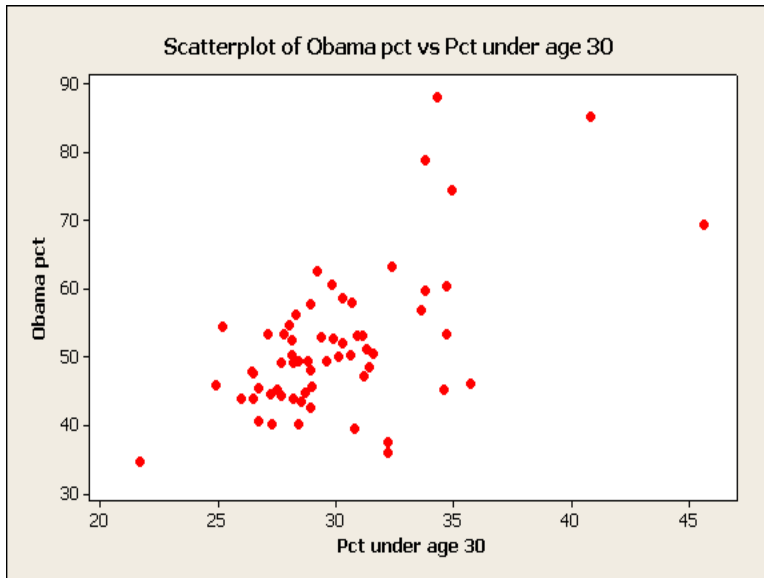


Youth and the 2008 presidential election

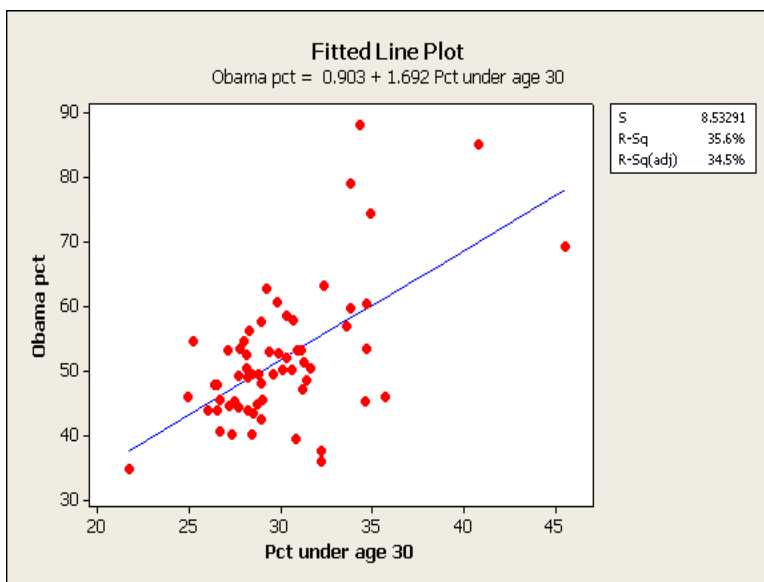
It is well-known that young voters had a large effect on the 2008 presidential election. Indeed, exit polls showed that voters aged 18 to 29 preferred Barack Obama to John McCain by a margin of 66 to 32 percent, a much wider margin between the candidates than for any other age group. If this was a one-time occurrence it is interesting, but if this reflects an underlying pattern of young voters preferring candidates “like” President Obama, it could have far-reaching implications about the future of the country.

From a more practical point of view, this question is also important for politicians, and those who try to get them elected. If the relationship between voting pattern and the age profile of a community is well-understood, that provides information that can be used to guide the strategic placement of money, manpower, and so on. We can explore this possibility by looking at the relationship between the percentage of the voters who voted for President Obama and the percentage of the population under the age of 30 (the latter figure is estimated by the Census Bureau) for the 62 counties in New York State. As we do this, we should recognize that any inferential statements only make sense if we view this observed relationship as a snapshot reflecting an ongoing age/voter preference process.

Here is a scatter plot of the Obama vote versus the percentage of the population under 30 for the New York counties:



It is clear that there is a direct relationship between the two variables, although it is not that strong. A fitted line plot reinforces this:



Here is regression output for this model.

Regression Analysis: Obama pct versus Pct under age 30

The regression equation is

$$\text{Obama pct} = 0.90 + 1.69 \text{ Pct under age 30}$$

Predictor	Coef	SE Coef	T	P
Constant	0.903	8.885	0.10	0.919
Pct under age 30	1.6919	0.2937	5.76	0.000

S = 8.53291 R-Sq = 35.6% R-Sq(adj) = 34.5%

Analysis of Variance

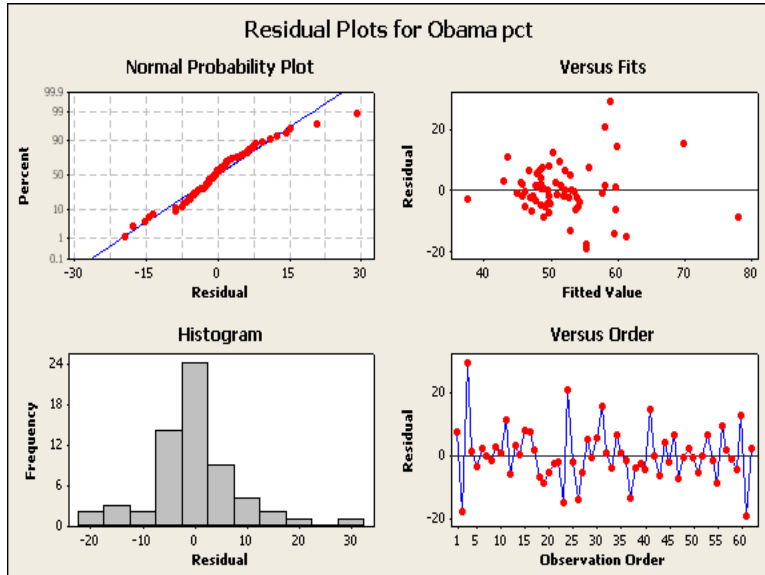
Source	DF	SS	MS	F	P
Regression	1	2416.7	2416.7	33.19	0.000
Residual Error	60	4368.6	72.8		
Total	61	6785.3			

Roughly 35% of the variability in Obama vote percentage is accounted for by the youth population percentage. The t -test for the slope and the F -test are of course equivalent here, and they imply that there is a strongly statistically significant relationship between the percentage of voters voting for Obama and the percentage of the population under 30, as the p -value is quite small. The estimated intercept is 0.903, which would correspond to the estimated Obama vote percentage for a county with no people under the age of 30. Since that is impossible, this number is not meaningful in this case. The estimated slope of 1.69 implies that an increase of one percentage point in the percentage of people under the age of 30 in a county is associated with an estimated expected increase of 1.69 percentage points in the percentage of people voting for Obama (note that these are *percentage points*, not *percents*).

Does this model provide useful predictive power? The standard error of the estimate is 8.5, implying (by the standard assumptions) that this model can predict the Obama vote percentage to within $\pm(2)(8.5) = 17$ percentage points roughly 95% of the time. Since the middle 95% of the Obama vote percentages cover roughly 43 percentage points (35%

to 78%; this comes from a boxplot of the distribution, which I have not included), the reduction in the predictive range is noticeable, but not overwhelmingly impressive.

How well does this model fit the data? Here is a four-in-one plot of residuals:



These plots don't look quite as nice as we would like. There is some evidence of nonconstant variance (although it is fairly weak), but there is also evidence of three unusual observations. The point to the left in the residuals versus fitted values plot is a potential leverage point, and corresponds to Hamilton County, which only has 21.7% of its population under the age of 30. Hamilton County is unusual, in that it is the least populous New York county, and lies completely within Adirondack State Park. It is also the most consistently Republican county in the state, although note that it is apparently not an outlier in this regression fit, since its low youth proportion is consistent with its low vote for Obama (also the lowest in the state).

There are also two apparent outliers in the data, which are apparent in the normal plot of the residuals. These are the Bronx and Brooklyn (Kings County), which had unusually high Obama votes given their youth population. These two counties had two of the top three Obama votes in the state, but it is interesting to note that Manhattan had a higher Obama percentage than Brooklyn, but was not flagged as an outlier because of its high youth population (more than 40% under the age of 30).

What should we do about this? Since we're interested in this relationship for the bulk of population, we should check to see if these unusual observations have had a strong effect on the regression. If they have, we should note that in any discussion we have about the model. Here is the output for the model omitting these three counties:

Regression Analysis: Obama pct versus Pct under age 30

The regression equation is

$$\text{Obama pct} = 9.31 + 1.39 \text{ Pct under age 30}$$

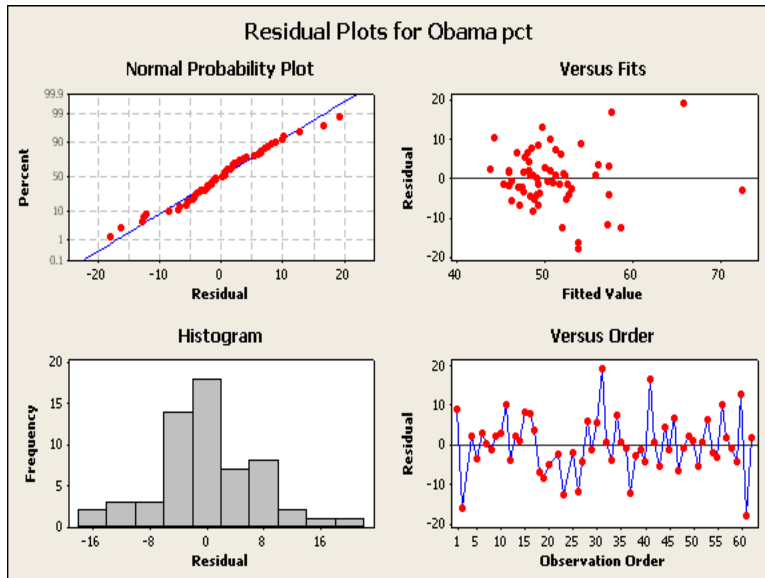
Predictor	Coef	SE Coef	T	P
Constant	9.315	8.004	1.16	0.249
Pct under age 30	1.3853	0.2647	5.23	0.000

S = 7.21106 R-Sq = 32.5% R-Sq(adj) = 31.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1424.6	1424.6	27.40	0.000
Residual Error	57	2964.0	52.0		
Total	58	4388.6			

The strength of the relationship hasn't changed very much (it has decreased slightly), but the coefficients have changed noticeably, with the intercept increasing and the slope decreasing. Still, of course, the basic character of the relationship has remained the same. The residuals look reasonable now, although there is still some evidence of nonconstant variance.



This is a situation where it is natural to consider the use of confidence and prediction intervals. Consider the case of Los Angeles County in California. Can this model help us to understand the Obama vote in a county like that? Its percentage of people under the age of 30 was 35.2. What does the regression model say about a county with that value? Here are confidence and prediction interval results:

Predicted Values for New Observations

New

Obs	Fit	SE Fit	95% CI	95% PI
1	58.076	1.659	(54.755, 61.397)	(43.259, 72.893)

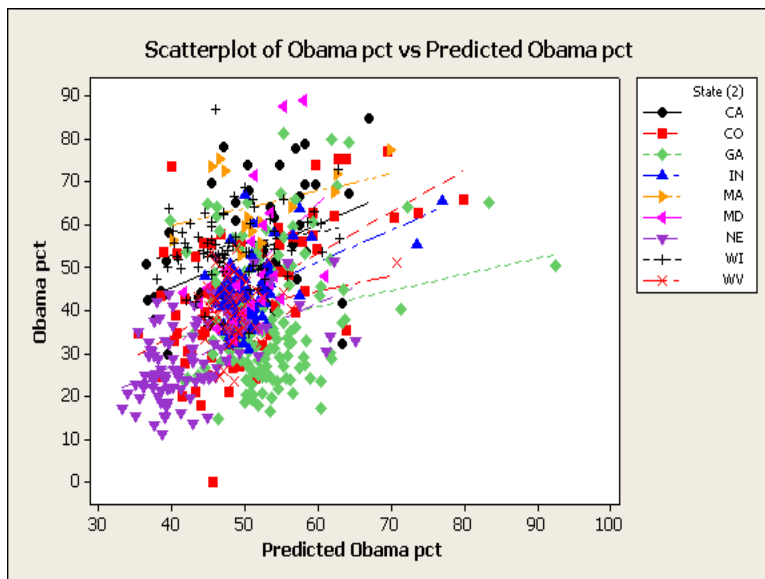
Values of Predictors for New Observations

	Pct
New	under
Obs	age 30
1	35.2

The confidence interval of (54.8%, 61.4%) corresponds to an interval estimate for the average Obama percentage for all counties that had (or could have, remembering that we are viewing these data as a snapshot of an ongoing process) a youth percentage of 35.2. What is probably of more interest is the prediction interval, which corresponds to our prediction of the Los Angeles Obama vote based on data from New York. The predicted

vote is 58.1%, with an interval (43.3%, 72.9%). The actual vote was 69.3%, so the interval does contain the actual value, but the width of the interval reinforces that this model is not precise enough to be very useful. Of course, a real model of this type would include many other demographic variables besides the youth proportion.

The plot below expands on this idea. I took the model above based on New York counties, and applied it to the counties in California, Colorado, Georgia, Indiana, Massachusetts, Maryland, Nebraska, Wisconsin and West Virginia (by “applied it” I mean that I just plugged the youth population figures for each of the new counties into the formula above, and got a predicted Obama vote for that county). The plot below is a scatter plot of the observed Obama votes versus the predicted Obama votes, with different characters for each state and the regression lines relating the observed and predicted values for each state superimposed. It is clear that the model does not do a great job of predicting overall, and there are clear state-to-state differences. The model is not too biased for California, Colorado, Maryland, Wisconsin, and West Virginia, but seriously underestimates the Obama vote in Massachusetts, and seriously overestimates it in Georgia and Nebraska, corresponding to the generally different level of support for Democrats in those states compared to New York.



Thus, overall, we can see that there is evidence of the relationship between a young population and voting for Obama, but the large state-to-state variation suggests that the model is unlikely to be very useful as a practical political tool.

Minitab commands

To get a four-in-one residual plot, click on **Graphs** while fitting a regression, and click the radio button next to **Four in one**.

Although it is possible to omit observations in a sample by simply highlighting them in the data worksheet and pressing the delete key, this is generally not advisable, since then the observation cannot be recovered without reopening the original file (and if you save the data before doing that, the observation is gone completely). A better approach is to create a subset of the worksheet that has the observations you want; this will create a new worksheet that can be analyzed, but the original worksheet will still be there as well. Click on **Data** → **Subset Worksheet**. You can give the new worksheet an identifying name if you like under **Name:**. Click the radio button next to **Specify which rows to exclude**, click the radio button next to **Specify rows:**, and enter the row numbers of the outliers in the associated box. Note that there is a good deal of flexibility in the subsetting; you can identify rows to include or exclude, identify them by some condition (for example, observations with values of a predictor greater than 10), or brush them on a scatter plot and identify them that way, in addition to specifying them by row number(s).