HW Unit 9

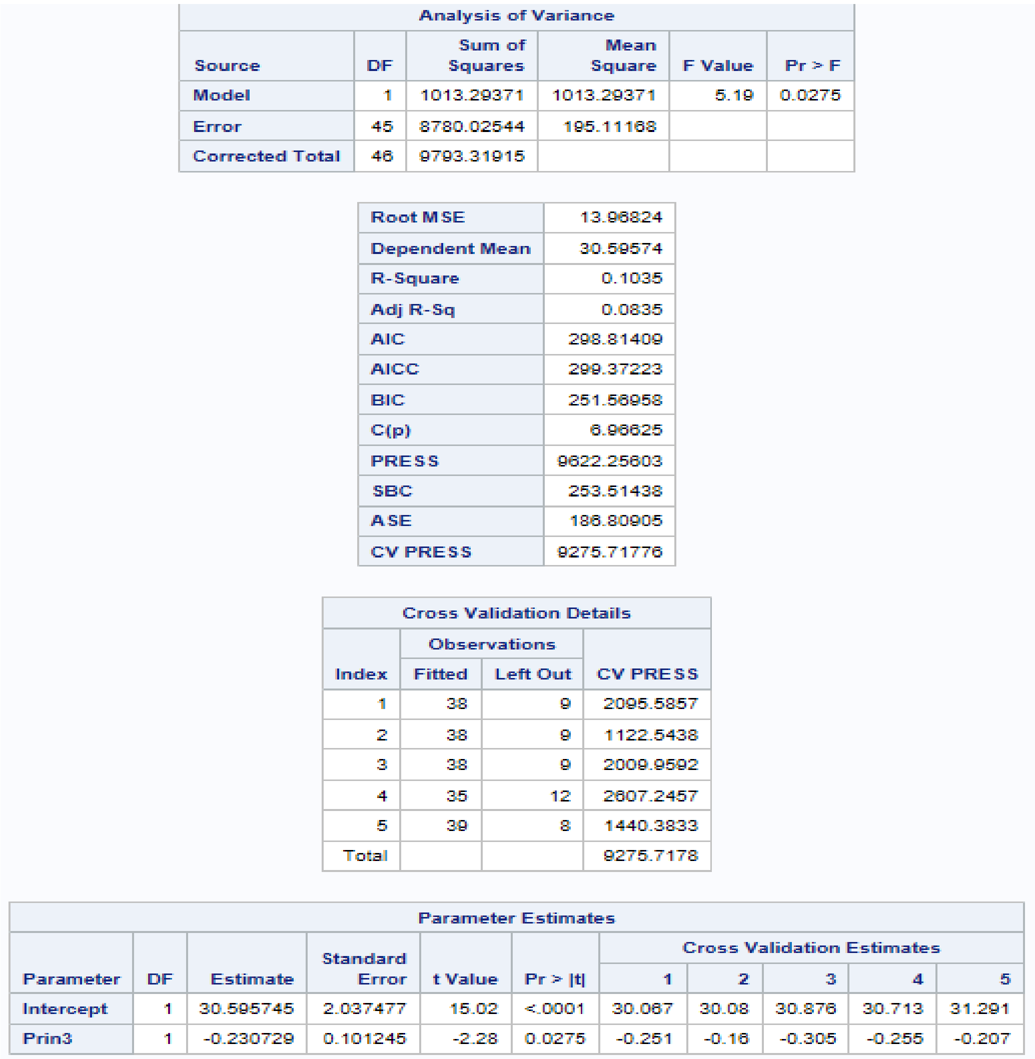
Consider data in file hw9CourseData.xlsx. This data from Chicago area zip codes in the 1970s is described below:

- General Zip Code Features
  - Fire = fires / 1,000 households
  - Theft = thefts / 1,000 population
  - Age = percentage of housing units built prior to 1940
  - Income = median family income
  - Race = percentage minority
  - Zip = zip code
- Insurance Companies New Policies
  - Vol = number of voluntary policies issued by insurance companies / 100 households
  - Invol = number of involuntary policies issued by insurance companies / 100 households

Run a principal components analysis in SAS on this data with the goal of using the components to understand the effect of the variables on the insurance companies' voluntary policies. Use the variance/covariance matrix of the variables when calculating eigenvectors and values.

1. Use proc glmselect to regress the voluntary insurance sales on PCs 1 – 5 again. Use a stepwise regression with the select = CV, choose = CV and stop = AIC. Report the SBC and CVPress from the selected model.

```
proc glmselect data=pca plots(stepAxis=number)=(criterionPanel ASEPlot CRITERIONPANEL);

model zip = prin1-prin5 / selection=stepwise(select=CV choose=CV stop=AIC) cvdetails=all showpvalues stats=all;

run;
```



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1013.29371	1013.29371	5.19	0.0275
Error	45	8780.02544	195.11168		
Corrected Total	46	9793.31915			

Root MSE	13.96824
Dependent Mean	30.59574
R-Square	0.1035
Adj R-Sq	0.0835
AIC	298.81409
AICC	299.37223
BIC	251.56958
C(p)	6.96625
PRESS	9622.25603
SBC	253.51438
ASE	186.80905
CV PRESS	9275.71776

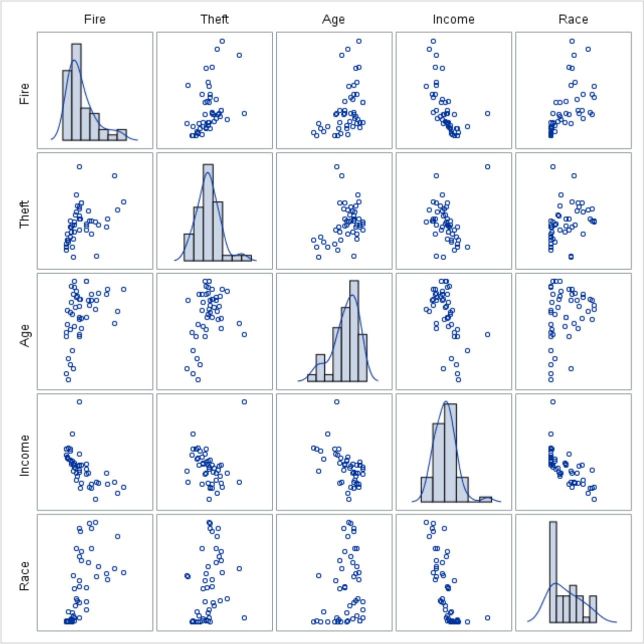
Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	38	9	2095.5857
2	38	9	1122.5438
3	38	9	2009.9592
4	35	12	2607.2457
5	39	8	1440.3833
Total			9275.7178

Parameter Estimates										
						Cross Validation Estimates				
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	1	2	3	4	5
Intercept	1	30.595745	2.037477	15.02	<.0001	30.067	30.08	30.876	30.713	31.291
Prin3	1	-0.230729	0.101245	-2.28	0.0275	-0.251	-0.16	-0.305	-0.255	-0.207

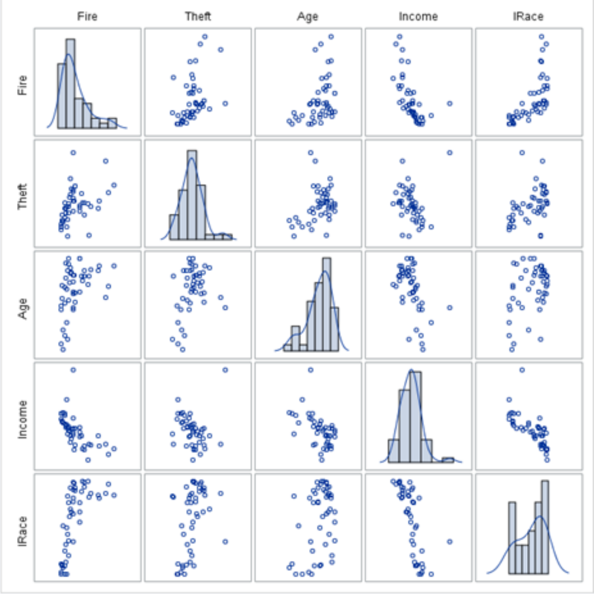
CV press of 9622.25603 and SBC of 253.51438

2. Create and display a matrix of scatterplots here for all the variables in the data set with histograms down the diagonal. Transform the Race percent to log(Race). Provide an additional matrix of scatterplots here for all the variables in the data (with log(race) this time) with histograms down the diagonal. Below is the code to do this.

```
proc sgscatter data=insurance;

   matrix Fire Theft Age Income Race/diagonal=(histogram kernel);

run;
```



```
data insurance2;
set insurance;
lRace = log(Race);
run;


proc sgscatter data=insurance2;
      matrix Fire Theft Age Income lRace/diagonal=(histogram kernel);
run;
```



3. Why did we perform the log transform?

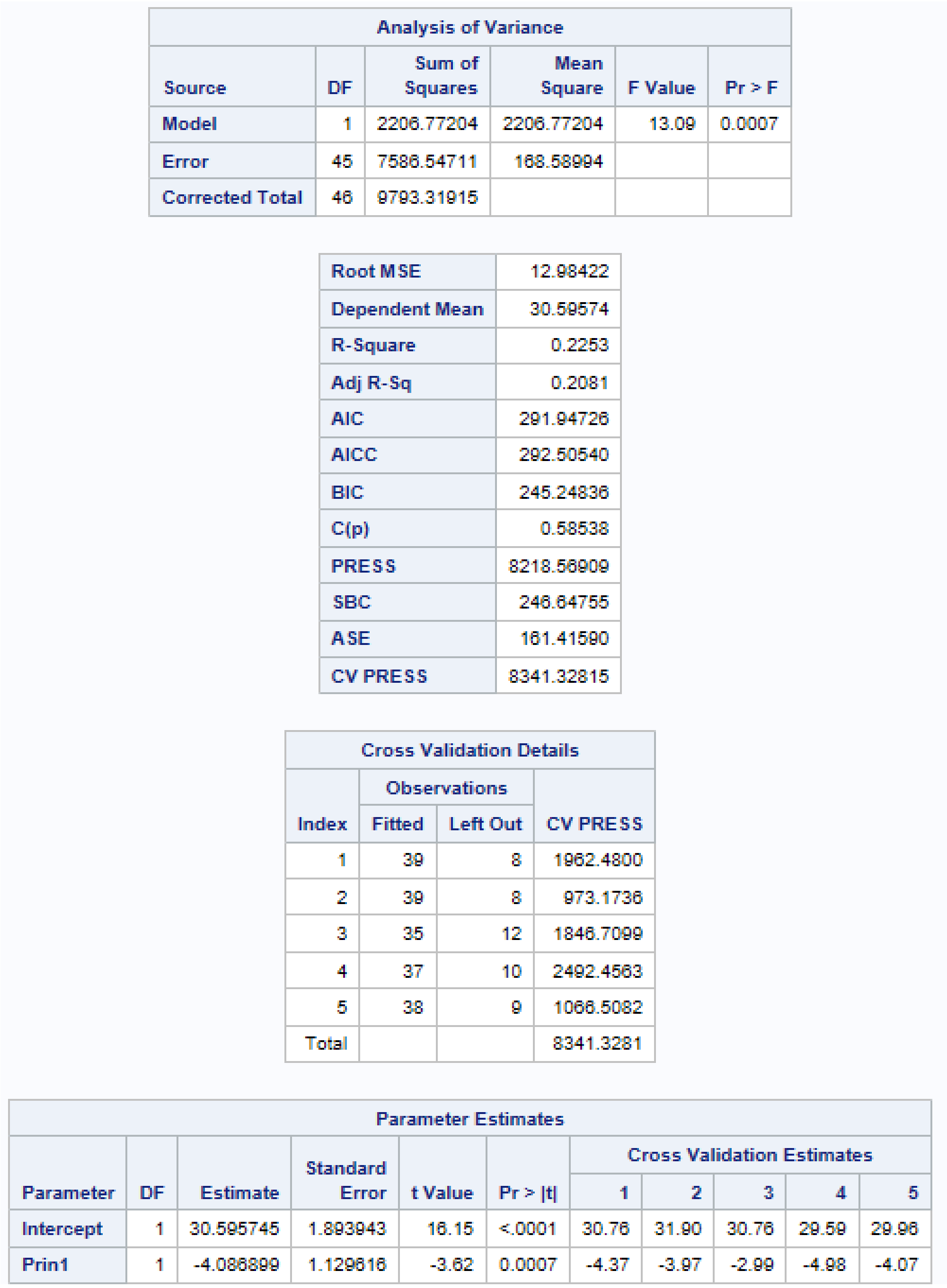
Make it the Race variable more normally distributed.

4. Did the log transform help?

Yes it helped since it was more right skewed before.

5. Re-conduct the above model selection procedure in question 1, this time with the log of the race percent. Report the SBC and the CVPRESS.

```
proc princomp plots=all data=insurance2 out=pca2;

var Fire Theft Age Income lRace ;

run;


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

model zip = prin1-prin5 / selection=stepwise(select=CV choose=CV stop=AIC) cvdetails=all showpvalues stats=all;

run;
```



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2206.77204	2206.77204	13.09	0.0007
Error	45	7586.54711	168.58994		
Corrected Total	46	9793.31915			

Root MSE	12.98422
Dependent Mean	30.59574
R-Square	0.2253
Adj R-Sq	0.2081
AIC	291.94726
AICC	292.50540
BIC	245.24836
C(p)	0.58538
PRESS	8218.56909
SBC	246.64755
ASE	161.41590
CV PRESS	8341.32815

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	39	8	1962.4800
2	39	8	973.1736
3	35	12	1846.7099
4	37	10	2492.4563
5	38	9	1066.5082
Total			8341.3281

Parameter Estimates										
						Cross Validation Estimates				
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	1	2	3	4	5
Intercept	1	30.595745	1.893943	16.15	<.0001	30.76	31.90	30.76	29.59	29.96
Prin1	1	-4.086899	1.129616	-3.62	0.0007	-4.37	-3.97	-2.99	-4.98	-4.07

6. Compare the two models found in question 1 and 5. Which do you prefer and why?

I would prefer the log one since the CV press and SBC(8218.56909 and 246.64755) are lower than the unlogged data.