# What they don't tell you about regression analysis

*Lauren Rosenshein and Lauren Scott*

Regression analysis is used to understand, model, predict and/or explain complex phenomena. It helps you answer *why* questions like "Why are there places in the United States with test scores that are consistently above the National average?" or "Why are there areas of the city with such high rates of residential burglary?" Regression analysis is all about explaining something, like childhood obesity for example, using a set of related variables such as income, education, or accessibility to healthy food. Typically, regression analysis helps us answer these why questions so that we can do something about them. If, through regression analysis, we discover that childhood obesity is lower in schools that serve fresh fruits and vegetables at lunch, for example, we can use that information to guide policy and to make decisions about school lunch programs. Likewise, if we know the variables that help to explain high crime rates, we can use that information to make predictions about future crime, and consequently allocate crime prevention resources more effectively. These are the things they do tell you about regression analysis.

What they don't tell you about regression analysis is that it isn't always easy to find a set of explanatory variables that will allow you to answer your question or to explain the complex phenomenon you are trying to model. Childhood obesity, crime, test scores and almost all of things that you might want to model using regression analysis are complicated issues that rarely have simple answers. Chances are if you have ever tried to build your own regression model, this is nothing new to you. But they *also* don't tell you about the techniques you can use to help you find a good model.

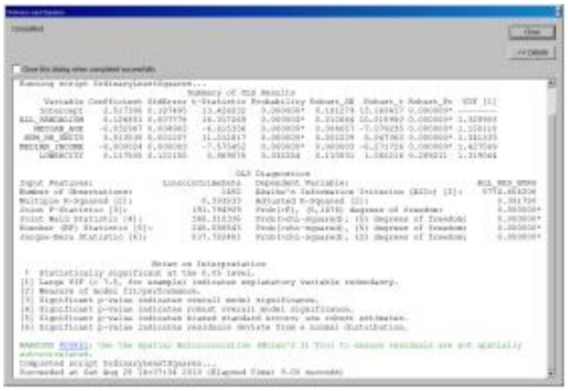
When you run Ordinary Least Squares (OLS) regression from the Spatial Statistics toolbox you are presented with a set of diagnostics that can help you figure out why you aren't finding a good model. In addition, the OLS documentation includes a guide to the 6 things that you need to check before you can feel confident that you have a properly specified model (a model you can trust). Those 6 checks, and the techniques that you can use to solve some of the most common regression analysis problems, are little-known resources that will make your work easier.

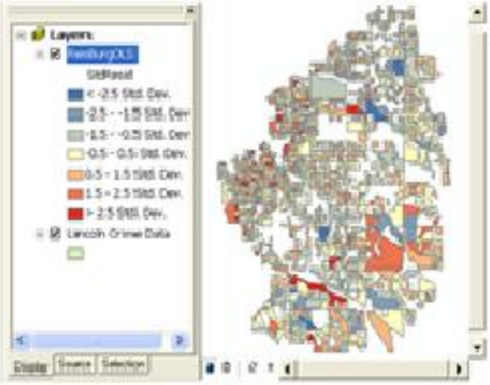
## Getting Started

The first thing you do when you run a regression analysis is choose the variable that you want to understand, predict, or model. This variable is known as the dependent variable. In the examples described above, childhood obesity, crime, and test scores are the dependent variables being modeled.

The next thing that you have to do is decide which factors you think might help explain your dependent variable. These variables are known as your explanatory variables. In the childhood obesity example, the explanatory variables would be things like income, education, and accessibility to healthy food. You will need to do your research in order to identify all of the explanatory variables that might be important here. You'll want to consult theory and existing research, talk to experts, and always rely on your common sense. This preliminary research increases your chances of finding a good model.

Once you've chosen your dependent variable and your candidate explanatory variables, the next step is to run your analysis. In ArcGIS, the regression tools can be found in the Spatial Statistics toolbox, in the Modeling Spatial Relationships toolset. You will always start with Ordinary Least Squares regression because it provides you with the important diagnostic tests that let you know if you've found a good model or if you still have some work to do. There are two important outputs from the OLS tool. The first output is largely numeric, and includes a lot of the diagnostics that you will use when going through the 6 checks below. The other important output is the map of the regression residuals; these are the over and under predictions from your model. Analyzing the residual map is an important step in finding a good model.



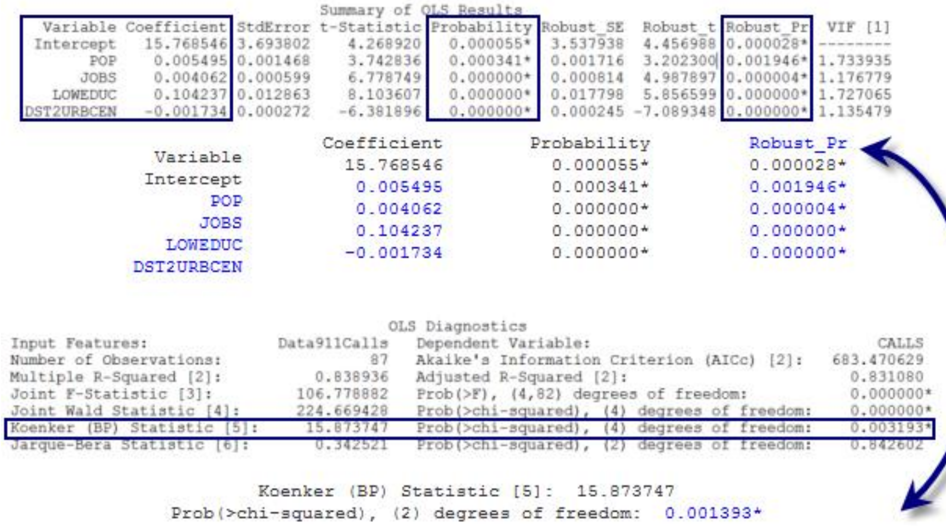


## The 6 Checks

### *Check 1: Are these explanatory variables helping my model?*

If you follow the steps outlined above, you will have already consulted theory and existing research, and have identified a set of candidate explanatory variables. You'll have good reason for including each of these variables in your model. When you run your model, you will find that some of your explanatory variables are statistically significant and some are not. *How will you know that?* Well, the OLS tool calculates a coefficient for each explanatory variable in the model, and then performs a statistical test to determine if that variable is helping your model or not. The statistical test computes the probability that the coefficient is actually zero. If a coefficient is zero (or very near zero), the associated explanatory variable has very little impact on your model; it is not helping your model. The smaller the probability is, however, the smaller the likelihood that the coefficient is, in fact, zero. When the probability is smaller than 0.05, an asterisk next to the probability on the OLS summary report indicates the associated explanatory variable *is* helping your model (its coefficient is statistically significant at the 95% confidence level). So you are looking for explanatory variables associated with statistically significant probabilities (look for the asterisks).

The OLS tool computes both the probability and the robust probability for each explanatory variable. It is not unusual with spatial data for the relationships you are modeling to vary across your study area. This is called non-stationarity. When the relationships you are modeling are non-stationary, you can *only* trust the robust probabilities to determine if an explanatory variable is statistically significant or not. *How will you know if the relationships in your model are non-stationary?* Well, another diagnostic reported by OLS is the Koenker test for non-stationarity; if there is an asterisk next to the Koenker p-value – the relationships you are modeling do exhibit statistically significant non-stationarity, so be sure to consult the robust probabilities.



```
                                   Summary of OLS Results
   Variable Coefficient StdError t-Statistic Probability Robust_SE  Robust_t Robust_Pr  VIF [1]
  Intercept   15.768546 3.693802    4.268920   0.000055*  3.537938  4.456988 0.000028* --------
        POP    0.005495 0.001468    3.742836   0.000341*  0.001716  3.202300 0.001946* 1.733935
       JOBS    0.004062 0.000599    6.778749   0.000000*  0.000814  4.987897 0.000004* 1.176779
    LOWEDUC    0.104237 0.012863    8.103607   0.000000*  0.017798  5.856599 0.000000* 1.727065
DST2URBCEN    -0.001734 0.000272   -6.381896   0.000000*  0.000245 -7.089348 0.000000* 1.135479
```

```
                 Coefficient       Probability      Robust_Pr
     Variable     15.768546         0.000055*       0.000028*
    Intercept      0.005495         0.000341*       0.001946*
          POP      0.004062         0.000000*       0.000004*
         JOBS      0.104237         0.000000*       0.000000*
      LOWEDUC     -0.001734         0.000000*       0.000000*
   DST2URBCEN
```

```
                                         OLS Diagnostics
Input Features:                  Data911Calls   Dependent Variable:                            CALLS
Number of Observations:                    87   Akaike's Information Criterion (AICc) [2]: 683.470629
Multiple R-Squared [2]:              0.838936   Adjusted R-Squared [2]:                     0.831080
Joint F-Statistic [3]:             106.778882   Prob(>F), (4,82) degrees of freedom:        0.000000*
Joint Wald Statistic [4]:          224.669428   Prob(>chi-squared), (4) degrees of freedom: 0.000000*
Koenker (BP) Statistic [5]:         15.873747   Prob(>chi-squared), (4) degrees of freedom: 0.003193*
Jarque-Bera Statistic [6]:           0.342521   Prob(>chi-squared), (2) degrees of freedom: 0.842602
```

```
               Koenker (BP) Statistic [5]:  15.873747
     Prob(>chi-squared), (2) degrees of freedom:  0.001393*
```

You will try a variety of OLS regression models, attempting to find a model where all of the explanatory variables have statistically significant coefficients. These coefficients (and their statistical significance) can change radically depending on the combination of variables you include in your model. You will typically remove explanatory variables from you model if they are not statistically significant. If theory indicates a variable is very important, however, or if a particular variable is the focus of your analysis, you might keep a variable even if it is not statistically significant; it will have very little impact on how your model performs.

### *Check 2: Are the relationships what I expected?*

Not only is it important to determine whether an explanatory variable is actually helping your model, it is also important to think about the relationship it has to the dependent variable. The coefficient associated with each explanatory variable is either a negative number or a positive number. Suppose you were modeling crime, for example, and one of your statistically significant explanatory variables is neighborhood income. If the coefficient associated with the income variable is negative, it means that crime goes up as neighborhood incomes go down (a negative relationship). If you were modeling childhood obesity and access to fast food was an explanatory variable with a statistically significant positive coefficient, this would indicate that childhood obesity increases when the number of fast food opportunities also increases (a positive relationship).

When you create your list of candidate explanatory variables, you should include for each the relationship (positive or negative) you are expecting. You would have a hard time trusting a model that told you the opposite of what theory and/or common sense dictated. Suppose you were predicting forest fire frequency and your regression model returned a positive coefficient for your precipitation variable. You would not expect forest fires to increase in locations with lots of rain.

Unexpected coefficient signs usually indicate other problems with your model that will surface as you continue through the 6 checks. You can only trust the sign and strength of your explanatory variable coefficients if your model passes all of these checks. If you *do* find a model that passes all of the checks despite the unexpected coefficient sign, it might be an opportunity to learn something new. Perhaps there is a positive relationship between forest fire frequency and precipitation because the primary source of forest fires in your study area is lightening. It may be worthwhile to try to obtain data about lightning for your study area to see if it improves model performance.

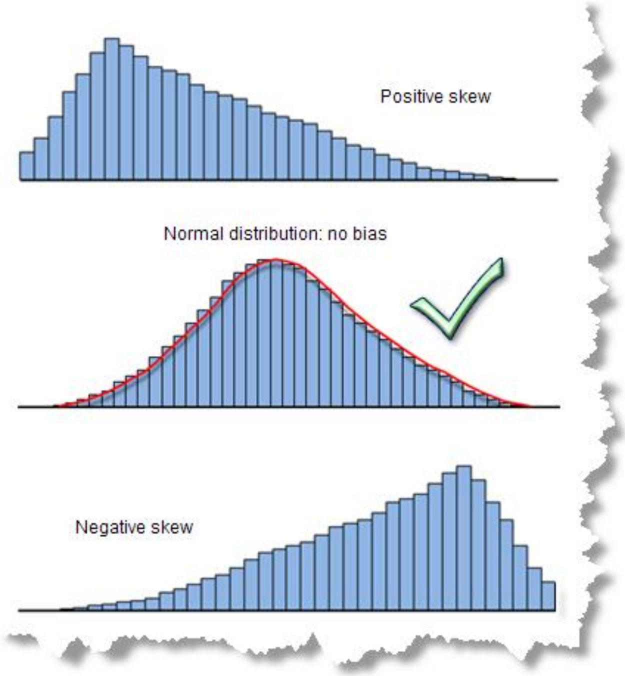
### *Check 3: Are there explanatory variables that are redundant?*

When choosing explanatory variables to include in your analysis, look for variables that get at difference aspects of what you are trying to model. Avoid variables that are redundant. For example, if you were trying to model home values, you probably wouldn't include explanatory variables for both home square footage and the number of bedrooms. Both of these variables are related to the size of the home, and including them both could make your model unstable. Ultimately, you cannot trust your model if you have multiple variables that tell the same story.

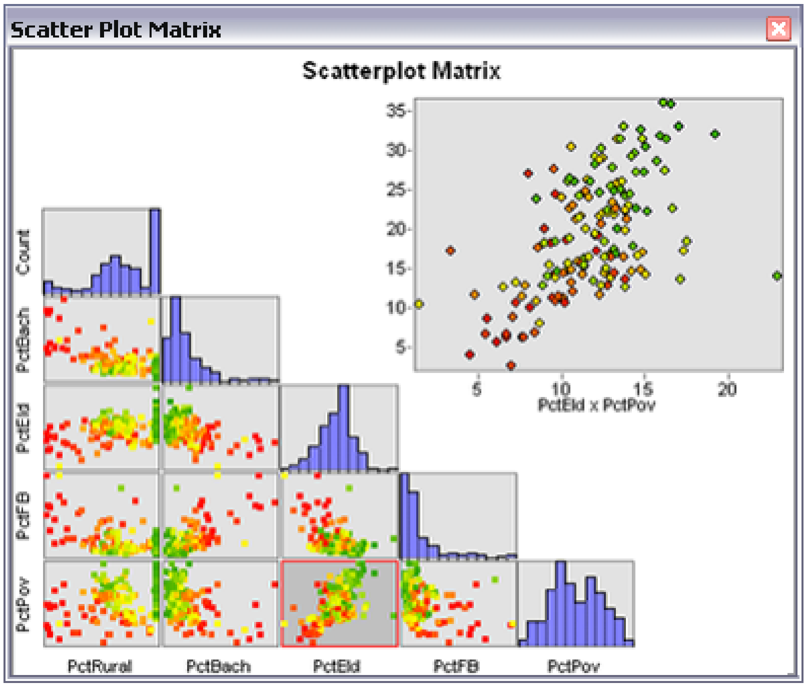
*How will you know if two or more variables are telling the same story?* Fortunately, the OLS tool computes a Variance Inflation Factor (VIF) value for each explanatory variable in your model. The VIF value is a measure of variable redundancy, and can help you decide which variables can be removed from your model without losing explanatory power. As a rule of thumb, a VIF value above 7.5 is considered problematic. If you have two or more variables with VIF values above 7.5, you should start by removing one variable at a time and rerunning OLS until you have removed all of the redundancy. But remember, you do not want to remove all variables with high VIF values. In our example of modeling home values, square footage and number of bedrooms would likely both have high VIF values, but we would only want to remove one of them from the model. Including one variable that gets at home size is important; we just don't want to model that aspect of home values redundantly.

### *Check 4: Is my model biased?*

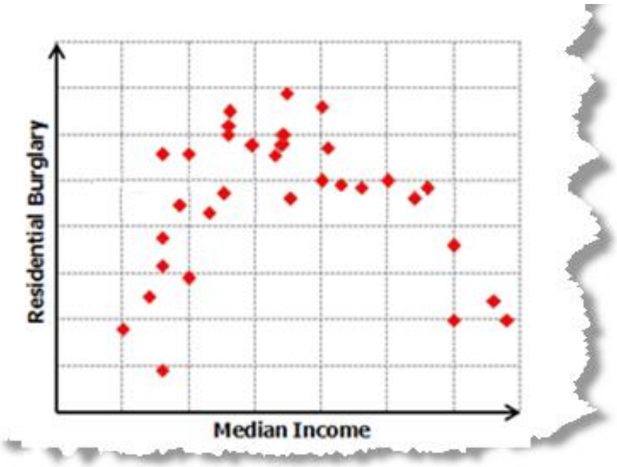
This may seem like a tricky question to answer, but it is actually very simple! When we have a properly specified OLS model, the model residuals (over and under predictions) have a normal distribution with a mean of zero (think bell curve). When our model is biased, however, the distribution of the residuals is unbalanced (see graphic). We cannot trust our predicted results when the model is biased. Luckily, there are a couple strategies we can use to try to correct this problem.



First we need to understand *how* our model is biased. One possibility is that our model is doing a good job for low values, but is not predicting well for high values (or vice versa). With the childhood obesity example, this would mean that where we have low childhood obesity our model is doing a great job, but in the areas with high childhood obesity our predictions are off. In this case, creating a Scatterplot Matrix can be a very useful tool for several reasons.



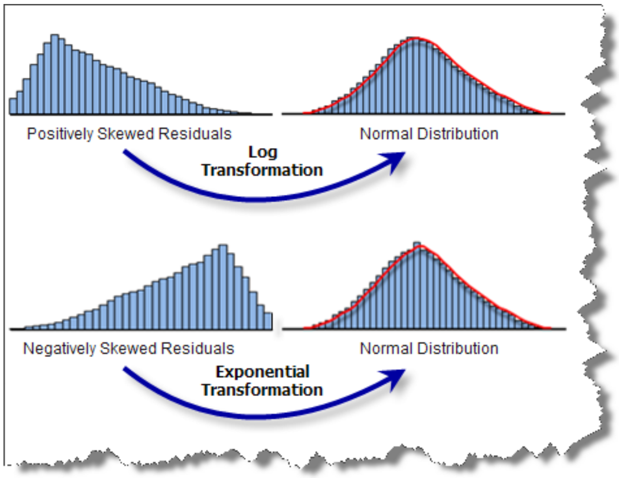
You can use the Scatterplot Matrix to evaluate all of the relationships between the variables in your data. Linear relationships would look like diagonal lines in the scatterplot matrix. Non-linear relationships could look more like curved lines, or take some other shape. If you see that your dependent variable has a non-linear relationship with one of your explanatory variables then you have some work to do! OLS is a linear regression model that assumes that the relationships between your variables are linear. If they aren't linear, you can try to transform your variables so that the relationships become linear. Common transformations include Log and Exponential functions.



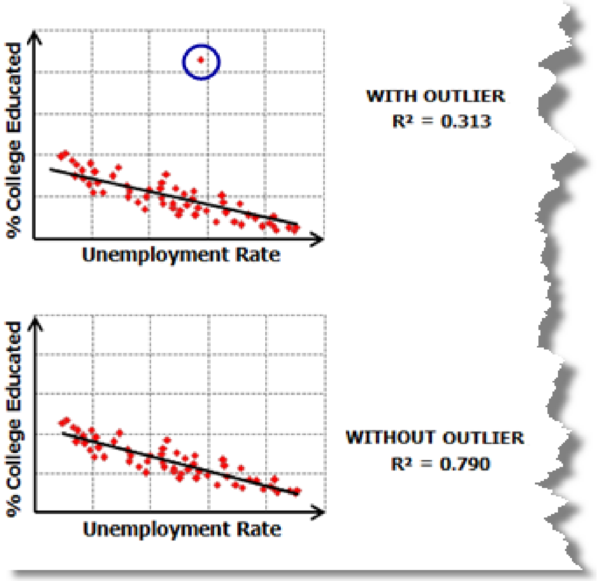
You should also check each of your variables to see if you have represented them in a way that will truly reflect linear relationships. Suppose, for example, one of your variables is Aspect, recorded either as compass directions or degrees (0 to 360). Circular variables need to be represented in a linear fashion. If you want to understand the relationship between bird nest densities and aspect, for example, you should first create a histogram showing the number of nests associated with each compass direction. If you find that a southwest (SW) aspect has the highest density of birds, for example, you need to give that the largest value. Locations with aspects near SW might also have high densities; in that case you would assign the next highest values to both West (W) and South (S). The directions with the fewest nests should have the smallest values. If the fewest nests are in the North (N), followed by Northeast (NE) and then East (E), you can assign N the lowest value, NE the next lowest value, and E the third lowest value. In this way you convert the circular representation of aspect to a linear ranking: from the direction the birds like most (SW) down to the direction they like least (N).

Would some of your variables be represented better by ranking categories? Suppose you suspect that housing values are high when they are near to commuter train stations, but not if they are TOO near. Rather than representing this relationship using a simple distance variable (a negative relationship: the larger the distance the smaller the home value), you might get a better model by assigning the highest rankings both to locations very near stations and to locations very far away from stations.

Another useful output of the scatterplot matrix is the histogram that is created for each of the variables. If some of your explanatory variables are strongly skewed, you may be able to remove model bias by transforming them. The image below shows how different types of transformations can help you get your data into its most useful form.



Model bias may also be a result of outliers that are influencing your model estimation, and you can also use the scatterplot matrix to find these. Try running OLS both with and without the outliers to see how much impact they are having on model results. In some instances (especially if you think that the outliers represent bad data), you may be able to drop the outliers and continue your analysis without them.



***Check 5: Have I found all of the key explanatory variables?***

Often times we go into an analysis with a hypothesis about what variables are going to be important. Maybe we think there are 5 variables that are definitely good predictors of what we are trying to model… or maybe there are 10 that we think could be related. While it is important to approach a regression analysis with a hypothesis, it is also important to allow your creativity and insight to dig a little deeper. You shouldn't limit yourself to your initial variable list. Really consider all of the variables that might impact what you are trying to model. Create thematic maps of each of your candidate explanatory variables and compare those to a map of your dependent variable. Hit the books again and scan the relevant literature. Use your intuition to look for relationships in your mapped data. Your goal is to try to come up with as many candidate *spatial* explanatory variables as you can as these will often be critical to your analysis. These are variables like distance from the urban center, proximity to major highways, or access to large bodies of water. These kinds of spatial variables can be very important in an analysis where you believe geographic processes influence the relationships in your data. Until you find explanatory variables that effectively capture the spatial structure of your dependent variable, you will be missing key explanatory variables and will not be able to pass all 6 diagnostic checks.

Sure evidence that you are missing one or more key explanatory variables is statistically significant spatial autocorrelation of your model residuals. For regression residuals, spatial autocorrelation usually takes the form of clustering: the over-predictions cluster together and the under-predictions cluster together. *How do you know if you have statistically significant spatial autocorrelation in your model residuals?* You can run the Spatial Autocorrelation (Moran's I) tool from the Spatial Statistics toolbox on your regression residuals. A statistically significant z-score indicates you are missing key explanatory variables.

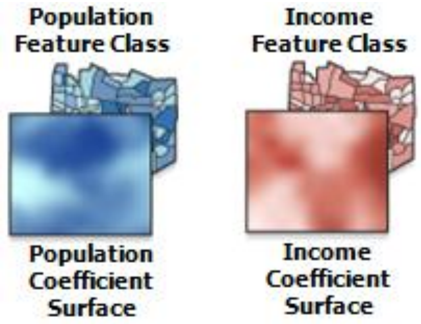
Finding those missing explanatory variables is often as much of an art as it is a science. Try these strategies to see if they give you clues about what might be missing:

***Examine the OLS residual map**.* The standard output from OLS is a map of the regression residuals. Red areas indicate your actual observed values are higher than the model predicted. Blue areas show were actual values are lower than the model predicted. Examining where the under and over predictions occur can often help you figure out which spatial variables might be missing. For instance, if you notice that you are consistently over-predicting in the urban areas, you might want to think about a variable representing the distance to the urban center or perhaps some kind of urban density variable. If it looks like the over predictions are associated with mountain peaks or valley bottoms, you might be missing an elevation variable. Do you see regional clusters of over/under predictions or can you recognize a trend in the residual data? Sometimes creating dummy variables to capture regional differences or trends is needed. A classic example of a dummy variable would be an urban/rural variable. By assigning all urban features a value of 0 and all rural features a value of 1 you are able to capture a spatial relationship in the landscape that could be influencing your model. Sometimes creating a hot spot map of regression residuals will help you to visualize broad regional patterns.

Figuring out the missing spatial variables not only have the potential to improve your model, they can also help you understand the phenomenon that you are modeling in a new and innovative way!

[Note: While spatial regime dummy variables are great to include in your OLS model, you will want to remove them when you run Geographically Weighted Regression (GWR). Since GWR takes these spatial relationships into consideration in its mathematics, these dummy variables aren't needed and, in fact, they create local redundancy between the dummy variable and the intercept, often making it impossible for GWR to solve].

***Examine coefficient surfaces**.* You can also try running Geographically Weighted Regression (GWR) and creating a coefficient surface for each of your explanatory variables. Select one of your OLS models to use for this exercise. A good choice would be a model with a high Adjusted $R^2$ value that is passing all or most of the other diagnostic checks. Because GWR creates a regression equation for each feature in your study area, the GWR coefficient surfaces illustrate how relationships between the dependent variable and each explanatory variable fluctuate geographically. Sometimes you see clues about missing spatial variables in these surfaces: a dip in prediction power near major freeways, a decline with distance from the coast, a sharp decrease in an industrial region, or a strong east to west trend or boundary.



Examining the coefficient surfaces also gives you a feel for how your modeled relationships are changing across space. You may find, for example, that a certain variable is really important in one part of your study area, but not important at all in another part. In some cases you might see that the coefficients for a variable actually switch signs, from a positive to a negative relationship. This is important to notice because OLS will likely discount the predictive potential of these highly non-stationary variables. OLS is a global model and is expecting relationships to be consistent (stationary) across the study area.

Consider, for example, the relationship between childhood obesity and access to healthy food options. It may be that in low income areas with poor access to cars, being far away from a supermarket is a real barrier to buying healthy food. In high income areas, however, having a supermarket within walking distance might actually be undesirable, and with potentially better access to vehicles, the distance to the supermarket might not act as a barrier to buying healthy foods at all. While GWR is capable of modeling these kinds of complex relationships, OLS is not. Instead, the positive relationship between distance and supermarkets for high income areas cancels out the negative relationship for low income areas so that the distance variable isn't significant at all in global model like OLS. Think of it as (+1) + (-1) = 0. When you examine the coefficient surfaces, look for areas where the coefficients are changing dramatically, especially if they are switching signs. If you believe strongly that these variables are important to your model, you should keep them even if OLS says they are not helping your model. These types of variables will be effective when you move to GWR.

***Try fitting OLS to smaller, subset study areas**.* GWR is tremendously useful when dealing with non-stationarity of explanatory variables, and it can be very tempting to move directly to GWR without first finding a properly specified OLS model. We do not recommend this. The reason we always start with OLS is that it gives us all sorts of great diagnostics to help us figure out if our explanatory variables are significant, if our residuals are normally distributed, and ultimately if we have a good model. GWR will not fix an improperly specified model unless you can be sure that the only reason your OLS model is failing the 6 checks is due to non-stationarity. Evidence of this would be finding explanatory variables that are strong predictors, but switch signs (they reflect a strong positive relationship in some parts of your study area, and a strong negative relationship in other parts of your study area). Alternatively, the problem may be that the set of key explanatory variables is inconsistent. It may be that one set of variables provides a great model in one part of the study area, and another set of variables provide a great model in another part of your study area. To see if this is the case, you can pick several small sample areas within your broader study area and then see if the explanatory variables for each subarea change. Pick sample areas where you think the processes associated with what you are trying to model might be different (high vs. low income areas, old vs. new housing, etc.). Alternatively map the Local R2 results from GWR and select areas where GWR performed well and where it performed badly. These might be areas with different sets of explanatory variables. It can be very useful to look at these areas individually using OLS.

If you do find good OLS models in several small study areas, then you can conclude that you've found the proper variables (overall), but just aren't getting a good global OLS model because of regional variation. You can move to GWR with the FULL set of variables from the combined smaller study area analyses, because GWR will adjust the coefficients to reflect that regional variation. If you don't get a good OLS model in any of the smaller study areas, it may be that the key explanatory variables are too complex to represent as a simple series of numeric measurements or the relationships between variables are not linear (and cannot be made linear through simple transformations). In that case, you will need to look for other analytical methods.

Okay, so all of this is a bit of work, but it really is a great exercise in exploratory data analysis, and will help you understand your data better, find new variables to use, and could even help you find a great model!

### ***Check 6: How well am I explaining my dependent variable?***

Now it's finally time to evaluate model performance. The Adjusted R-Squared ($R^2$) value is an important measure of how well your explanatory variables are modeling your dependent variable. The $R^2$ value is also one of the first things they tell you about regression analysis. So why are we leaving this important check until the end? What they *don't* tell you is that you cannot trust your $R^2$ value unless you have passed all of the other checks we've talked about. If your model is biased, it doesn't matter how high your $R^2$ value is. Likewise, if you have spatial autocorrelation of your residuals you cannot trust the coefficient relationships or be sure your model will predict well for new observations.

Once you have gone through the other checks and feel confident that you have met all of the necessary criteria, however, it is time to figure out how well your model is explaining your dependent variable by assessing the adjusted $R^2$ value. $R^2$ values range between 0 and 1, and represent a percentage. Suppose we are modeling crime rates and find a model that passes all 5 of the checks above with an Adjusted $R^2$ value of 0.65. This means that the explanatory variables in our model tell 65% of the crime rate story (or more technically, the model explains 65% of the variation in the crime rate variable). Adjusted $R^2$ values have to be judged rather subjectively. In some areas of science, explaining 23% of a complex phenomenon is very exciting! In other fields, an $R^2$ value may have to be closer to 80% or 90% to be considered notable. Either way, the adjusted $R^2$ values will help you judge the performance of your model.

Another important diagnostic for assessing model performance is known as Akaike's Information Criterion (AICc). The AICc value is a useful measure for comparing models that have the same dependent variable. For example, you might want to try modeling student test scores using several different sets of explanatory variables. In one model you might use only socioeconomic variables, while in another model you might use variables that represent the class size, teacher to student ratio, or other school-related variables. As long as the dependent variable for all models is the same (in this case student test scores), you can use the AICc values from each model to determine which performs better. The model with the smaller AICc value is the better model – that is, taking into account model complexity, the model with the smaller AICc provides a better fit for the observed data.

## And don't forget…

The most important thing to remember when you are going through these steps of building a properly specified regression model is that the goal of your analysis is to understand your data and to use that understanding to solve problems and answer questions. The truth is that you may try a number of models, with and without transformed variables, explore several small study areas, analyze your coefficient surfaces, and still not find a properly specified OLS model. But, and this is important, you will still be contributing to the body of knowledge on the phenomenon you are modeling! If the model you hypothesized would be a great predictor is not significant at all, discovering that is incredibly helpful information. If one of the variables you thought would be a strong predictor has a positive relationship in some areas and a negative relationship in other areas, knowing this certainly increases your understanding of the issue. The work that you do here, trying to find a good model using OLS, and using GWR to understand your data and improve your model, is always going to be valuable.

For more information about regression and other tools in the Spatial Statistics Toolbox, go to [www.bit.ly/spatialstats](www.bit.ly/spatialstats)