Final Examination:

Data on poverty rates and determinants across 58 California counties

The purpose of this study is to use the data taken from the 1980 and 1990 census studies of 58 California counties to determine which recorded variables affect the poverty level of the counties. In doing so, we will look at the main differences between the 1980 census and the 1990 census, as well as determine what is the best statistical model to represent this data and what this model means. The data taken during each census includes:

- *POVERTY*: Percentage of families with income below poverty level
- *URBAN*: Percentage of urban population
- *FAMILY SIZE*: Number of persons per households
- *UNEMPLOYMENT*: Percentage of unemployment rate
- *HIGH SCHOOL*: Percentage of population with high school education
- *COLLEGE*: Percentage of population with four or more years of college education
- *MEDINE*: Median family income in thousands of dollars

For this study, we will be using statistical methods such as model building, correlation analysis, regression analysis, scatter plots, covariance analysis, and basics such as mean, median, standard deviation, and range to analyze and review the given data.

Once we are able to determine which variables have a significant and direct affect on the poverty level of a specific area, we will be able to determine what goals should be set in order to work toward decreasing the poverty level as well as what steps should be taken to meet those goals.

We will begin by first taking a look at the data gathered during each census and see what changes occurred from the 1st census to the 2nd one.

POVERTY LEVEL	**1980**	**1990**
Count	58	58
Mean	9.12	9.90
Median	9.05	9.8
Sum	529	574.4
Minimum	4.5	3
Maximum	18.1	20.8
Range	13.6	17.8
Standard Deviation	2.50	3.96

By looking at the poverty levels for 1980 and 1990, we can see that the average percentage of income for families that are at below poverty

level has increase slightly. The range and standard deviation of the data has also increased showing a wider range of percentages spread out over the 58 counties.

Although the percentage of families below poverty level has increase, the following chart shows that the overall average income from 1980 to 1990 has significantly increased. This information may not, however, has great influence in our study, as inflation could be the cause for the overall increase in income.

Median Income (in $1000)	**1980**	**1990**
Count	58	58
Mean	19.24	35.34
Median	18.51	32.57
Sum	1115.95	2049.59
Minimum	13.52	24.36
Maximum	29.72	59.15
Range	16.20	34.78
Standard Deviation	3.30	8.26

We will now take a look at the data from some of the factors included in the census that may have an affect on the poverty level, given from each of the census years:

	Family Size		**Urban Population**		**Unemployment Rate**	
	1980	**1990**	**1980**	**1990**	**1980**	**1990**
Count	58	58	58	58	58	58
Mean	3.14	2.69	9.29	9.95	58.76	34.10
Median	3.14	2.64	8.95	9.7	66.15	32.3
Sum	182.13	156.07	539	577.2	3408.1	1977.9
Minimum	2.76	2.29	3.5	4	0	2.7
Maximum	3.73	3.26	17.6	21.3	100	94.3
Range	.97	.97	14.1	17.3	100	91.6
Standard Deviation	.18	.24	3.31	3.93	31.82	19.48

	High School Educated		**College Educated**	
	1980	**1990**	**1980**	**1990**
Count	58	58	58	58
Mean	55.99	57.57	16.65	18.79
Median	56.7	58.7	14.95	16.35
Sum	3247.3	3338.8	965.5	1089.9
Minimum	41.3	43	9	9
Maximum	65.5	68.5	38.3	44
Range	24.2	25.5	29.3	35
Standard Deviation	5.64	6.22	6.28	7.70

After reviewing all of this data, we can see that changes over the years in each of the data sets. The family size has decrease, while the

percentage of urban population has increased slightly. The unemployment rate has significantly decreased from a median of 66.15 in 1980 to 32.3 in the 1990 census. The percentages of high school and college educated individuals have increase by about 2% each. Overall, it looks as if the 58 California counties have improved based on this data. The median income has increase and the unemployment rate has gone down. We have seen an increase in the number of educated individuals on both a high school and college level, yet the poverty rate still shows that it increase from the 1980 to the 1990 census. It is important now to use further analysis to determine which of these factors are influencing the poverty rate. Once this is determined, we will be able to decide on what action will be needed in order to decrease the poverty level.

I have started out forming three models to look at the significance of the data. By using regression analysis, I have come up with a model for all of the data, a model showing just the 1980 census data, and a 3rd model showing only the 1990 census data. We begin by taking a close look at the overall model, which includes the data from both censuses.

MODEL 1

Regression Statistics	
Multiple R	0.863738203
R Square	0.746043683
Adjusted R Square	0.729583552
Standard Error	1.724508575
Observations	116

This means that 75% of our data can be explained by this model

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	943.5386823	134.7912403	45.32428412	2.45237E-29
Residual	108	321.1844211	2.973929825		
Total	115	1264.723103			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
POVERTY	21.20362892	5.750858214	3.687037331	0.000356819	9.804430603	32.60282724
URBAN	-0.00211732	0.006951733	-0.30457568	0.761275466	-0.01589686	0.011662211
FAMSIZE	1.961590663	1.27582312	1.537509888	0.127093175	-0.56731202	4.490493345
UNEMPL	0.073485678	0.060093628	1.222853071	0.224046943	-0.04563031	0.192601671
HIGHSCHL	-0.19957748	0.039728484	-5.02353629	2.0199E-06	-0.27832622	-0.12082874
COLLEGE	0.02249674	0.046158964	0.487375328	0.626980193	-0.06899833	0.113991811
MEDINE	-0.41588105	0.046883898	-8.87044447	1.70439E-14	-0.50881303	-0.32294899
D90	8.524662504	1.049858447	8.119820849	8.18611E-13	6.443660992	10.60566402

For our first model, we can determine that this is a significant model, as the F calculated is a large number at 45. By look at R^2, we also see that about 75% of our data can be explained by the model. This is a significant percentage and shows this is a good model. If we can find another model with a higher value for R^2, then it would be a better model to use for our data.

By looking at the t-Stat values, we can determine from this model that the high school education and the median income appear to be the other variables that significantly affect the poverty level. Since the other calculations are so low, we conclude that they are insignificant variables for this model.

We will now take a look at the models for each of the individual census years to see if we get similar results, and to determine which one of these three models are more of a fit for the data.

Model for 1980 Census

Regression Statistics	
Multiple R	0.78919799
R Square	0.622833467
Adjusted R Square	0.558853091
Standard Error	1.620654945
Observations	58

62% of the 1980 Census data fits into this model

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	221.2025273	31.60036105	14.03646402	6.60827E-10
Residual	51	133.9526451	2.626522452		
Total	58	355.1551724			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
POVERTY	27.85149435	8.675176068	3.210481739	0.002294694	10.43535203	45.26763666
URBAN	0.023835945	0.010875595	2.191691201	0.032993373	0.002002278	0.045669612
FAMSIZE	-0.46666034	1.980475987	-0.23563039	0.814663147	-4.44263170	3.509311022
UNEMPL	0.111309811	0.070249709	1.584487857	0.119264713	-0.02972236	0.252341984
HIGHSCHL	-0.16168361	0.062782403	-2.57530149	0.012958081	-0.28772454	-0.03564268
COLLEGE	-0.01635059	0.061705416	-0.26497831	0.792094225	-0.14022938	0.107528191
MEDINE	-0.53927692	0.132809254	-4.06053725	0.000168779	-0.80590262	-0.27265122

The R squared and F calculated values show that this is a significant model, however, it is not as significant as the previous model that included all of the data. We also see by looking at the t-Stat values that none of the variables are showing to have a significant relationship with the poverty level.

Model for 1990 Census

Regression Statistics	
Multiple R	0.914430918
R Square	0.836183904
Adjusted R Square	0.816911422
Standard Error	1.692492541
Observations	58

83% of the 1990 Census Data fit into this model

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	745.7082292	124.2847049	43.38745322	2.39147E-18
Residual	51	146.0910811	2.864531003		
Total	57	891.7993103			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
POVERTY	16.81756584	8.50256299	1.977940753	0.053350474	-0.25204126	33.88717295
URBAN	-0.01873498	0.014757271	-1.26954269	0.210010247	-0.04836144	0.010891471
FAMSIZE	6.091761128	1.881073037	3.23845008	0.002116234	2.315349507	9.868172748
UNEMPL	-0.01179635	0.11945747	-0.09874937	0.921724383	-0.25161722	0.228024522
HIGHSCHL	-0.11855182	0.068100371	-1.74083958	0.087742129	-0.25526901	0.018165375
COLLEGE	0.171105551	0.098165436	1.743032564	0.087354645	-0.02596978	0.368180882
MEDINE	-0.53599203	0.070354007	-7.61850054	5.76052E-10	-0.67723359	-0.39475048

Now looking at the 1990 Census Model, we see that it is significant and even has a higher R squared value than Model 1 did. It also has a significant F calculated value, although slightly less than in Model 1. When we look at the t-Stat values, we again find that there are no variables showing that they have a significant relationship with the poverty level.

After comparing the three different models, we will now focus on Model 1 since it contains the most significant data and holds the highest value for F calculated.

By looking at the coefficients in Model 1, we have the following regression equation:

y-hat = 21.20 - .002(urban) + 1.96(family size) + .07(unemployment rate) -
1.99(high school) + .022(college) - .416 (medine) + 8.52(D90)

For each increase in poverty, there is a decrease of .002 in urban population, a 1.96 percent increase in family size, a .73 percent increase in unemployment, a 1.99 percent decrease in the amount of high school graduates, a .022 percent increase in the amount of college graduates, and a .416 percent decrease in the median income for that county.

As mentioned above, the t-stats in Model 1 shows that the median income and the percentage of high school educated show that they are significant variables in the model. Since this is the case, we will take a look at Model 2, which only includes the poverty level and these two variables.

Model 2

Regression Statistics	
Multiple R	0.442971917
R Square	0.19622412
Adjusted R Square	0.181997998
Standard Error	2.999342001
Observations	116

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	248.1691776	124.0845888	13.79322648	4.36642E-06
Residual	113	1016.553926	8.996052441		
Total	115	1264.723103			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Poverty	21.206885	2.792775854	7.593479247	9.68963E-12	15.673893	26.73987798
High School	-0.1505831	0.046911517	-3.20993899	0.001728625	-0.24352328	-0.05764293
Median	-0.115255	0.027352288	-4.21372335	5.07625E-05	-0.16944479	-0.06106516

Although Model 2 shows to be a significant model, with an F calculated value at 13.79, the R squared is 19.6 %, which is significantly less than we had with Model 1. We can determine that Model 1 is still a better fitting model for our data. By transforming the variable data using the natural log (ln), we will form Model 3 to calculate the significance of that data.

Model 3

Regression Statistics	
Multiple R	0.745136554
R Square	0.555228485
Adjusted R Square	0.530745649
Standard Error	2.271710368
Observations	116

Model 3 Continued

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	702.2102921	117.0350487	22.67827514	3.31323E-17
Residual	109	562.5128114	5.160667994		
Total	115	1264.723103			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
POVERTY	90.68206055	14.33554471	6.325679449	5.70806E-09	62.26947785	119.0946432
URBAN	0.002845019	0.009167005	0.310354296	0.756884185	-0.01532368	0.021013725
FAMSIZE	-11.1776565	3.533003155	-3.16378335	0.002018363	-18.1799537	-4.17535938
UNEMPL	2.687184661	0.720795022	3.728084378	0.000307849	1.258592404	4.115776918
HIGHSCHL	-14.6881413	2.449246675	-5.99700369	2.65968E-08	-19.5424684	-9.83381422
COLLEGE	-3.81478968	0.855288427	-4.46023769	1.9989E-05	-5.50994348	-2.11963587
MEDINE	-1.65371868	0.903274721	-1.83080367	0.069861626	-3.44397976	0.136542395

The data from Model 3 shows that it is a significant model, but it again does not have a higher R squared than Model 1 does. It also shows by the t-Stat values that there are no significant variables.

Based on the information given in Models 1, 2, 3, and the 1980 and 1990 census models, Model 1 has shown to be the most significant and best fitting model to use for this data.

Although our F calculated for Model 1 shows 45.32, it is still important to conduct a significance test to make sure that the model is statistically significant and that all of the coefficients are not equal to zero. The hypotheses for the test are as follows:

H_O: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$
H_A: at least one β_i does not equal zero

Looking at the Significance F with an $\alpha = .05$, we find that p-value $\approx 0 < .005$, so we reject H_O and conclude that at least one of the coefficients is not equal to zero. This shows that the model is significant.

We can also test the significance each coefficient using the following tests:

Urban Population:
H_O: $B_1 = 0$ H_A: $B_1 \neq 0$ $\alpha = .05$
$t_{\alpha/2} \approx 1.96$
t = (-.002 -0) / 2.70 = -.00074
since 1.96 > -.00074, we accept the H_O and conclude that the urban population variable is insignificant in this model

Family Size:
H_O: $B_2 = 0$ H_A: $B_2 \neq 0$ $\alpha = .05$
$t_{\alpha/2} \approx 1.96$
t= (1.96-0)/ 0.0288 = 68.056
since 1.96 < 68.056, we reject the H_O and conclude that the family size variable is significant in this model

Unemployment Rate:
H_O: $B_3 = 0$ H_A: $B_3 \neq 0$ $\alpha = .05$
$t_{\alpha/2} \approx 1.96$
t = (.07-0) / .337 = .208
since 1.96 < .208, we reject the H_O and conclude that the unemployment rate variable is significant in this model

High School Educated:
H_O: $B_4 = 0$ H_A: $B_4 \neq 0$ $\alpha = .05$
$t_{\alpha/2} \approx 1.96$
t=(-1.99 -0)/ .55 = -3.62
since 1.96 > -3.62, we accept the H_O and conclude that the high school educated variable is insignificant in this model

College Educated:
H_O: $B_5 = 0$ H_A: $B_5 \neq 0$ $\alpha = .05$
$t_{\alpha/2} \approx 1.96$
t = (.022-0)/ .657 = .0335
since 1.96 > .0335, we accept the H_O and conclude that the college educated variable is insignificant in this model

Median Income:
H_O: $B_6 = 0$ H_A: $B_6 \neq 0$ $\alpha = .05$
$t_{\alpha/2} \approx 1.96$
t= (-.416-0)/.95 = -.438
since 1.96 > -.438, we accept the H_O and conclude that the median income variable is insignificant in this model

Each individual coefficient test has shown us that the family size and the unemployment rate are the only variables in Model 1 that are significant. We have found all other variables to be insignificant. Using only the poverty level, family size, and the unemployment rate, we have come up with Model 4 to see if these significant variables show a stronger model than Model 1.

Model 4

Regression Statistics	
Multiple R	0.577071
R Square	0.333011
Adjusted R Square	0.321206
Standard Error	2.732233
Observations	116

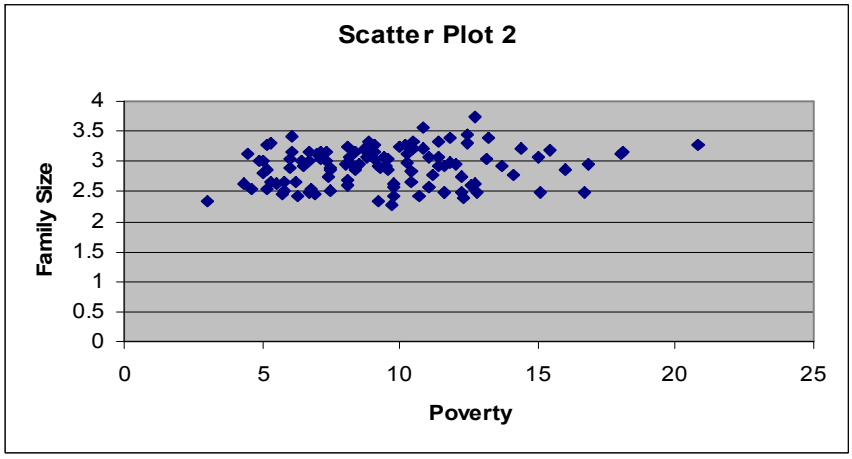
Model 4 Continued

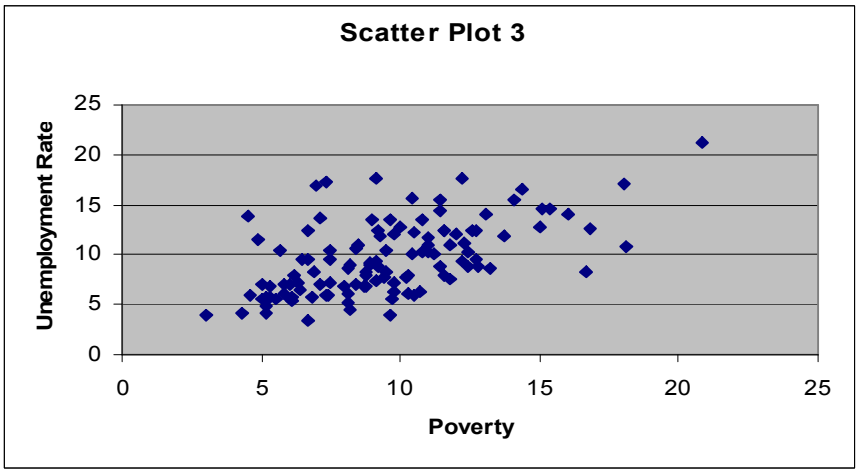
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	421.1669	210.5834	28.20906	1.16E-10
Residual	113	843.5562	7.465099		
Total	115	1264.723			

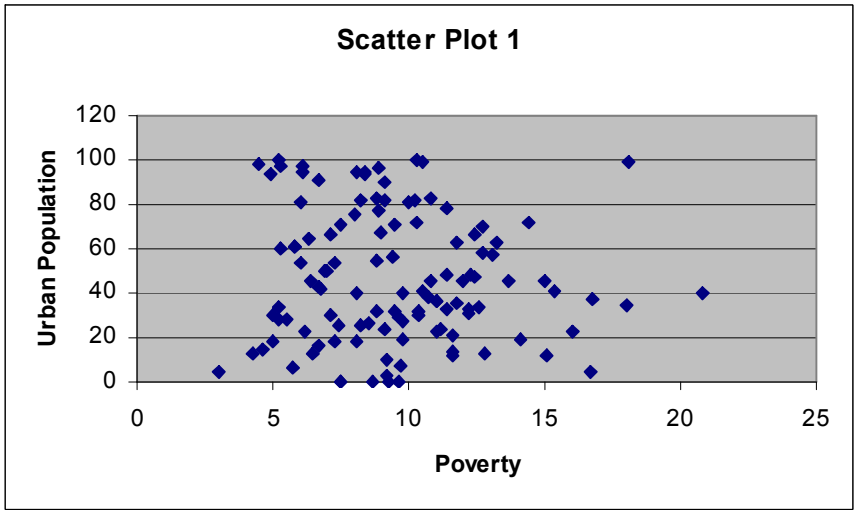
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
POVERTY	2.423566844	2.434193638	0.995634368	0.321555085	-2.39900962	7.246143316
FMLY SIZE	0.732974433	0.829686543	0.883435364	0.378876754	-0.91078419	2.376733064
UNEMP RT	0.514580087	0.070821695	7.265853821	5.11315E-11	0.374269539	0.654890634

The F significant of 28.209 shows that the model is significant, and the t-Stat values for variables also indicate the variables are all significant. The following Scatter Plots show the relationships between the poverty level and the two significant variables, of family Size and the unemployment rate.

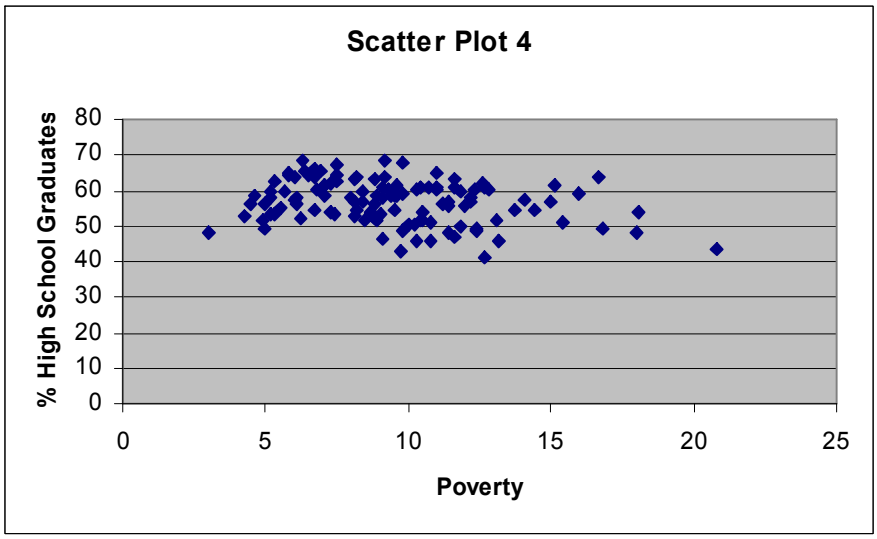




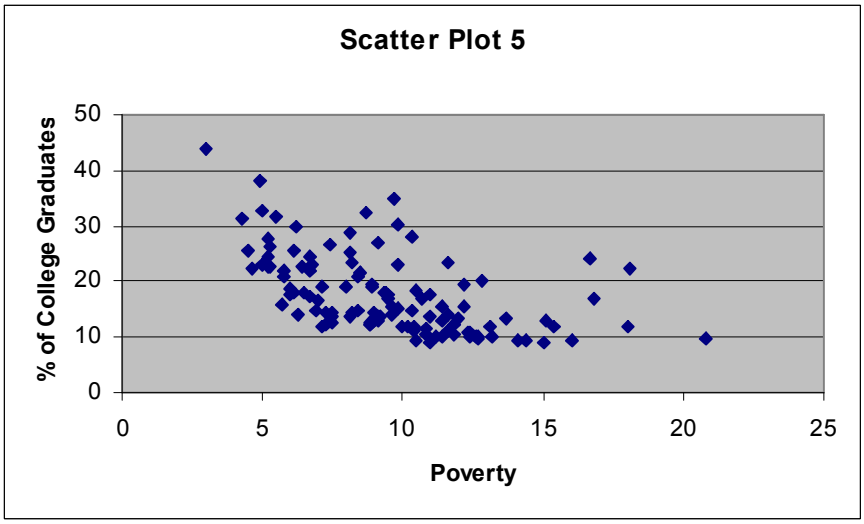
By the appearance of these two scatter plots, we can see the visual layout of the relationships between these variables and the poverty level. They both appear to have positive relationships. In comparison, we will look at the scatter plots for the insignificant variables and the poverty level to see the difference.



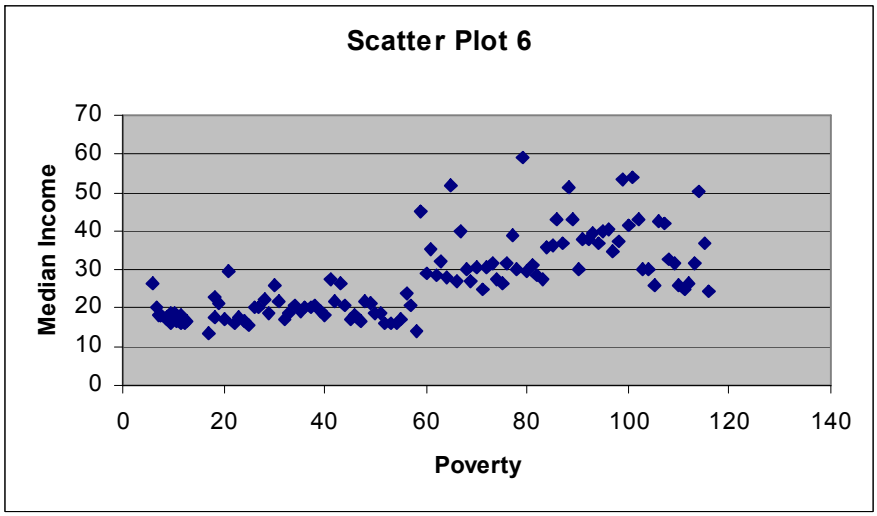
Scatter Plot 1 shows that there is not a clear relationship between the urban population and the poverty level



Scatter Plot 4 shows that there is no relationship between the high school educated and the poverty level



Scatter Plot 5 has the appearance of a negative linear relationship between the college educated and the poverty level, however our significance test and data analysis show that there is not a significant relationship



Scatter Plot 6 indicates that there may be a positive, linear relationship between the median income and the poverty level, however our significance test and data analysis show that there is not a significant relationship

Looking at the scatter plots can give us an idea of whether or not there is a significant relationship between two variables, but we must always look at the statistical data and test to see if there are actually significant relationships.

We can also take a look at the correlation for the variables to see if there are significant relationships between different variables.

Correlation

	POVERTY	URBAN	FAMSIZE	UNEMPL	HIGHSCHL	COLLEGE	MEDINE	D90
POVERTY	1							
URBAN	-0.07778	1						
FAMSIZE	0.146286	0.55685	1					
UNEMPL	0.573066	-0.1215	0.13796	1				
HIGHSCHL	-0.26444	-0.2172	-0.49429	-0.04416	1			
COLLEGE	-0.540271	-0.023	-0.27456	-0.56471	-0.202489	1		
MEDINE	-0.350618	-0.251	-0.57488	-0.27972	-0.017798	0.594149	1	
D90	0.11853	-0.4264	-0.72773	0.09106	0.1328547	0.152187	0.79041	1

Looking at the Correlation, we can see that there is a significant, positive, correlation between the Poverty level and the Unemployment Rate, there is a significant negative correlation between the poverty level and the # of college educated individuals, there is a significant positive correlation between the Urban population and the family size, there is a significant negative correlation between the unemployment rate and the # of college educated individuals, and there is a significant positive correlation between the median income and the # of college educated individuals.

By reviewing all of our statistical data and tests, we have concluded that the main variables included in the 1980 and 1990 censuses of 58 California counties that are significant in influencing the poverty level are the family size and the unemployment rate. This means that as family sizes are larger in each county, the poverty level is higher. This could come from the fact that families are unable to make enough money to support large families, leaving those families below the poverty level. It also makes sense that as the unemployment rate increases, so does the percentage of families living below poverty. As individuals are unemployed, this directly affects their household income. If the providers of a household are not working, that family could be living below poverty level.

If the counties that were included in the census are working towards decrease the percentage of families considered to be below poverty, they will need to focus some energy on working to decrease the unemployment rate. When we are able to employ those who are currently counted in the unemployment rate, we will decrease that rate and at the same time increase the overall family income. The increase in income will help to bring many families to an above poverty level.

One of the important steps in decreasing the unemployment rate is to find out why the unemployed are not currently working. It could be due

to lack of education, lack of skills or training, and/or lack of opportunity. Although high school and college education did not show to be a significant factor in our statistical model, it may be necessary to educate or train individuals with the set of knowledge or skills needed for the jobs that are currently available.

One our goals should be to work with state colleges and universities to have programs available so that individuals in the community can obtain the necessary skills they need to be employable in areas that are in need to employees. These programs should be assessable to anyone in the community and be tailored to what each community is in need of. Programs should be set up to work directly with employers who are looking for individuals to work for them. It may be that employers can work directly with the institutions in the area to set up programs specifically for a particular job or occupation.

Another goal would be to assist companies in advertising for job openings and what skills and knowledge is need for specific jobs. Information on where a potential employee can go for training if they are interested in a job, can help to bring forth opportunities for those seeking jobs. If individuals know what is needed to do a job and where they can go to learn so skills, they are more likely to see an attainable goal. This goal will help them to earn more income for their families and help to bring them out of poverty.

When we look back over the data comparisons between the 1980 and 1990 census data, we show that the unemployment rate has decrease greatly. If we keep making an effort to reduce the rate, we should help to further improve our economy and the family lives of those living in the California counties. We also see that the percentages of individuals who have high school and college educations have increased. This increase in education should help to continue to decrease the unemployment rate and bring the overall average income up for families and the communities in general.

Overall, it looks like we are on the right path to decreasing the poverty levels and should continue to work to make sure individuals are educated are able to support their families and live at an above poverty level.