Predicting Poverty from Census Data with Multiple Regression

Jack Hiegel, Long Tran, and Beomchang Kim

Introduction

According to UNESCO, poverty is when a family's income fails to meet a certain threshold that differs across countries. Within the United States, the percentage of people in poverty varies from state to state. Our project is going to look at creating a linear regression model to predict the percentage of people in poverty in a particular state by using a number of other variables. We collected data from all 50 states for a total of six variables. This data came from the U.S. Census Bureau. Our response variable is the percent of people in poverty within a particular state in 2018. Our five explanatory variables are as follows:

X1: The percent of persons aged 25+ who have a bachelor's degree or higher (2013-2017)
X2: The percent of households with a computer (2013-2017)
X3: The average number of persons per household (2013-2017)
X4: The percent of people with a disability under age 65 (2013-2017)
X5: The mean travel time to work in minutes for workers age 16+ (2013-2017)

Using this data, we hope to see what kind of variables affect the poverty level in a particular state. Other questions of interest deal with figuring out the specific nature of the relationship between poverty level and our explanatory variables and determining poverty level for a state with specific explanatory variable values.

Full Model

Calculating X^TX

Before fitting our full model, we should calculate the value of matrix X^TX, with the matrix X being a matrix of the levels of the regressor variables. This matrix assists us in obtaining a linear model. Its diagonal elements are the sums of squares of the columns of X, and the off-diagonal elements are the sums of cross products of the elements in the columns of X. The following table (Table 1) is X^TX of the regression model obtained by using R.

	x1	x2	x3	x4	x5
x1	46595.490	131391.55	3877.6010	13520.420	36895.900
x2	131391.550	378137.45	11188.2610	39701.740	105333.180
x3	3877.601	11188.26	332.2891	1175.347	3126.882
x4	13520.420	39701.74	1175.3470	4375.100	11059.880
x5	36895.900	105333.18	3126.8820	11059.880	30044.700

Table 1

Global F-Test

In the following table, which we shall call Table 2, we have the regression output from Minitab for the model including all of our explanatory variables. Our first step is to conduct the global F-test, which will help us to determine whether our model is useful in predicting poverty percentage. Our hypotheses are as follows:

Ho: $\beta j = 0$ for $j = 1, 2, 3, 4, 5$

Ha: At least one $\beta j \neq 0$

We will set our alpha value equal to 0.05 for this test and for all remaining tests. As found in Table 2, our test statistic is 49.29, and our p-value is 0.000. Because our p-value is less than our alpha, we reject the null and have statistically significant evidence that our model is useful in predicting our response.

Regression Analysis: Percent of People in Poverty, 2 ... to work (minut

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	333.013	66.6026	49.29	0.000
Bachelor_s degree or higher, pe	1	3.583	3.5828	2.65	0.111
Households with a computer, per	1	52.748	52.7485	39.04	0.000
Persons per household, 2013-201	1	12.115	12.1148	8.97	0.005
With a disability, under age 65	1	7.374	7.3738	5.46	0.024
Mean travel time to work (minut	1	0.135	0.1348	0.10	0.754
Error	44	59.451	1.3512		
Total	49	392.464			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.16240	84.85%	83.13%	78.76%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	49.94	7.92	6.31	0.000	
Bachelor_s degree or higher, pe	-0.1120	0.0688	-1.63	0.111	4.38
Households with a computer, per	-0.5600	0.0896	-6.25	0.000	3.00
Persons per household, 2013-201	4.25	1.42	2.99	0.005	2.05
With a disability, under age 65	0.384	0.164	2.34	0.024	3.15
Mean travel time to work (minut	0.0198	0.0626	0.32	0.754	2.04

Regression Equation

Percent of People in Poverty, 2 = 49.94 - 0.1120 Bachelor_s degree or higher, pe
- 0.5600 Households with a computer, per
+ 4.25 Persons per household, 2013-201
+ 0.384 With a disability, under age 65
+ 0.0198 Mean travel time to work (minut

Fits and Diagnostics for Unusual Observations

Obs	Percent of People in Poverty, 2	Fit	Resid	Std Resid		
11	8.800	12.399	-3.599	-3.43	R	
31	19.500	16.869	2.631	2.51	R	
44	9.000	10.334	-1.334	-1.48		X

R Large residual
X Unusual X

Table 2

Individual T-Tests

Having confirmed the usefulness of our model, we should now move on to checking the significance of each individual explanatory variable within the model. We will accomplish this through the individual t-tests, which evaluate the effect of a particular explanatory variable on the response given that the other variables are in the model. We will need to do five individual t-tests, one for each of our explanatory variables. Our hypotheses will be:

Ho: $\beta j = 0$ for j = the number of the particular explanatory variable we are testing

Ha: At least one $\beta j \neq 0$

Looking at our coefficients table in Table 2, we obtain our test statistics (the t-values) and the p-values for these tests. For example, for X1, the t-value is -1.63, and the p-value is 0.111. Comparing the p-values to the alpha value of 0.05, we find that three of them (the ones for X2, X3, and X4) are less than 0.05, and the ones for X1 and X5 are greater than 0.05. Therefore, we reject the null hypotheses in the cases of X2, X3, and X4. Thus, we have statistically significant evidence that these three variables all significantly affect the percentage of people living in poverty in a state given that the other variables are included in the model. On the other hand, we fail to reject the null hypotheses in the cases of X1 and X5. Thus, we lack statistically significant evidence that the percentage of people with a bachelor's degree or higher and the mean travel time to work have a significant effect on our response variable given that the other variables are included in the model. Because of this, we want to look into the possibility of removing X1 and X5 from our model in order to get a more appropriate one.

Partial F-Tests

In order to determine whether or not X1 and X5 should be included in our model, we will take a look at partial F-tests. These tests will compare our full model with all variables included to models where certain variables are excluded in order to determine the significance of the excluded variables. First, let us remove X5 from the model and determine whether or not it is necessary. The Minitab output for the model, which includes only the first four explanatory variables, is included in Table 3. Our hypotheses are as follows:

Ho: $\beta 5 = 0$

Ha: $\beta 5 \neq 0$

Our test statistic of F is equal to the change in SSR divided by the number of excluded variables divided by the MSE for the full model. Upon calculating this value, we get a test statistic of 0.0999. Our cutoff value for this test, which is equal to $F_{(0.05, 1, 44)}$, is approximately 4.08. Because our test statistic is less than our cutoff value, we fail to reject the null hypothesis and lack statistically significant evidence that mean travel time to work significantly impacts the percentage of a state's population living in poverty given that our other variables are present in the model.

Regression Analysis: Percent of People in Poverty, 2 ... y, under age 65

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	332.878	83.219	62.85	0.000
Bachelor_s degree or higher, pe	1	4.837	4.837	3.65	0.062
Households with a computer, per	1	64.627	64.627	48.81	0.000
Persons per household, 2013-201	1	20.465	20.465	15.46	0.000
With a disability, under age 65	1	8.768	8.768	6.62	0.013
Error	45	59.586	1.324		
Total	49	392.464			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.15071	84.82%	83.47%	79.97%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	50.13	7.82	6.41	0.000	
Bachelor_s degree or higher, pe	-0.0976	0.0511	-1.91	0.062	2.47
Households with a computer, per	-0.5710	0.0817	-6.99	0.000	2.55
Persons per household, 2013-201	4.51	1.15	3.93	0.000	1.37
With a disability, under age 65	0.399	0.155	2.57	0.013	2.87

Regression Equation

Percent of People in Poverty, 2 = 50.13 - 0.0976 Bachelor_s degree or higher, pe
- 0.5710 Households with a computer, per
+ 4.51 Persons per household, 2013-201
+ 0.399 With a disability, under age 65

Fits and Diagnostics for Unusual Observations

Obs	Percent of People in Poverty, 2	Fit	Resid	Std Resid	
11	8.800	12.414	-3.614	-3.48	R
31	19.500	16.967	2.533	2.34	R

R Large residual

Table 3

Having determined that X5 is not necessary, we will now conduct a partial F-test on a model which excludes X1. We will use the model in Table 3 as our new full model and the model in Table 4 as our reduced model. Our hypotheses are as follows:

Ho: $\beta 1 = 0$

Ha: $\beta 1 \neq 0$

We calculate our F statistic in a similar fashion as before and wind up with a test statistic of 3.6541. Our cutoff value is $F_{(0.05,\ 1,\ 45)}$, which is roughly 4.08. Because our test statistic is less than our cutoff value, we fail to reject the null hypothesis and lack statistically significant evidence that the percentage of people with a bachelor's degree or higher significantly impacts a state's poverty percentage given that our other variables are present in the model.

R^2 and R^2_{adj}

While our partial F-test suggests that it would be acceptable to remove X1 from the model, we have indications to believe that it might be better to keep it in the model. This comes from the fact that our t-test for X1 just barely rejected the null hypothesis as our p-value of 0.111 was very close to our alpha. Additionally, the p-value for X1 in Table 3 is 0.062, which is even closer to alpha. Because of this, we want to take a closer look at the effect of X1 on the model.

We will do this by taking a look at the value of R^2 and R^2_{adj} for the models in Table 3 and Table 4. R^2 is the percentage of the variation in the response that can be explained by the explanatory variables in the model. In Table 3, we see that R^2 is equal to 84.82%, and in Table 4, this value is 83.58%, making them both quite useful in predicting the response. As we can see, the model with X1 included does a slightly better job at explaining the variation in poverty percentage than the one without X1. However, R^2 can only increase for each additional variable included in the model, so we should find a more suitable method to compare the two models. This is where R^2_{adj} comes in as it takes into consideration the effects of additional variables and provides us with a value more suited for model comparison. In Table 3, we see that R^2_{adj} is equal to 83.47%, and in Table 4, this value is 82.51%. Our full model in Table 2 has an R^2_{adj} value equal to 83.13%. As we can see, the model with the largest R^2_{adj} is the one with X1 and without X5. Because of this and because of our initial doubts about the supposed insignificance of X1, we decide to retain X1 in our model. Thus, we have our final model, which predicts the percentage of the population in a state living in poverty given the percentage of people with bachelor's degrees or higher, the percentage of households with a computer, the average household size, and the percentage of people with a disability. Table 3 will serve as our final model table.

Regression Analysis: Percent of People in Poverty, 2 ... y, under age 65

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	328.04	109.347	78.08	0.000
Households with a computer, per	1	93.84	93.840	67.00	0.000
Persons per household, 2013-201	1	31.37	31.372	22.40	0.000
With a disability, under age 65	1	20.37	20.367	14.54	0.000
Error	46	64.42	1.401		
Total	49	392.46			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.18343	83.58%	82.51%	78.95%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	49.35	8.03	6.15	0.000	
Households with a computer, per	-0.6325	0.0773	-8.19	0.000	2.15
Persons per household, 2013-201	5.26	1.11	4.73	0.000	1.21
With a disability, under age 65	0.538	0.141	3.81	0.000	2.24

Regression Equation

Percent of People in Poverty, 2 = 49.35 - 0.6325 Households with a computer, per
+ 5.26 Persons per household, 2013-201
+ 0.538 With a disability, under age 65

Fits and Diagnostics for Unusual Observations

Obs	Percent of People in Poverty, 2	Fit	Resid	Std Resid		
11	8.800	12.430	-3.630	-3.40	R	
31	19.500	17.199	2.301	2.05	R	
44	9.000	10.376	-1.376	-1.38		X

R Large residual
X Unusual X

Table 4

Final Model

Assumptions

Before moving on to evaluating our final model, we should first verify the four regression assumptions. By doing this, we can guard against a variety of problems that could crop up should our assumptions not be met. To verify these assumptions, we use Graph 1, which displays residual plots for our model, and Graph 2, which displays plots comparing our explanatory variables to our residuals.

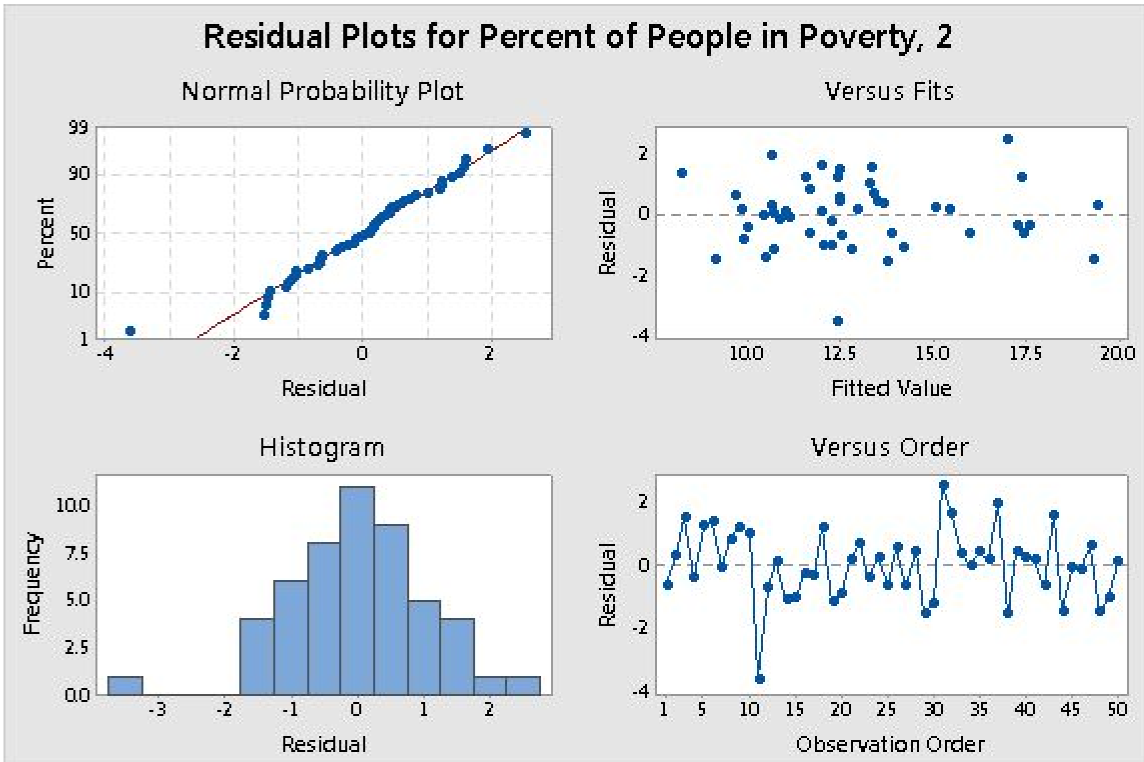
The first of our assumptions is linearity, which will help prevent predictions from being biased and the model from fitting the data poorly. To verify this assumption, we check our Residuals vs. Fits. graph in Graph 1 and all of our graphs in Graph 2 for curvilinear scatter of points. The points appear to form an equal spread without curves around the horizontal lines in all graphs. Therefore, the linearity assumption is satisfied.

The second of our assumptions is constant variation, which will help prevent the estimate of the population standard deviation from being biased and the inference procedures from being invalidated. We take a look at the same graphs as before to see if the distributions take on a fan or eyeball shape, which would indicate that there is a problem here. We can see that the residuals are relatively equally spread above and below the horizontal lines, not taking on the aforementioned shapes. Therefore, we can also assume constant variation for the data.

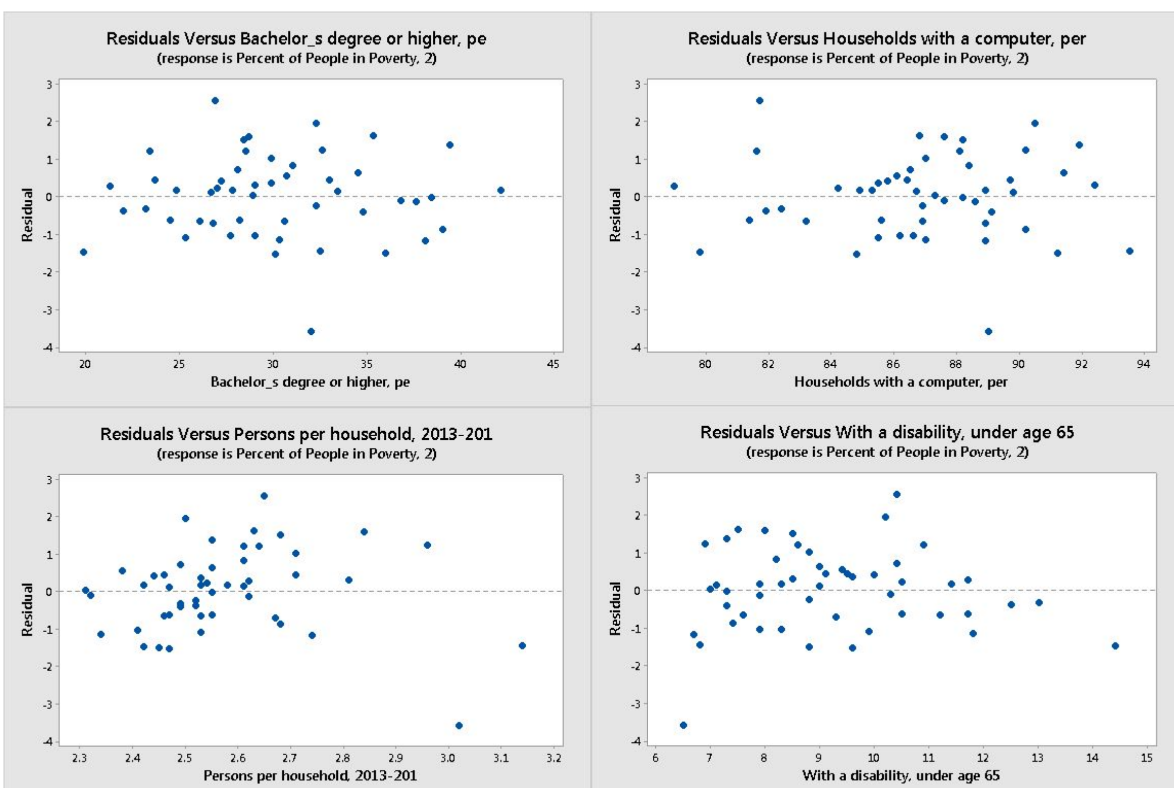
Our third assumption is normality. If the normality assumption is not met, then we cannot guarantee that conclusions based on inferential procedures will be valid. To check this assumption, we take a look at the Normal Probability Plot and the histogram in Graph 1. The first of these figures shows a relatively linear spread. There is some curve at the end but not enough to dissuade us from our conclusion that it is linear. The histogram has a roughly symmetric distribution barring the presence of a few possible outliers. Based on this evidence, we can assume normality.

Finally, we need to meet the independence assumption. If there was a violation here, then our estimated standard errors of regression coefficients may be biased and our hypothesis tests and confidence intervals may be suspect. To check this, we look at the Residuals vs. Order plot on Graph 1. We see that the residuals vary randomly around the horizontal line. The large number of runs is a potential indication of negative autocorrelation, but we have no other indication that independence would be violated for this data set. Therefore, we can conclude that this is a well-behaved Residuals vs. Order plot and that the independence assumption is satisfied.

Graph 1



Graph 2



Interpretation of Coefficients

Now that we have determined that our final model satisfies our assumptions, we shall now begin to take a closer look at it. First of all, we shall state the model itself, which is as follows:

$\hat{y} = 50.13 - 0.0976x_1 - 0.5710x_2 + 4.51x_3 + 0.399x_4$

These values come from the coefficient table in Table 3, and we can use them to observe the marginal effect each variable has on the response assuming the other variables are accounted for. We can do this by interpreting the coefficients. For example, let's look at β3, which is the coefficient for X3. For every one-person increase in the average number of persons per household in a state, the predicted percentage of people within that state who live in poverty increases by 4.51% assuming the other variables are in the model. In other words, X3 and our response have a positive relationship in this model. Interpretations of the other explanatory variables take a similar form. We will take a closer look at the direction of all of these relationships when looking at our scatterplots.

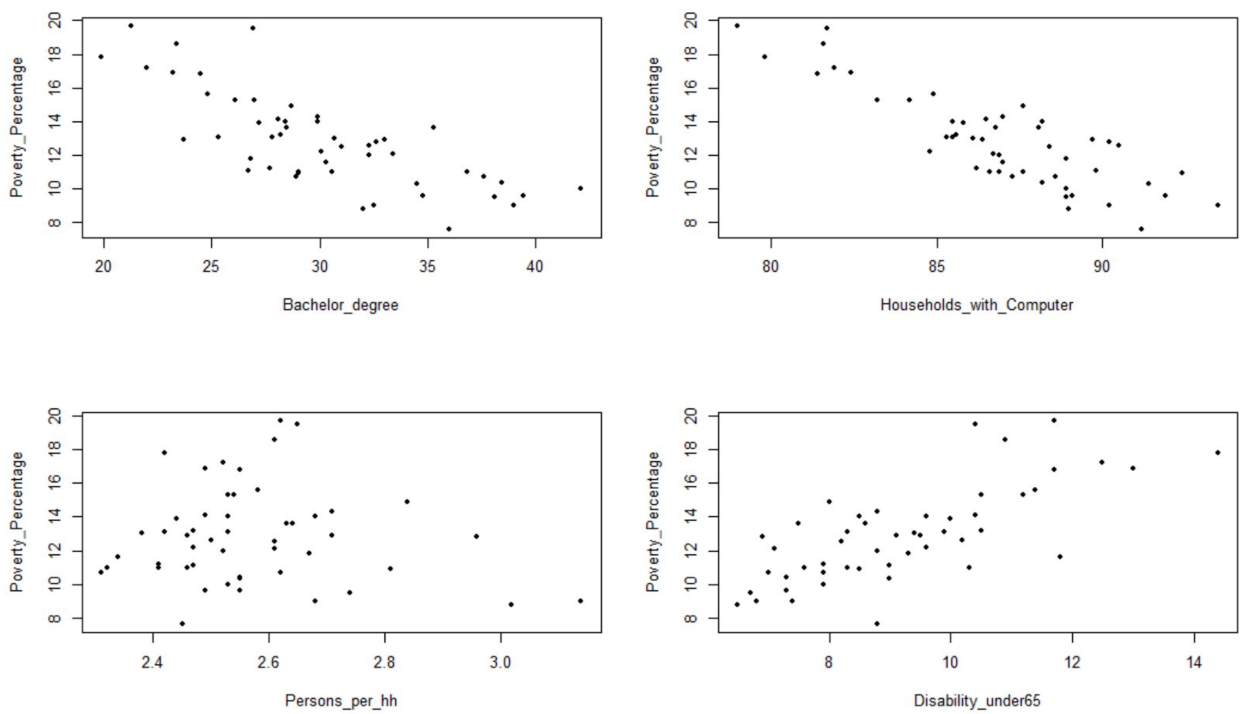
We also calculated the 95% confidence interval for the marginal effect of X3 to get a better idea of how this variable affects our response. The calculation for this interval is shown below:

$\beta_3 \pm t_{\alpha/2}\, se(\beta_3) = 4.51 \pm (2.0141)(1.15) = (2.194, 6.826)$

Therefore, there is a 95% chance that the marginal effect for X3 is between 2.194 and 6.826. Because the entire interval is positive, we have more proof that the effect of X3 upon the response is positive.

It would also be wise to interpret our y-intercept and determine if this interpretation is valid. When all four of our variables register values of 0 in a state, then the percentage of people living in poverty in that state is predicted to be 50.13%. However, this interpretation does not mean much because none of our explanatory variables would likely be recorded at the value of 0 in real life. In particular, X3 is not going to record a value of 0 because it is impossible for a state's average number of persons per household to be 0.

Scatterplots



Graph 3

We used R to make scatterplots of the 4 explanatory variables against the response variable to get a better idea of how their relationships look. These scatterplots are shown in Graph 3. In the first plot, we can see that X1 has a pretty strong negative correlation with the response. In other words, as the percentage of persons who have a bachelor's degree or higher in the state increases, the population percentage of that state in poverty decreases. In the next plot, we can see that X2 also has a strong negative correlation with the response. As the percent of households with a computer in the state increases, the population percentage of that state in poverty decreases. In the third plot, it is harder to tell what kind of relationship exists between X3 and the response, but as the coefficient for X3 is 4.51, we can tell there is a slight positive relationship between them. Lastly, the graph with X4 and the response indicates a moderately strong positive correlation between the 2 variables. A state with a higher percent of people with a disability under the age 65 tends to have a higher population percentage of that state in poverty. Overall, the scatterplots confirmed what we knew about our data and helped us to visualize it.

Multicollinearity

One potential issue which we want to address is multicollinearity, which occurs when one explanatory variable can be expressed as a linear combination of other explanatory variables. Among the problems raised by the presence of multicollinearity are the inflation of standard errors of coefficients and the instability of coefficient estimates. To determine whether or not multicollinearity exists between our explanatory variables, we chose to use the VIF method. VIFs, or Variance Inflation Factors, measure how much a variable's variance is affected by multicollinearity. The VIF values are located in Table 3 on the coefficients table.

There are three ways to use VIF values to test for multicollinearity. We performed all three. First, we checked the individual VIF values of each explanatory variable in Table 3 and compared them to 10. If they are greater than 10, then there is indication of multicollinearity. The VIF values of X1, X2, X3, and X4 are 2.47, 2.55, 1.37, and 2.87 respectively, which are all less than 10, so no multicollinearity was detected this way.

Second, we checked the average of VIFs and compared this new value to 5. If greater than 5, then there could be an indication of multicollinearity. We found that there was no indication of multicollinearity between the variables since the average of VIFs, 2.315, is less than 5.

Third, we compared each VIF with the guideline $1/1-R^2$. As R^2 is 0.8482, $1/1-R^2$ is approximately 6.5876. All values of VIF are less than 6.5876, which means that no indication of multicollinearity is present here. Therefore, from the VIF method, we can conclude that there is no indication of multicollinearity.

Unusual Points

We now move on to locating unusual points that are present in our data. These are points that don't match up with the rest of our data, and they might influence the model and its coefficients. We want to look at two types of unusual points in particular, which are regression outliers and influential points. Table A contains all relevant information obtained from Minitab and is included at the back of this report.

Regression outliers are data points that do not match the pattern of other points. In order to locate outliers, we need to use the deleted t residuals method. With this method, we take a look at the standardized residual for each observation and compare it to the value of +/- $t_{(n-p)}$. If the residual falls outside of this value, then we can say that the point in question is an outlier. Using Minitab, we calculated the standardized residuals for each state. Then, we calculated our cutoff value, which, with n = 50, p = 5, and alpha = 0.05, is equal to +/- 2.0141. Since the standardized residuals for Hawaii and New Mexico are outside the critical value (-4.0251 and 2.4701 respectively), we have evidence that these two states were outliers.

Next, we should check for influential points, which are points which have some undue influence on our model. To determine influential points, first we use the Cook's Distance method. The value of Cook's Distance for a particular observation measures the change in the set of coefficients if the observation is removed. If this value is greater than $F_{(0.5, p, n-p)}$, then the observation it represents is considered an influential point. We calculated the Cook's Distances for all states in Minitab. Our cutoff value is equal to $F_{(0.5, 5, 45)}$ = 0.8835. Using this method, none of our observations have Cook's Distance values greater than the cutoff. Therefore, we lack indications of influential points from this approach.

We can also use the DFFIT approach. The DFFIT value measures change in predicted value if the observation in question is removed. The absolute value of an observation's DFFIT value is considered significant if it is greater than $2\sqrt{\frac{p}{n}}$. We use Minitab to acquire the DFFIT values for all the states, and our cutoff value comes out to be $2\sqrt{\frac{p}{n}}$ = 0.6325. Looking at the values shown in Table A, we determine that Hawaii, New Mexico, Oregon, Utah, and West Virginia are influential points because their DFFIT absolute values are greater than the cutoff.

Point Prediction

Having acquired a list of unusual points, we want to use to check how accurate our model estimates the poverty percentage for each state. We will do this through the calculation of point predictions. Essentially, we will be taking our model and plugging in the values for a particular state to see what the predicted response value is. Then, we will compare the predicted response to the observed response. So, we decided to calculate point predictions for both non-unusual point states (Missouri, California and Maryland) and unusual point states (New Mexico, Oregon and Hawaii).

For the non-unusual point states, upon plugging their explanatory variable values into our model, we received values of 13.8, 11.6 and 9.9, respectively. These are not identical to the respective observed values of 13.2, 12.8 and 9. However, we can see that the differences between predicted and observed are very small. On the other hand, the unusual point states' estimated values are, respectively, 17.0, 10.7, and 12.4. When we compared them with their respective observed values of the states, which are 19.5, 12.6 and 8.8, we found that the error in all cases is slightly larger than the others.. However, none of these errors is extremely large. Thus, our model does a pretty good job at matching up to what we see in real life.

Confidence Interval

Our final point of order is to calculate 95% confidence and prediction intervals using our model. We will start with the confidence interval. We wish to find the mean confidence interval for states which display average values for all of our explanatory variables. We calculate the means of our explanatory variables as 30.114% of persons aged 25+ who have a bachelor's degree or higher, 86.906% of households with a computer, 2.5726 persons per household, and 9.184% of people with a disability under age 65. We then plug these values into Minitab to obtain a 95% confidence interval for the mean of the response, which is shown in Table 5.

Prediction

Fit	SE Fit	95% CI	95% PI
12.846	0.162735	(12.5182, 13.1738)	(10.5053, 15.1867)

Table 5

From the Minitab output, the 95% confidence interval for the mean response would be (12.5182, 13.1738). Therefore, we are 95% confident that the average poverty percentage for states with mean values for each of our explanatory variables is between 12.5182% and 13.1738%.

Prediction Interval

Using the same data and output, we determined the prediction interval for an individual state displaying the mean values of our explanatory variables. From Table 5, we determine that we are 95% confident that the percentage of this particular state's population which is living in poverty is between 10.5053% and 15.1867%.

Conclusion

After having reviewed this data, we have learned much about the relationship that a state's percentage of population living in poverty has with a number of other socioeconomic variables related to the state. Chiefly, we found that computer usage, household population, and disability rates have a significant impact on poverty percentage while the effect of education is more muddled and the effect of travel time to work is not significant. This last point regarding travel time's lack of effect is worth some speculation. In our opinion, this variable was not significant because it really does not relate to a person's socioeconomic status. It does not really affect the amount of money someone makes, so it would not impact their chances of being above the poverty line. Thus, it would not impact the percentage of the population of a particular state living in poverty.

As for the effect of the percentage of a state's population with a bachelor's degree, we obviously felt that it was significant enough to be included given our analysis of its impact on our model. However, the initial t-test would have had us reject this variable altogether. This could be due to correlation between this variable and other variables, which would hide the significance of education in our model. We can see traces of this when the p-value for X1 decreased by almost half going from our full model in Table 2 to the model without X5 in Table 3, showing that there was some correlation between X1 and X5. Study of these correlations would make a good topic for further research into the matter.

Looking back on the collection of our data and subsequent analysis, we felt that there was nothing too difficult about our efforts. However, we do think that it might have been better to perform all of our analysis together and in one sitting so as to avoid confusion between the various parts of the project. This may have led to some difficulties towards the end of our project. We also felt that it might have been better to wait and create the final project outline until we had gone over more topics in class, which would have given us more things to discuss.

Regarding the project itself, we enjoyed that it allowed us to review the semester and all that we learned throughout it. We also felt that it served as a good review for the final exam!

Finally, we were able to develop a better appreciation for the work that statisticians do and for the field as a whole.

Table A

	State	Y	$\widehat{Y}$	Residuals	Deleted(t)	Cook'sD	DFFIT
1	Alabama	16.80	17.4400	-0.6400	-0.5791	0.0068	-0.1834
2	Alaska	10.90	10.6160	0.2840	0.2677	0.0029	0.1201
3	Arizona	14.00	12.4861	1.5139	1.3576	0.0164	0.2886
4	Arkansas	17.20	17.5824	-0.3824	-0.3464	0.0027	-0.1141
5	California	12.80	11.5592	1.2408	1.1618	0.0407	0.4529
6	Colorado	9.60	8.2343	1.3657	1.2608	0.0357	0.4255
7	Connecticut	10.40	10.4445	-0.0445	-0.0403	0.0000	-0.0132
8	Delaware	12.50	11.6825	0.8175	0.7171	0.0031	0.1239
9	Florida	13.60	12.3928	1.2072	1.0717	0.0092	0.2150
10	Georgia	14.30	13.2800	1.0200	0.9007	0.0060	0.1732
11	Hawaii	8.80	12.4141	-3.6141	-4.0251	0.5515	-1.9207
12	Idaho	11.80	12.5167	-0.7167	-0.6425	0.0065	-0.1795
13	Illinois	12.10	11.9799	0.1201	0.1079	0.0002	0.0329
14	Indiana	13.10	14.2122	-1.1122	-0.9887	0.0092	-0.2145
15	Iowa	11.20	12.2385	-1.0385	-0.9710	0.0300	-0.3873
16	Kansas	12.00	12.2454	-0.2454	-0.2141	0.0003	-0.0367
17	Kentucky	16.90	17.2440	-0.3440	-0.3149	0.0027	-0.1148
18	Louisiana	18.60	17.3846	1.2154	1.1120	0.0252	0.3561
19	Maine	11.60	12.7686	-1.1686	-1.1013	0.0412	-0.4548
20	Maryland	9.00	9.8706	-0.8706	-0.7925	0.0135	-0.2586
21	Massachusetts	10.00	9.8329	0.1671	0.1589	0.0012	0.0754
22	Michigan	14.10	13.3870	0.7130	0.6275	0.0031	0.1245

23	Minnesota	9.60	10.0112	-0.4112	-0.3656	0.0018	-0.0950
24	Mississippi	19.70	19.4386	0.2614	0.2480	0.0027	0.1154
25	Missouri	13.20	13.8407	-0.6407	-0.5623	0.0023	-0.1061
26	Montana	13.00	12.4661	0.5339	0.4720	0.0024	0.1085
27	Nebraska	11.00	11.6616	-0.6616	-0.5982	0.0071	-0.1873
28	Nevada	12.90	12.4632	0.4368	0.4117	0.0068	0.1833
29	New Hampshire	7.60	9.1132	-1.5132	-1.4101	0.0487	-0.4990
30	New Jersey	9.50	10.6921	-1.1921	-1.1013	0.0302	-0.3892
31	New Mexico	19.50	16.9669	2.5331	2.4701	0.1434	0.8936
32	New York	13.60	11.9873	1.6127	1.4818	0.0381	0.4422
33	North Carolina	14.00	13.6435	0.3565	0.3108	0.0005	0.0511
34	North Dakota	10.70	10.6828	0.0172	0.0172	0.0000	0.0100
35	Ohio	13.90	13.4892	0.4108	0.3611	0.0012	0.0753
36	Oklahoma	15.60	15.4279	0.1721	0.1528	0.0003	0.0393
37	Oregon	12.60	10.6584	1.9416	1.8623	0.1018	0.7328
38	Pennsylvania	12.20	13.7529	-1.5529	-1.3908	0.0153	-0.2798
39	Rhode Island	12.90	12.4712	0.4288	0.3776	0.0014	0.0818
40	South Carolina	15.30	15.0731	0.2269	0.1986	0.0003	0.0382
41	South Dakota	13.10	12.9474	0.1526	0.1389	0.0005	0.0485
42	Tennessee	15.30	15.9662	-0.6662	-0.5908	0.0040	-0.1407
43	Texas	14.90	13.3219	1.5781	1.4490	0.0368	0.4341
44	Utah	9.00	10.4571	-1.4571	-1.5226	0.1823	-0.9685

45	Vermont	11.00	11.1027	-0.1027	-0.0963	0.0004	-0.0422
46	Virginia	10.70	10.8496	-0.1496	-0.1339	0.0003	-0.0392
47	Washington	10.30	9.6765	0.6235	0.5676	0.0074	0.1915
48	West Virginia	17.80	19.2935	-1.4935	-1.4652	0.1016	-0.7219
49	Wisconsin	11.00	12.0428	-1.0428	-0.9475	0.0172	-0.2928
50	Wyoming	11.10	10.9903	0.1097	0.1018	0.0004	0.0415

Critical Value

Deleted t :

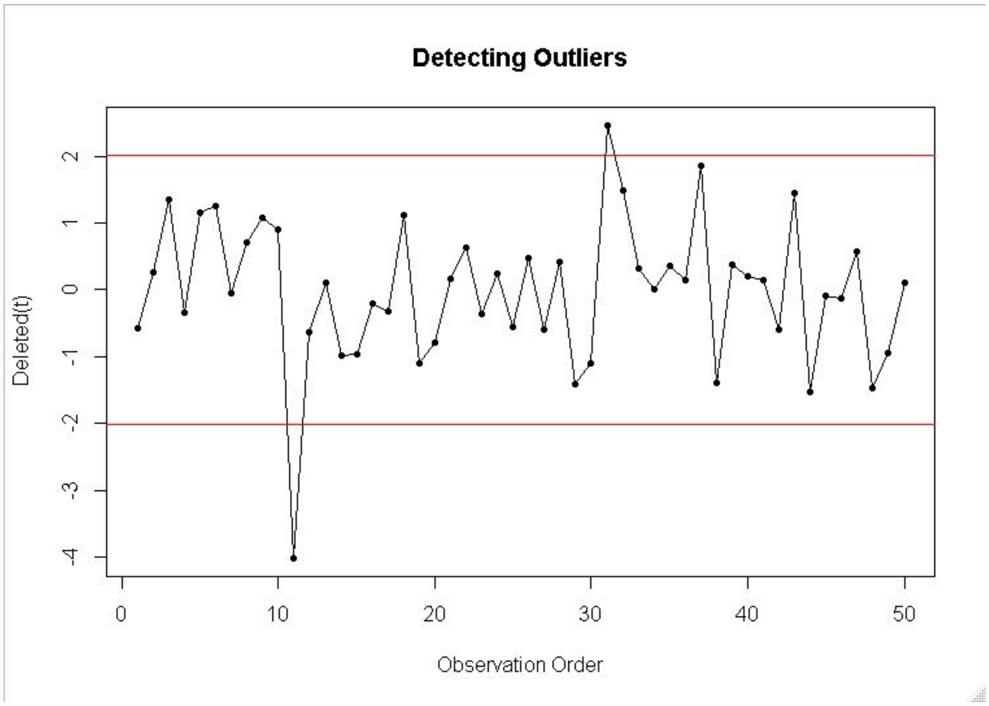
- $\pm t_{n-p}$ $\alpha = 0.05$, n=50, p= 5
- $\pm t_{45} = \pm 2.0141$

Cook's Distance :

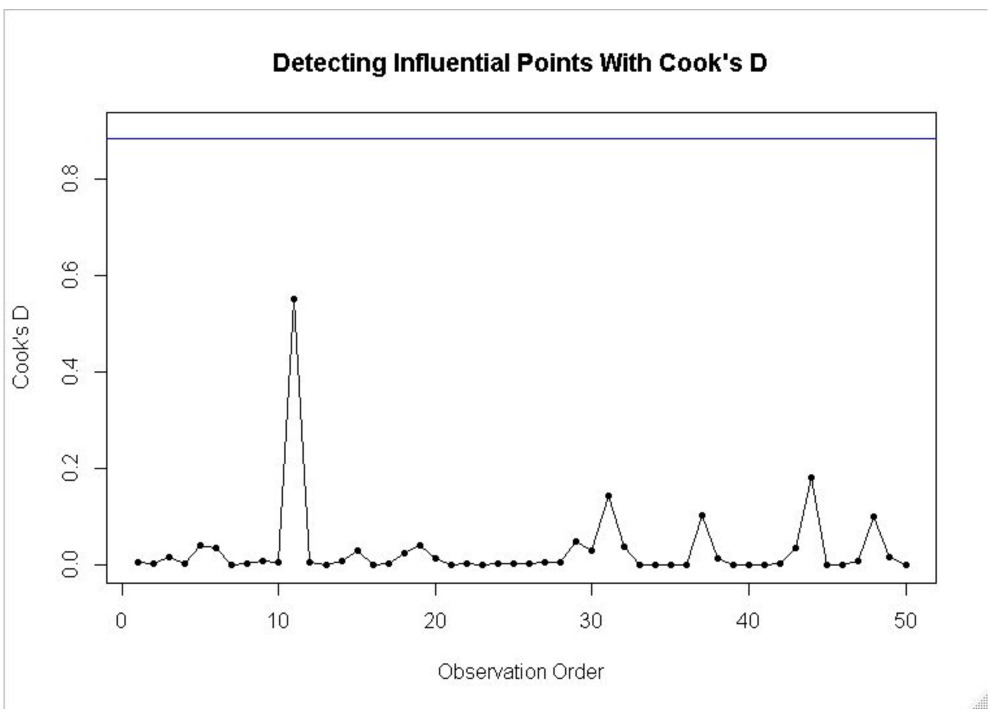
- $D_i > F_{0.50,\, p,\, n-p}$, n=50, p=5
- $F_{0.50,5,45} = 0.8835$

DFFIT :

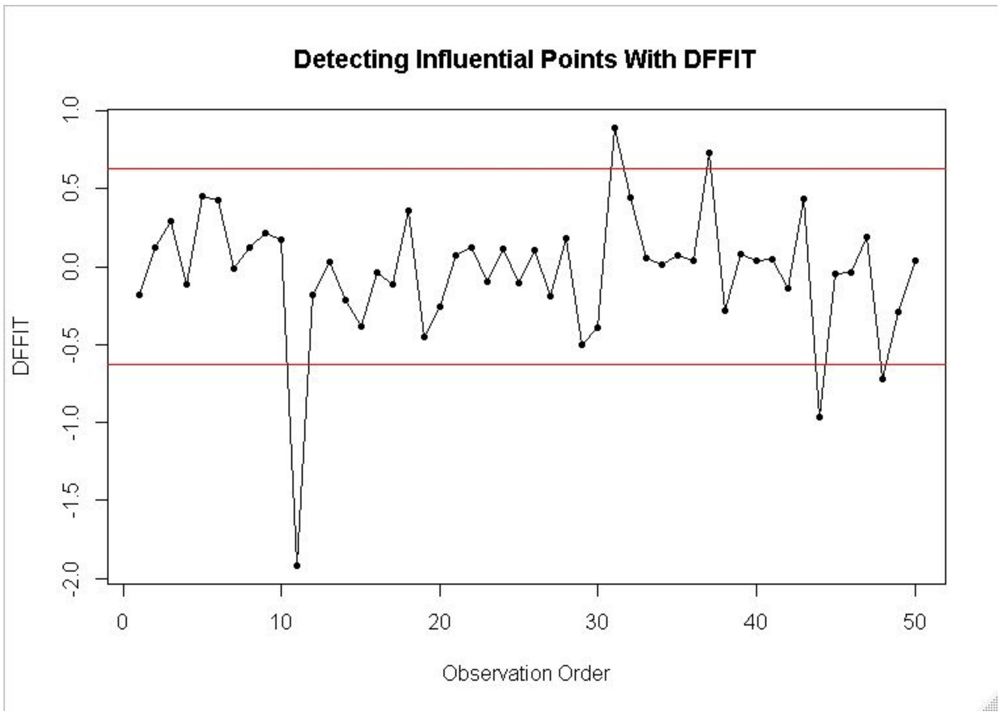
- $|DFFIT| > 2\sqrt{\frac{p}{n}}$, n=50, p=5
- $2\sqrt{\frac{p}{n}}$ =0.6325



Detecting Outliers
Deleted(t)
2
1
0
-1
-2
-3
-4
0
10
20
30
40
50
Observation Order



Detecting Influential Points With Cook's D
Cook's D
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
Observation Order



Detecting Influential Points With DFFIT
DFFIT
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
0
10
20
30
40
50
Observation Order