

Spatial Data Analysis in Archaeology
Anthropology 589b

Fraser D. Neiman
2.18.07

University of Virginia
Spring 2007

**Evaluation Goodness of Fit in Trend-Surface Analysis:
r-square, sums of squares, mean squares, and F statistics**

As we have seen, it can be helpful to think about spatial variation as being comprised of three additive components, a global trend, spatially dependent local variation, and independently and identically distributed noise. If we think this model might be useful in a particular case, how can we go about identifying the trend? There are many ways of doing this. The oldest and most widespread is trend-surface analysis. Trend-surface analysis attempts to extract the global trend underlying a spatial distributed variable (z) by predicting that variable's values as a function of its Cartesian coordinates in space. The function in question is a polynomial equation of some degree. The first degree is the simplest case. Here the predictions are based the X and Y coordinates of the location in question (*e.g.* the northings and eastings). In the second-degree case the predictions are a function of the X and Y coordinate and their squares and cross products. Third, fourth, fifth, and higher-degree equations are possible as well. In deciding what order of trend surface is the "right" one, it would be helpful to know how well the predictions account to the actual values of z . These notes describe how this is done.

Trend surface analysis is one flavor what's known in statistics as "the general linear model." This phrase denotes a single framework that can accommodate a variety of statistical methods, including *t*-tests, regression, analysis of variance, analysis of covariance. What unites these different methods under the GLM banner is that can all be seen as attempts to predict the values of a dependent variable as a function of one or more independent variables. They all assume that errors in the model predictions of the value of the dependent variable follow a Gaussian ("normal") distribution. They also assume that predictions are linear functions of the independent variables. So GLMs all look like this:

$$\hat{z} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \dots + b_pX_p$$

Here the *b*'s are the coefficients estimated as part of the model. \hat{z} is the predicted value of the dependent variable. The *X*'s are the *p* independent variables. (Geek note: if we relax the Gaussian-distribution assumption, we end up with *generalized* linear models, GLIM's). The math behind GLMs is designed so that the estimates of the values of the *b*'s that result produce predictions of \hat{z} that minimize the sum of squared deviations from the predicted values to the actual ones. I

Two issues are critical to understanding GLMs. The first is how the goodness of fit of the model to the data is judged: just how good are the predictions? This is usually measured using r-square, the proportion of the variation in the dependent variable accounted for by the GLM. The second issue is the statistical significance of the fitted model: does the model account for more variation than we could expect as a result chance alone?

An odd question? It simply means this: Imagine that we had a very, very large population of observations on the variables in our model. Imagine too that in the population as a whole the independent variables included in our model had no predictive relevance to the dependent variable at all. Suppose we drew multiple, independent samples from the population of the size we actually have. Suppose that we then estimated the parameters of our model, based on each sample. The question of statistical significance of the model boils down to an estimate of the proportion of times, under repeated random sampling from such a population, we would get a model that fit as well as ours actually does (has as high an r-square), simply as a result of sampling error.

R-square

How can we measure variation? A simple way is the sum of squared deviations (SS) of the \hat{z} value (where \hat{z} is the dependent variable). But deviations have to be deviations FROM some quantity. What the deviations are from depend on what kind of variation we want to measure.

When we are trying to build a GLM to predict the value of some variable (\hat{z}) based on the value or one or more independent variables, there are three kinds of variation that are of interest:

1. Total SS: the sum of squared deviations of the \hat{z} values from their mean. This is the total variation in the \hat{z} variable.
2. Model SS: the sum of squared deviations of the values of \hat{z} predicted by the model from the mean of \hat{z} . This is the amount of variation accounted for by the model.
3. Residual (or Error) SS: the sum of squared deviations of the \hat{z} values from the values for predicted by the model. This is the quantity that is minimized in coming up with the estimates of the b 's for the model.

In GLMs, it turns out that

$$\text{Total SS} = \text{Model SS} + \text{Error SS}.$$

Cool, huh?

So it makes sense to summarize how well the \hat{z} values predicted by the model match the actual \hat{z} values by computing:

$$\text{SS Model} / \text{SS Total}$$

This is r-squared -- the proportion of the total variation accounted for by the model

You can also compute r-square like this:

$$1 - (\text{Error SS} / \text{Total SS}).$$

Can you see why?

Statistical Significance

Now that we know how to measure variation in general and the proportion of variation accounted for by a GLM in particular, the next question is: is the amount of variation accounted for by the model greater than what I could expect to result from sampling error – or from chance alone? In other words, is the model doing any real work, or just coasting on chance effects?

The answer to that question depends on four quantities:

1. How much variation does the model account for. The more variation accounted for the less likely chance alone is responsible for the fit of the model (summarized by r -squared).
2. How much variation was not accounted for, the more variation NOT accounted for, the more likely chance alone is responsible for the fit of the model.
3. The sample size (number of observations on z). The greater the sample size, the less likely chance effects account for the fit of the model. Bigger samples mean less sampling error, right?
4. The number of parameters contained in the model (the number of independent variables for which the model estimates b coefficients). The greater the number of parameters we estimate to get the fit, the more likely those parameters are capitalizing on chance effects. To think about this, realize you can always get a perfect fit (r square = 1) by including as many parameters in the model as observations.

1 and 2 are measured using SS as described above. 3 and 4 are measured using the concept of degrees of freedom.

Just as there is a SS associated with the total variation in z , the model, and the residuals, There are degrees of freedom associated with each:

1. Total DF: The number of observations -1. We lose one DF because in order to estimate the Total SS, we had to estimate a parameter: the mean of z . Say there are 41 observations. If we have our estimate of the mean, we can perfectly predict the value of the 41'st observation, given knowledge of only 40 of them. So there are 40 degrees of freedom associated with the deviations from the mean.
2. Model DF: The number of independent variables in the model for which you have b coefficient estimates. (For a linear trend surface, this is 2. For a quadratic, it's 5, for a cubic it's 9.)

3. Error DF (or Residual DF): This is the number of observations minus the number of parameters you had to estimate to compute the residuals. To get the residuals, we had to estimate the b coefficients and the mean of z . So Error DF = Number of observations - Model DF - 1. The error DF is the number of observations that are free to vary, once we have made the predictions. Say there are 41 observations and 2 parameters estimated by our model. The error DF is 39 (=41-2-1). This means that, if we know the mean and the 2 parameters, we can perfectly predict the value of z for each of the (3) remaining observations, once we know the value of the other 39 of them.

Now it turns out that:

$$\text{Total DF} = \text{Model DF} + \text{Error DF}.$$

Now we have two sets of quantities: the three sums of squares (SS) and their corresponding degrees of freedom. Here is the final step. So far we have just measured variation in terms of sums of squared deviations. This is helpful. But it does not take into account the number of degrees of freedom we had to burn to account for that variation. It's kind of like saying you bought 4 bananas – good to know. But it is also good to know how much each banana cost.

In the banana case, you divide the total number of bananas by the cost. In the sum of squares case, we divide the amount of variation accounted for by the model by the number of degrees of freedom we had to spend to build the model (the Model DF). In doing this division, we have converted the SS into a variance (recall that a variance is just a MEAN of squared deviations). This variance is special, so it has a special name "the mean square" – or the mean of the sum of squares accounted for by the model. Actually it's the mean amount of variation accounted for by each parameter estimated in the model.

Now we are almost there. The question is: Is the model mean square (variance) bigger than what we would expect as a result of sampling error? Answering this question requires estimating the size of the variance that we could expect to account for as a result of sampling error.

It turns out that we already have the ingredients to build an estimate of the variance that is likely to arise as a result of sampling error in a sample like the one we have, given the model we are trying to evaluate. One ingredient of the residual sum of squares – the variation NOT accounted for by the model. The other ingredient is the residual degrees of freedom. By dividing the Error SS by the Error DF, we get an error variance – the mean amount of variation from the predictions of the model, associated with those observations of z , whose values are free to vary, once we have made the predictions.

So we have two variances: the Model variance and the Error variance, and the question is, is the Model variance significantly greater than the Error variance? To answer this we take the ratio of the two:

$$F = \text{Model variance} / \text{Error variance}$$

or

$$F = \text{Model mean square} / \text{Error mean square}.$$

This new quantity is a ratio of variances, usually abbreviated with the letter F (in honor of R.A. Fisher, the guy who figured all this out in the 1920's). Our question now becomes: is the observed value of F greater than what we should expect to arise as a result of sampling error. To answer this question, we need to check out the sampling distribution of the F statistic.

This distribution gives the probability of getting a variance ratio as big, or bigger, than we one we actually have, as a result of chance. The F distribution just like any other distribution with which you may be more familiar. For example, the binomial distribution gives the probability of getting K success in N trials, given an overall probability p of a success. Now the shape of the binomial distribution is governed by two parameters: N and p. The shape of the F distribution is also governed by two parameters: the DF associated with the numerator variance and the DF associated with the denominator variance.

Given an estimate of F (the ratio of the model to error mean squares) and the degrees of freedom associated with its numerator and denominator, the F distribution tells us the probability of getting an F value that big, or bigger, by chance. That quantity is the thing we have been looking for: the probability that the fit achieved by the model (summarized by r-square) could have arisen by chance.

Here is how SAS summarizes all this information in the results from Proc GLM (General Linear Model). The model is a 1-degree trend surface for the house diameter data from site AR 5. So the model looks like this:

$$Diameter = northing + easting$$

The GLM Procedure

Dependent Variable: diam

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.31132446	1.65566223	2.85	0.0611
Error	140	81.29448253	0.58067488		
Corrected Total	142	84.60580699			

R-Square	Coeff Var	Root MSE	diam Mean
0.039138	24.79722	0.762020	3.073007

The F value is the ratio of the two mean squares or variances: $1.66 / .54 = 2.85$. Given 2 and 140 degrees of freedom, the probability of getting an F value this big or bigger is .06. So there's only about a 1 in 20 chance that goodness of fit of the model to the data, as summarized in the r-square value of .039 could have arisen by chance alone. So we have a marginally significant result – in the statistical sense. But very little predictive power.

The GLM framework also allows us to partition the variation that is accounted for by the model into components, one associated with each of the independent variables. Here again the variation is measured in SS. This makes it possible to estimate how much of the variation in \hat{z} is accounted for by each of the independent variables. And it is possible to test the hypothesis that we could have gotten a SS this big by chance alone. We proceed exactly as

we did before: we compute a corresponding mean square (or variance) by dividing the SS by the appropriate DF. We then form an F-ratio of the independent-variable mean square with the Error mean square ($5.70 = 3.31 / .58$). And we compute the corresponding probability from the F distribution with (in this case) 1 and 140 DF:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
northing	1	3.31090958	3.31090958	5.70	0.0183
easting	1	0.00041488	0.00041488	0.00	0.9787

Here we encounter an ambiguity of sorts. The amount a variation (SS) that each independent variable accounts for *depends on the order in which it appears in the model equation*. This makes intuitive sense if you think about each successive independent variable in the equation only being able to account for that variation which the previously entered variables have not accounted for. SAS calls the SS that are computed for each independent variable, *in the order specified in the GLM equation* (specified in the Proc GLM Model statement) the Type-I SS. These quantities are shown above.

But it is also almost always helpful to know how much variation each variable would account for if it were entered LAST, after all the other variables have had a shot. SAS calls this the Type-III SS. Again, the corresponding mean squares, DF, F-ratios and p-values are given as well:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
northing	1	1.95012383	1.95012383	3.36	0.0690
easting	1	0.00041488	0.00041488	0.00	0.9787

So far we have talked glibly about variation accounted for – and not accounted for– without actually looking under the hood, to check out the parameter estimates that are doing the predicting and accounting for. Here they are:

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	0.2252949403	1.76269471	0.13	0.8985
northing	0.0035544790	0.00193960	1.83	0.0690
easting	-.0000246172	0.00092096	-0.03	0.9787

The t statistics and corresponding p values are tests of the hypothesis that the actual values of the b coefficients in the underlying population are 0. In other words, that the independent variable has no effect, *once the effects of all the other independent variable have been taken into account*. Note these probability values are identical to those associated with the Type-III SS.

Now the problem with the significance tests in the traditional GLM framework is that they all depend on the assumption that the errors are identically and *independently* distributed. This assumption is built into the use of error mean square in the denominator of the F-ratio. Spatial autocorrelation among the residuals from the model's predictions is a violation of the independence assumption. It means that the Error DF overstates the actual number of degrees of freedom in the data – if the residuals are correlated with one another, the number of values that are free to vary is smaller than it would be if the residuals were independent.

As a result, the estimate of the error mean square is too small and the F-ratio is inflated, which means that results that look significant in the statistical sense, may not be.

A Final Note: Loess

A loess model is not a GLM. In fact it is composed of multiple *local* GLM.s. Because the loess predictions do not come from a single GLM, the GLM framework does not apply, strictly speaking. For example it is no longer the case that

$$SS \text{ Total} = SS \text{ Model} + SS \text{ Error.}$$

In fact, because the predictions in loess are based on multiple, local GLMs, $SS \text{ Total} > SS \text{ Model} + SS \text{ Error}$. The reason lies in the local fits. Local fitting means that each of the GLMs that comprise the loess surface minimizes the Error SS, and thereby maximizes the Model SS, *in its local neighborhood*. As a result, the predicted values are optimized for the neighborhood, NOT the entire surface.

But this does not change the fact that it still makes sense to compare the Error SS from the loess surface to the Total SS. So the proportion of variation NOT accounted for the loess surface is $\text{Error SS} / \text{Total SS}$.

And this means it is still possible to compute a *pseudo* r-square value that summarizes the goodness of fit of the entire surface:

$$1 - (\text{Error SS} / \text{Total SS})$$

We call this a "pseudo r-square" to distinguish it from the real thing, in the GLM case, where $SS \text{ total} = SS \text{ model} + SS \text{ Error}$ and $r\text{-square} = \text{Model SS} / \text{Total SS}$.

Like I said, Cool, huh?