:**

**Item 1:** <http://www.npr.org/sections/health-shots/2010/08/26/129456941/annual-flu-death-average-fluctuates-depending-on-how-you-slice-it>

**PROC MIXED Reference:** <https://support.sas.com/documentation/cdl/en/statugmixed/61807/PDF/default/statugmixed.pdf>

**data** project2raw;

infile '\\Client\C$\SASDATA\InterferonFluData.csv' dlm=',' firstobs=**2**;

input Subject Hours Gender Age Response $ IFITM3;

**run**;

**proc** **print** data = project2raw; **run**;

/\* Descriptive Statistics - Means etc. > distributions / boxplots by Response \*/

**proc** **means** data = project2raw;

**run**;

/\* Scatter Plot Matrix \*/

**proc** **sgscatter** data = project2raw;

matrix Subject Hours IFITM3 Age Gender / diagonal=(histogram) group = Response;

**run**;

/\* Need the fit diagnostics and residuals for IFITM3\*/

ods graphics on;

**proc** **mixed** data = project2raw;

class Response Hours;

model IFITM3 = Response|Hours / residual influence;

repeated / type=un subject=Subject r;

LSMeans Response / PDiff;

**run**; **quit**;

ods graphics off;

/\* Creating IFITM3 Mean Time Series Plot \*/

**proc** **sort** data = project2raw;

by Response Hours;

**run**;

**proc** **summary** data = project2raw;

by Response Hours;

var IFITM3;

output out = means (drop = \_:) mean = mean n = n stderr = stderr;

**run**;

**data** means;

set means;

label mean = "IFITM3 Mean +/- SD";

elower = mean - stderr;

eupper = mean + stderr;

**run**;

**proc** **sgplot** data = means;

xaxis type = discrete;

series x=Hours y=mean / group = Response;

scatter x=Hours y = mean / group = Response yerrorlower = elower yerrorupper = eupper;

**run**; **quit**;

/\* Interaction Term Investigation \*/

ods graphics on;

**proc** **mixed** data = project2raw;

class Response Hours Subject;

model IFITM3 = Response Hours Response\*Hours / residual influence;

repeated Hours/ type=CS subject=Subject;

estimate 'Contrast'

Response

-**1** /\* Asymptomatic \*/

**1** /\* Symptomatic \*/

Response\*Hours

**0** **0** **0** **0** -**.5** -**.5** **0** **0** **0** **0** **0** **0** **0** **0** **0** /\* Asymptomatic\*hour = 0, 5, 12, 21, 29, 36, 45, 53, 60, 69, 77, 84, 93, 101, 108 \*/

**0** **0** **0** **0** **0.5** **0.5** **0** **0** **0** **0** **0** **0** **0** **0** **0** /\* Symptomatic\*hour = 0, 5, 12, 21, 29, 36, 45, 53, 60, 69, 77, 84, 93, 101, 108 \*/;

**run**; **quit**;

ods graphics off;

/\* Creating relevant interaction terms as variables \*/

**data** project2raw2;

set project2raw;

A0 = **.**;

A5 = **.**;

A12 = **.**;

A21 = **.**;

A29 = **.**;

A36 = **.**;

A45 = **.**;

A53 = **.**;

A60 = **.**;

A69 = **.**;

A77 = **.**;

A84 = **.**;

A93 = **.**;

A101 = **.**;

A108 = **.**;

if (Hours=**0** & Response = 'Asymp') then A0 = **1**;

else A0 = **0**;

if (Hours=**5** & Response = 'Asymp') then A5 = **1**;

else A5 = **0**;

if (Hours=**12** & Response = 'Asymp') then A12 = **1**;

else A12 = **0**;

if (Hours=**21** & Response = 'Asymp') then A21 = **1**;

else A21 = **0**;

if (Hours=**29** & Response = 'Asymp') then A29 = **1**;

else A29 = **0**;

if (Hours=**36** & Response = 'Asymp') then A36 = **1**;

else A36 = **0**;

if (Hours=**45** & Response = 'Asymp') then A45 = **1**;

else A45 = **0**;

if (Hours=**53** & Response = 'Asymp') then A53 = **1**;

else A53 = **0**;

if (Hours=**60** & Response = 'Asymp') then A60 = **1**;

else A60 = **0**;

if (Hours=**69** & Response = 'Asymp') then A69 = **1**;

else A69 = **0**;

if (Hours=**77** & Response = 'Asymp') then A77 = **1**;

else A77 = **0**;

if (Hours=**84** & Response = 'Asymp') then A84 = **1**;

else A84 = **0**;

if (Hours=**93** & Response = 'Asymp') then A93 = **1**;

else A93 = **0**;

if (Hours=**101** & Response = 'Asymp') then A101 = **1**;

else A101 = **0**;

if (Hours=**108** & Response = 'Asymp') then A108 = **1**;

else A108 = **0**;

S0 = **.**;

S5 = **.**;

S12 = **.**;

S21 = **.**;

S29 = **.**;

S36 = **.**;

S45 = **.**;

S53 = **.**;

S60 = **.**;

S69 = **.**;

S77 = **.**;

S84 = **.**;

S93 = **.**;

S101 = **.**;

S108 = **.**;

if (Hours=**0** & Response = 'Symp') then S0 = **1**;

else S0 = **0**;

if (Hours=**5** & Response = 'Symp') then S5 = **1**;

else S5 = **0**;

if (Hours=**12** & Response = 'Symp') then S12 = **1**;

else S12 = **0**;

if (Hours=**21** & Response = 'Symp') then S21 = **1**;

else S21 = **0**;

if (Hours=**29** & Response = 'Symp') then S29 = **1**;

else S29 = **0**;

if (Hours=**36** & Response = 'Symp') then S36 = **1**;

else S36 = **0**;

if (Hours=**45** & Response = 'Symp') then S45 = **1**;

else S45 = **0**;

if (Hours=**53** & Response = 'Symp') then S53 = **1**;

else S53 = **0**;

if (Hours=**60** & Response = 'Symp') then S60 = **1**;

else S60 = **0**;

if (Hours=**69** & Response = 'Symp') then S69 = **1**;

else S69 = **0**;

if (Hours=**77** & Response = 'Symp') then S77 = **1**;

else S77 = **0**;

if (Hours=**84** & Response = 'Symp') then S84 = **1**;

else S84 = **0**;

if (Hours=**93** & Response = 'Symp') then S93 = **1**;

else S93 = **0**;

if (Hours=**101** & Response = 'Symp') then S101 = **1**;

else S101 = **0**;

if (Hours=**108** & Response = 'Symp') then S108 = **1**;

else S108 = **0**;

**run**;

**proc** **print** data = project2raw2; **run**;

/\* MLR \*/

/\* Changes Response from categorical to value variables \*/

**data** project2raw3;

set project2raw2;

Response2 = **.**;

if (Response = 'Asymp') then Response2 = **0**;

else Response2 = **1**;

**run**;

**proc** **print** data = project2raw3; **run**;

/\* Principal Component Analysis \*/

ods graphics on;

**proc** **princomp** plots=all data=project2raw3 out=pca;

var IFITM3 Age Gender Hours Response2 S45 S53 S60 S69 S77 S84 S93 S101 S108 A45 A53 A60 A69 A77 A84 A93 A101 A108;

**run**;

ods graphics off;

**proc** **print** data=pca; **run**;

**proc** **sgscatter** data=pca;

matrix Prin1 Prin2 Prin3 Prin4 Prin5 Prin6 Prin7 Prin8 Prin9 Prin10 Prin11 Prin12 Prin13 Prin14 / diagonal=(histogram);

**run**;

**proc** **glmselect** data = pca seed=**1** plots(stepAxis=number)=(criterionPanel ASEPlot CRITERIONPANEL);

model IFITM3 = Prin1 Prin2 Prin3 Prin4 Prin5 Prin6 Prin7 Prin8 Prin9 Prin10 Prin11 Prin12 Prin13 Prin14 / selection=LASSO; **run**; **quit**;

**data** pca2;

set pca;

Prin1Sq = Prin1\*Prin1;

**run**;

**proc** **print** data = pca2; **run**;

**proc** **sgscatter** data=pca2;

matrix Prin1Sq Prin2 Prin3 Prin4 Prin5 Prin6 Prin7 Prin8 Prin9 Prin10 Prin11 Prin12 Prin13 Prin14 / diagonal=(histogram);

**run**;

**proc** **glmselect** data = pca2 seed=**1** plots(stepAxis=number)=(criterionPanel ASEPlot CRITERIONPANEL);

model IFITM3 = Prin1Sq Prin2 Prin3 Prin4 Prin5 Prin6 Prin7 Prin8 Prin9 Prin10 Prin11 Prin12 Prin13 Prin14 / selection=LASSO; **run**; **quit**;

**proc** **reg** data = pca2 plots(label) = (rstudentbyleverage cooksd);

model IFITM3 = Prin1Sq Prin2 / AIC VIF CLI; \*CORRB INFLUENCE CLB;

**run**; **quit**;