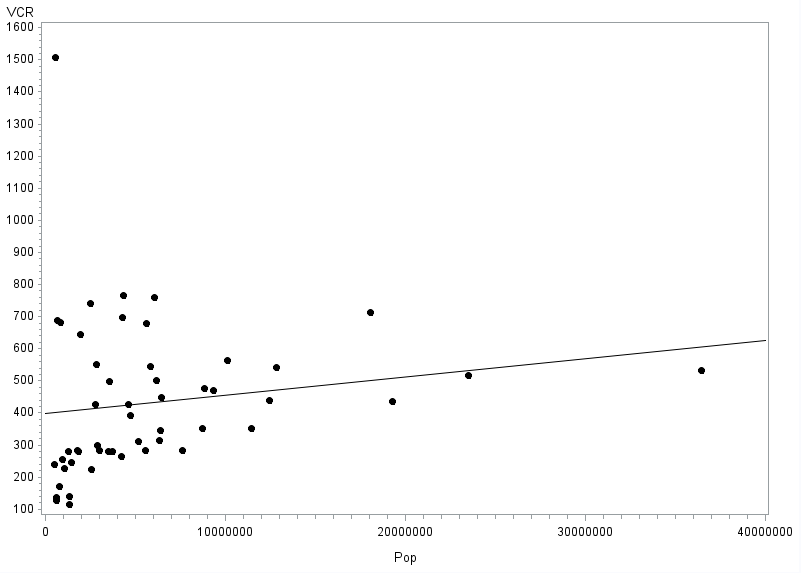
We are interested in determining factors for violent crime in the United States. One of the main factors often tested in determining a crime rate is population size. In this investigation, we will take observational data from the US census and crime statistics to determine if a relationship does indeed exist between population size and crime rate.

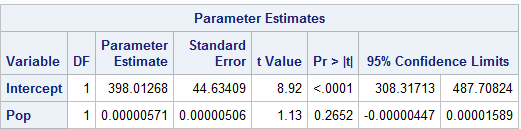
The initial scatter plot does not hold much evidence of a significant slope, however, there is evidence of a slight positive relationship:



The initial regression equation is equal to:



The slope of this equation is insignificant at p= .27 with a 95% confidence interval of -0.00000447, 0.00001589. These two factors indicate a lack of a relationship between population and crime rate:



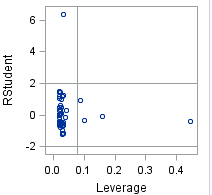
The assumptions for the linear regression model are linearity, constant variance across all values of Y given X, independence and lastly, normally distributed sub populations for all values of Y given X.

A residual plot should have a random characteristic, specifically, a random ‘cloud’ of sorts. We should not see any noticeable pattern in the residuals for predicted vs actual values. Further, we should also notice that a QQ plot is relatively straight, indicating normally distributed residuals at each Y given X.

If constant variance was violated, we’d likely see an increasing or decreasing pattern in our residuals as we move further away from the mean of X, specifically in the studentized and standardized residuals.

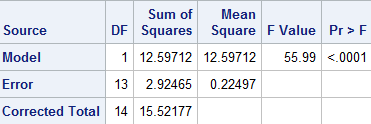
There is one extreme outlier: Washington DC with a violent crime rate of 1508 per 10,0000 and a population of 581K people.

Washington DC can be seen explicitly in a leverage plot at the top of the display as an outlier, along with California being extremely high leverage (far right of chart) due to its population size:

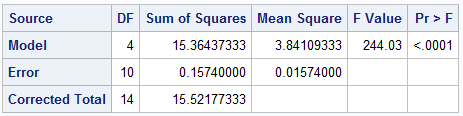


We are interested in modeling the decay in chemical concentration over time. 15 observations of chemical decay were taken after being randomly assigned to groups by time.

Based on the ANOVA table result:



By adding one variable to our model, we reduce our total sum of squares significantly (by over 12), along with an extremely large F of 55.99. Indicating that the additional variable (time) affords a much better model than the equal means model for this data set. Further, based on a hypothesis test of the slope, time is significant at p <.0001, confirming our ANOVA results.



The ANOVA table above shows an analysis of variance for the reduced vs full (separate means) models, using times as groups in the chemical decay experiment. Based on a tukey test of means, all group means are significantly different from one another except for a comparison of group 9 and group 7. The separate means model reduces the sum of squares by 15.364 – almost 99%.

The two previous models have the same total sum of squares. SST is based on the equal means model in both analyses. 15.52 is the total sum of squares based on the equal means model, while the errors represent the respective full (linear regression, separate means) models. The ‘model’ variable represents the amount of variance explained by the full model in the ANOVA table.

Each error MSE is different because of differences in models. Linear regression is using two parameters (d.f. = n-1), while the separate means model is using 5 parameters (d.f. = n-1). The generalization from the equal means model to linear regression reduces the sum of squares by 12.6, whereas the generalization of the equal means model to the separate means model in the ANOVA reduces the sum of squares by 15.4. The separate means model is more complex than the linear model and explains the variability between groups much better than the linear model in the case of chemical decay over time.

Pure error in the ANOVA model can be assumed ‘pure’ because each group mean is used in the model and differences from those means are only based on differences in the group mean. The regression model approximates group means for each x value using less parameters, using slope and intercept parameters instead of group means. Because of the lack of group means, the regression model not only has to account for intra group error, but also for how other values in different groups may affect the ‘fit’ of the line, or the ability to explain the variability in the response given a certain x value.



Based on a lack of fit f test with an F of 58.6 and a significant p, we can conclude the linear model is not a good fit for explaining the variability between group means. We should consider another variable or transformation and adjust.

SAS CODE:

**data** crime;

infile '\\Client\C$\Users\Patrickcorynichols\Desktop\Data Science\Stats\Data Sets\crime2006.csv' FIRSTOBS = **2** DLM = ',';

INPUT State $ Pop VCR Murder Rape Rob Assault Prop Burg Larc Vehicle;

logpop = log(pop);

logvcr = log(vcr);

**RUN**;

**PROC** **PRINT** data = crime;

**RUN**;

**PROC** **GPLOT** data = crime;

PLOT vcr\*pop;

SYMBOL1 V = Dot C = BLACK I= RL L =**1**;

**RUN**;

/\*x needs to be transformed, values bunched\*/

**PROC** **GPLOT** data = crime;

PLOT vcr\*pop;

**RUN**;

**PROC** **REG** data = crime;

MODEL vcr=pop / clb cli clm lackfit;

**RUN**;

**data** chemical;

input time conc;

datalines;

1 2.57

1 2.84

1 3.10

3 1.07

3 1.15

3 1.22

5 0.49

5 0.53

5 0.58

7 0.16

7 0.17

7 0.21

9 0.07

9 0.08

9 0.09

;

**RUN**;

**PROC** **REG** data = chemical;

MODEL conc = time / CLM CLI CLB lackfit;

**RUN**;

**PROC** **ANOVA** data = chemical;

CLASS time;

MODEL conc = time;

MEANS time / tukey cldiff;

**RUN**;