CrimeStat IV

Part V: Spatial Modeling II

Chapter 15:

OLS Regression Modeling

Ned Levine
Ned Levine & Associates
Houston, TX

Dominique Lord
Zachry Dept. of
Civil Engineering
Texas A & M University
College Station, TX

Table of Contents

Chapter 15:
OLS Regression Modeling[1]

The Regression I and Regression II modules are a series of routines for regression modeling and prediction. This chapter will lay out the basics of regression modeling and prediction and will discuss the Ordinary Least Squares (OLS) model in *CrimeStat*.

Functional Relationships

The aim of a regression model is to estimate a functional relationship between a dependent variable (call it y_i) and one or more independent variables (call them $x_{1i}, \cdots x_{Ki}$). In an actual database, these variables have unique names (e.g., ROBBERIES, POPULATION), but we will use general symbols to describe these variables. The functional relationship can be specified by an equation (15.1):

$$y_i = f(x_{1i}, \cdots, x_{Ki}) + \varepsilon_i \tag{15.1}$$

where Y is the dependent variable, $x_{1i}, \cdots x_{Ki}$ are the independent variables, $f(\,)$ is a functional relationship between the dependent variable and the independent variables, and ε_i is an error term (essentially, the difference between the actual value of the dependent variable and that predicted by the relationship).

Normal Linear Relationships

The simplest relationship between the dependent variable and the independent variables is *linear* with the dependent variable being normally distributed,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \tag{15.2}$$

---

[1] The regression chapters are the result of the effort of many persons. The maximum likelihood routines were produced by Ian Cahill of Cahill Software in Edmonton, Alberta as part of his MLE++ software package. We are grateful to him for providing these routines and for conducting quality control tests on them. The basic MCMC algorithm in *CrimeStat* for the Poisson-Gamma and Poisson-Gamma-CAR models was designed by Dr. Shaw-Pin Miaou of College Station, TX. We are grateful for Dr. Miaou for this effort. Improvements to the algorithm were made by us, including the block sampling strategy and the calculation of summary statistics. Dr. Dominique Lord of Texas A & M University provided technical advice on the Poisson-based models. Dr. Byung-Jung Park of the Korea Transport Institute expanded the MCMC algorithms to include various dispersion functions and a Simultaneous Autoregressive function. Dr. Ned Levine developed the block sampling methodology and provided overall project management. The programmer for the routines was Ms. Haiyan Teng of Houston, TX. We are also grateful to Dr. Richard Block of Loyola University in Chicago (IL) for testing the MCMC and MLE routines.

This equation can be written in a simple matrix notation: $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$ where $\mathbf{x}_i^T = (1, x_{1i}, \cdots, x_{Ki})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_K)^T$. The number one in the first element of $\mathbf{x}_i^T$ represents an intercept while T denotes that the matrix $\mathbf{x}_i^T$ is transposed.

This function says that a unit change in each independent variable, x_{ki}, for every observation, is associated with a unit change in the dependent variable, y_i. The coefficient of each variable, β_k, specifies the amount of change in y_i associated with that independent variable while keeping all other independent variables in the equation constant. The first term, β_0, is the intercept, a constant that is added to all observations. The error term, ε_i, is assumed to be *identically and independently* distributed (***iid***) across all observations, normally distributed with an expected mean of 0 and a constant standard deviation. If each of the independent variables has been standardized by

$$z_k = \frac{x_k - \bar{x}_k}{std(x_k)} \tag{15.3}$$

then the standard deviation of the error term will be 1.0 and the coefficients will be standardized, b_1, b_2, b_3, and so forth.

The equation is estimated by one of two methods, ordinary least squares (OLS) and maximum likelihood estimation (MLE). Both solutions produce the same results. The OLS method minimizes the sum of the squares of the residual errors while the maximum likelihood approach maximizes a joint probability density function.

Ordinary Least Squares

Appendix B by Luc Anselin discusses the method in more depth. Briefly, the intercept and coefficients are estimated by choosing a function that minimizes the residual errors by setting:

$$\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{k=1}^{K} \beta_k x_{ki} \right) x_{ki} = 0 \tag{15.4}$$

for *k=1* to *K* independent variables or, in matrix notation:

$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \tag{15.5}$$

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \tag{15.6}$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)^T$ and $\mathbf{y} = (y_1, y_2, \cdots, y_N)^T$.

The solution to this system of equations yields the familiar matrix expression for $\mathbf{b}_{OLS} = (b_0, b_1, \cdots, b_K)^T$

$$\mathbf{b}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{15.7}$$

An estimate for the error variance follows as

$$s^2_{OLS} = \sum_{i=1}^{N}\left(y_i - b_0 - \sum_{k=1}^{K} b_k x_{ki}\right)^2 /(N-K-1) \tag{15.8}$$

or, in matrix notation,

$$s^2_{OLS} = \mathbf{e}^T\mathbf{e}/(N-K-1) \tag{15.9}$$

Maximum Likelihood Estimation

For the maximum likelihood method, the *likelihood* of a function is the joint probability density of a series of observations (Wikipedia, 2010; Myers, 1990). Suppose there is a sample of *n* independent observations $(x_1, x_2, \cdots, x_N)$ that are drawn from an unknown *probability density* distribution but from a known family of distributions, for example the single-parameter exponential family. This is specified as $f(\cdot \mid \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the parameter (or parameters if there are more than one) that define the uniqueness of the family. The joint density function will be:

$$f(x_1, x_2, \cdots, x_N \mid \boldsymbol{\theta}) = f(x_1 \mid \boldsymbol{\theta}) \times f(x_2 \mid \boldsymbol{\theta}) \times \cdots \times f(x_N \mid \boldsymbol{\theta}) \tag{15.10}$$

and is called the ***likelihood*** function:

$$L(\boldsymbol{\theta} \mid x_1, x_2, \cdots, x_N) = f(x_1, x_2, \cdots, x_N \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i \mid \boldsymbol{\theta}) \tag{15.11}$$

where L is the likelihood and $\prod$ is the product term.

Typically, the likelihood function is interpreted in term of natural logarithms since the logarithm of a product is a sum of the logarithms of the individual terms. That is,

$$\ln\left\{\prod_{i=1}^{N} f(x_i \mid \boldsymbol{\theta})\right\} = \ln\left[f(x_1 \mid \boldsymbol{\theta})\right] + \ln\left[f(x_2 \mid \boldsymbol{\theta})\right] + \cdots + \ln\left[f(x_n \mid \boldsymbol{\theta})\right] \tag{15.12}$$

This is called the ***Log likelihood*** function and is written as:

$$\ln L(\boldsymbol{\theta} \mid x_1, x_2, \cdots, x_N) = \sum_{i=1}^{N} \ln\left[f(x_i \mid \boldsymbol{\theta})\right] \tag{15.13}$$

For the OLS model, the log likelihood is:

$$\ln L = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 \tag{15.14}$$

where N is the sample size and σ^2 is the variance. As a comparison, in Chapter 16 we discuss the Poisson model in which the log likelihood is:

$$\ln L = \sum_{i=1}^{N}\left[-\lambda_i + y_i \ln(\lambda_i) - \ln y_i!\right] \tag{15.15}$$

where $\lambda_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta})$ is the conditional mean for zone i, and y_i is the observed number of events for zone i. As mentioned, Anselin provides a more detailed discussion of these models in Appendix B.

The MLE approach estimates the value of θ that maximizes the log likelihood of the data coming from this family. Because they are all part of the same mathematical family and are distributed as a concave function, the maximum of a joint probability density distribution can be easily estimated. The approach is to, first, define a probability function from this family, second, create a joint probability density function for each of the observations (the Likelihood function); third, convert the likelihood function to a log likelihood; and, fourth, estimate the value of parameters that maximize the joint probability through an approximation method (e.g., Newton-Raphson or Fisher scores). Because the function is regular and known, the solution is relatively easy. Anselin discusses the approach in detail in Appendix B of the *CrimeStat* manual. More detail can be found in Hilbe (2008) or in Train (2009).

In *CrimeStat*, we use the MLE method. Because the OLS method is the most commonly used, a normal linear model is sometimes called an Ordinary Least Squares (OLS) regression. If the equation is correctly specified (i.e., all relevant variables are included), the error term, ε, will be normally distributed with a mean of 0 and a constant variance, σ^2.

The OLS normal estimate is sometimes known as a *Best Linear Unbiased Estimate* (BLUE) since it minimizes the sum of squares of the residuals errors (the difference between the

observed and predicted values of y). In other words, the overall fit of the normal model estimated through OLS or maximum likelihood will produce the best overall fit for a *linear* model. However, keep in mind that because a normal function has the best overall fit does not mean that it fits any particular section of the dependent variable better. In particular, for count data, the normal model usually does a poor job of modeling the observations with the greatest number of events. We will demonstrate this with an example below.

Assumptions of Normal Linear Regression

The normal linear model has some assumptions. When these assumptions are violated, problems can emerge in the model, sometimes easily correctable and other times introducing substantial bias.

Normal Distribution of Dependent Variable

First, the normal linear model assumes that the dependent variable is normally distributed. If the dependent variable is not exactly normally distributed, it has to have its peak somewhere in the middle of the data range and be somewhat symmetrical (e.g., a quartic distribution; see Chapter 10).

For some variables, this assumption is reasonable (e.g., with height or weight of individuals). However, for most variables that crime researchers work with (e.g., number of robberies, number of homicides, journey-to-crime distances), this assumption is usually violated. Most variables that are *counts* (i.e., number of discrete events) are highly skewed. Consequently, when it comes to counts and other extremely skewed variables, the normal (OLS) model will produce distorted results.

Errors are Independent, Constant, and Normally-distributed

Second, the errors in the model, the ε in equation 15.2, must be independent of each other, constant, and normally distributed. This fits the ***iid*** assumption mentioned above. Independence means that the estimation error for any one observation cannot be related to the error for any other observation. Constancy means that the amount of error should be more or less the same for every observation; there will be natural variability in the errors, but this variability should be distributed normally with the mean error being the expected value.

Unfortunately, for most variables that crime researchers and analysts work with, this assumption is usually violated. With count variables, the errors increase with the count and are much higher for observations with large counts than for observation with few counts. Thus, the assumption of constancy is violated. In other words, the variance of the error term is a function

of the count. The shape of the error distribution is also sometimes not normal either but may be more skewed. Also, if there is spatial autocorrelation among the error terms (which would be expected in a spatial distribution), then the error term may be quite irregular in shape; in this latter case, the assumption of independent observations would also be violated.

Independence of Independent Variables

Third, an assumption of the normal model (and any model, for that matter) is that the independent variables are truly independent. In theory, there should be zero correlation between any of the independent variables. In practice, however, many variables are related, sometimes quite highly. This condition, which is called *multicollinearity*, can produce distorted coefficients and overall model effects. The higher the degree of multicollinearity among the independent variables, the greater the distortion in the coefficients. This problem affects all types of models, not just the normal, and it is important to minimize the effects. We will discuss diagnostic methods for identifying multicollinearity later in the chapter.

Adequate Model Specification

Fourth, the normal model assumes that the independent variables have been correctly *specified*. That is, the independent variables are the correct ones to include in the equation and that they have been measured adequately. By 'correct ones', we mean that the independent variable chosen should be a true predictor of the dependent variable, not an extraneous one. With any model, the more independent variables that are added to the equation, in general the greater will be the overall fit. This will be true even if the independent variables are highly correlated with independent variables already in the equation or are mostly irrelevant (but may be slightly correlated due to sampling error). When too many variables are added to an equation, strange effects can occur. *Overfitting* of a model is a serious problem that must be seriously evaluated. Including too many variables will also artificially increase the model's variance (Myers, 1990).

Conversely, a correct specification implies that all the important variables have been included and that none have been left out. When important variables are not included, this is called *underfit*ting a model. Also, not including important variables lead to a biased model (known as the *omitted variables* bias). A large bias means that the model is unreliable for prediction (Myers, 1990). Also, the left out variables can be shown to have irregular effects on the error terms. For example, if there is spatial autocorrelation in the dependent variable (which there usually is), then the error terms will be correlated. Without modeling the spatial autocorrelation (either through a proxy variable that captures much of its effect or through a parameter adjustment), the error can be biased and even the coefficients can be biased.

In other words, adequate specification involves choosing the correct number of independent variables that are appropriate, neither overfitting nor underfitting of the model. Also, it is assumed that the variables have been correctly measured and that the amount of measurement error is very small.

Unfortunately, we often do not know whether a model is correctly specified or not, nor whether the variables have been properly measured. Consequently, there are a number of diagnostics tests that can be brought to bear to reveal whether the specification is adequate. For overfitting, there are tolerance statistics and adjusted summary values. For underfitting, we analyze the error distribution to see if there is a pattern that might indicate *lurking* variables that are not included in the model. In other words, examining violations of the assumptions of a model is an important task in assessing whether there are too many variables included or whether there are variables that should be included but are not, or whether the specification of the model is correct or not.

Example of Modeling Burglaries by Zones

For many problems, normal regression is an appropriate tool. However, for many others, it is not. Let us illustrate this point. A note of caution is warranted here. This example is used to illustrate the application of the normal model in CrimeStat and, as discussed further below, the normal model with a normal error distribution is not appropriate for this kind of dataset. For example, figure 15.1 shows the number of residential burglaries that occurred in 2006 within 1,179 Traffic Analysis Zones (TAZ) inside the City of Houston. The data on burglaries came from the Houston Police Department. There were 26,480 burglaries that occurred in 2006. They were then allocated to the 1,179 TAZ's within the City of Houston. As can be seen, there is a large concentration of residential burglaries in southwest Houston with small concentrations in southeast Houston and in parts of north Houston.

The distribution of burglaries by zones is quite skewed. Figure 15.2 shows a graph of the number of burglaries per zone. Of the 1,179 traffic analysis zones, 250 had no burglaries occur within them in 2006. On the other hand, one zone had 284 burglaries occur within it. The graph shows the number of burglaries up to 59; there were 107 zones with 60 or more burglaries that occurred in them. About 58% of the burglaries occurred in 10% of the zones. In general, a small percentage of the zones have the majority of the burglaries.

Example of Normal Linear Model

We can set up a normal linear model to try to predict the number of burglaries that occurred in each zone in 2006. We obtained estimates of population, employment and income from the transportation modeling group within the Houston-Galveston Area Council, the

Figure 15.1:

Burglaries in the City of Houston

Number in Each Traffic Analysis Zone: 2006

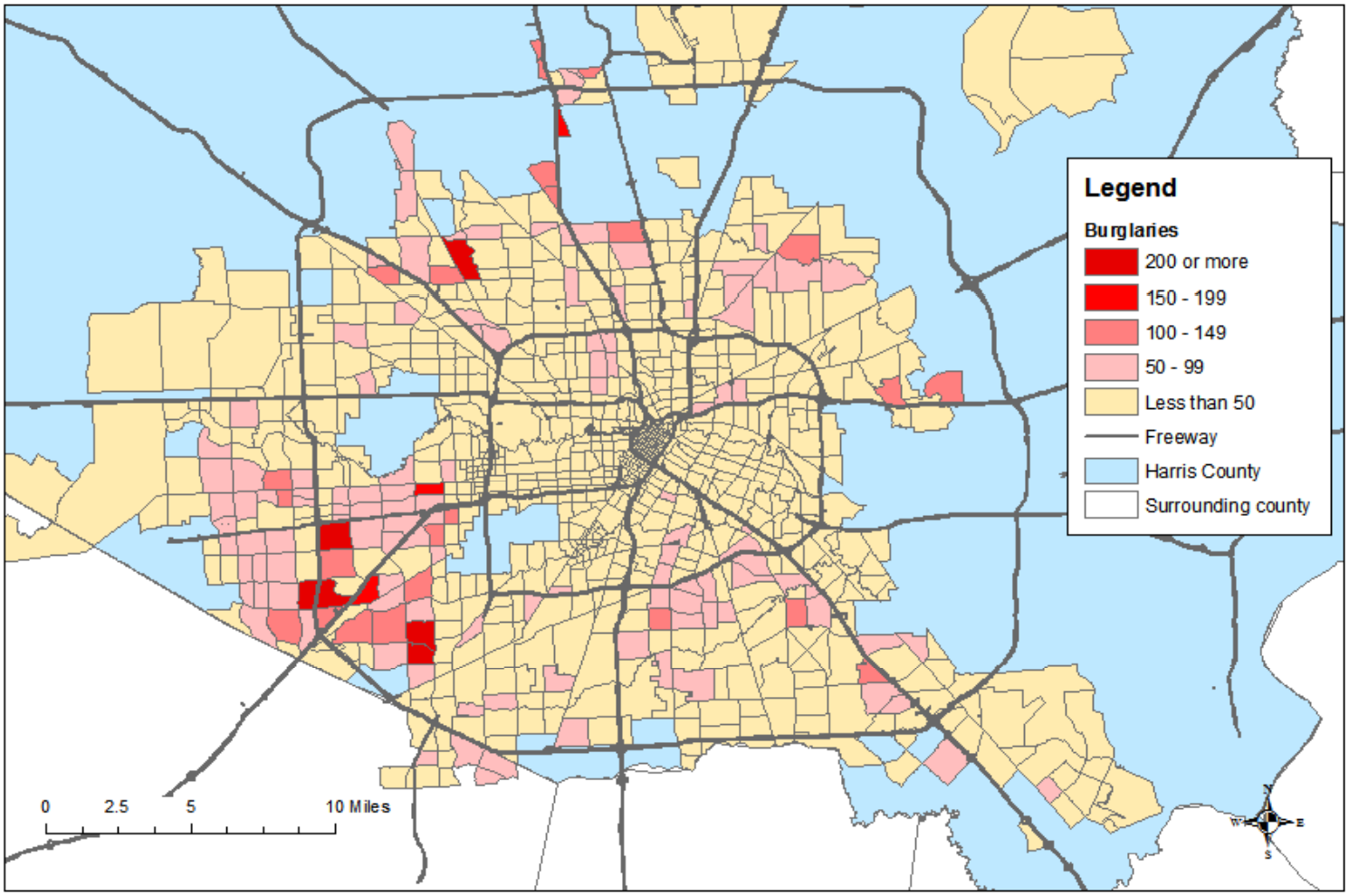
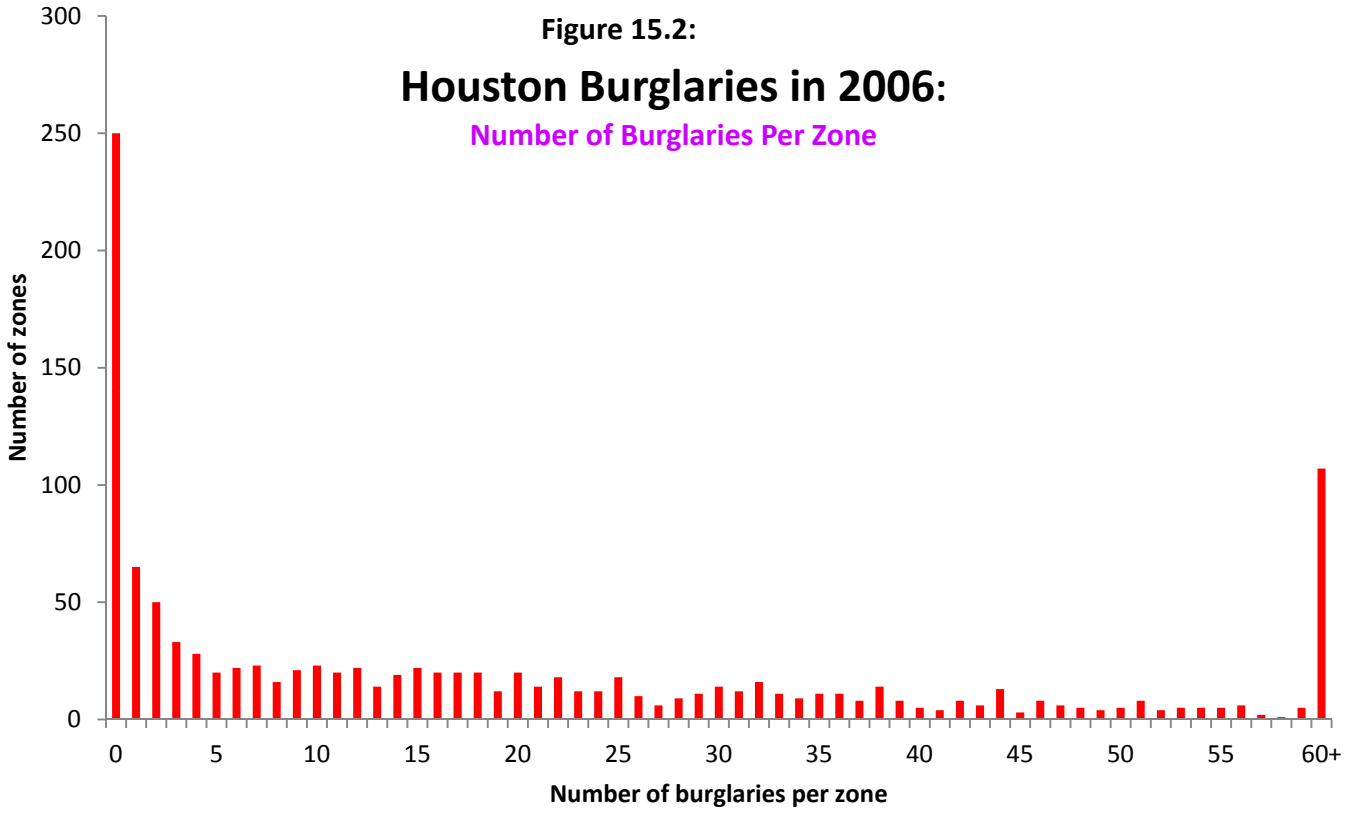


Figure 15.2:


Houston Burglaries in 2006:
Number of Burglaries Per Zone
Number of zones
300
250
200
150
100
50
0
0
5
10
15
20
25
30
35
40
45
50
55
60+
Number of burglaries per zone

Metropolitan Planning Organization for the area (H-GAC, 2010). Specifically, the model relates the number of 2006 burglaries to the number of households, number of jobs (employment), and median income of each zone. The estimates for the number of households and jobs were for 2006 while the median income was that measured by the 2000 census. Table 15.1 present the results of the normal (OLS) model.

Table 15.1:

Predicting Burglaries in the City of Houston: 2006
Ordinary Least Squares: Full Model

(N= 1,179 Traffic Analysis Zones)

DepVar:	**2006 BURGLARIES**
N:	1,179
Df:	1,174
Type of regression model:	Ordinary Least Squares
F-test of model:	357.2 $p \leq .0001$
R-square:	0.48
Adjusted r-square:	0.48
Mean absolute deviation:	13.5
1st (highest) quartile:	26.4
2nd quartile:	10.6
3rd quartile:	8.3
4th (lowest) quartile:	8.8
Mean squared predictive error:	505.1
1st (highest) quartile:	1,497.5
2nd quartile:	270.4
3rd quartile:	134.3
4th (lowest) quartile:	120.9

Predictor	DF	Coefficient	Stand Error	Tolerance	VIF	t-value	p
INTERCEPT	1	12.9320	1.269	-	-	10.19	0.001
HOUSEHOLDS	1	0.0256	0.0008	0.923	1.083	31.37	0.001
JOBS	1	-0.0002	0.0005	0.903	1.107	-0.453	n.s.
MEDIAN HOUSEHOLD INCOME	1	-0.0002	0.00003	0.970	1.031	-6.88	0.001

Summary Statistics for the Goodness-of-Fit

The table presents two types of results. First, there is summary information. Information on the size of the sample (in this case, 1,179) and the degrees of freedom (the sample size less one for each parameter estimated including the intercept and one for the mean of the dependent variable); in the example, there are 1,174 degrees of freedom (1,179 – 1 for the intercept, 1 for HOUSEHOLDS, 1 for JOBS, 1 for MEDIAN HOUSEHOLD INCOME, and 1 for the mean of the dependent variable, 2006 BURGLARIES).

The F-test presents an Analysis of Variance test of the ratio of the *mean square error* (MSE) of the model compared to the total mean square error (Kanji, 1993, 131; Abraham & Ledolter, 2006, 41-51). Next, there is the R-square (or R^2) statistic, which is the most common type of overall fit test. This is the percent of the total variance of the dependent variable accounted for by the model. More formally, it is defined as:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{15.16}$$

where y_i is the observed number of events for a zone, i, $\hat{y}_i$ is the predicted number of events given a set of *K* independent variables, and Mean $\bar{y}$ is the mean number of events across zones. The R-square value is a number from 0 to 1; 0 indicates no predictability while 1 indicates perfect predictability.

For a normal (OLS) model, R-square is a very consistent estimate. It increases in a linear manner with predictability and is a good indicator of how effective a model has fit the data. As with all diagnostic statistics, the value of the R-square increases with more independent variables. Consequently, an R-square adjusted for degrees of freedom is also calculated - the *adjusted r-square* in the table. This is defined as:

$$R_a^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2 / (N - K - 1)}{\sum (y_i - \bar{y})^2 / (N - 1)} \tag{15.17}$$

where N is the sample size and K is the number of independent variables.

The R^2 value is sometimes called the *coefficient of determination*. It is an indicator of the extent to which the independent variables in the model *predict* (or explain) the dependent variable. One interpretation of the R^2 is the percent of the variance of Y accounted for by the variance of the independent variables (plus the intercept and any other constraints added to the model). The *unexplained* variance is 1 - R^2 or the extent to which the model does not explain the

variance of the dependent variable. For a normal linear model, the R^2 is relatively straightforward. In the example, both the F-test is highly significant and the R^2 is substantial (48% of the variance of the dependent variable is explained by the independent variables). However, for non-linear models, it is not at all an intuitive measure and has been shown to be unreliable (Miaou, 1996).

The final two summary measures are *Mean Squared Predictive Error* (MSPE), which is the average of the squared residual errors, and the *Mean Absolute Deviation* (MAD), which is the average of the absolute value of the residual errors (Oh, Lyon, Washington, Persaud, & Bared, 2003). The lower the values of these measures, the better the model fits the data.

These measures are also calculated for specific quartiles. The 1st quartile represents the error associated with the 25% of the observations that have the highest values of the dependent variable while the 4th quartile represents the error associated with the 25% of the observations with the lowest value of the dependent variable. These percentiles are useful for examining how well a model fits the data and whether the fit is better for any particular section of the dependent variable. In the example, the fit is better for the low end of the distribution (the zones with zero or few burglaries) and less good for the higher end. We will use these values in comparing the normal model to other models.

It is important to point out that the summary measures are more useful when several models with a different number of variables are compared with each other than for evaluating a single model.

Statistics on Individual Coefficients

The second type of information presented is about each of the coefficients. The table lists the independent variables plus the intercept. For each coefficient, the degrees of freedom associated are presented (one per variable) plus the estimated linear coefficient. For each coefficient, there is an estimated standard error, a t-test of the coefficient (the coefficient divided by the standard error), and the approximate two-tailed probability level associated with the t-test (essentially, an estimate of the probability that the null hypothesis of zero coefficient is correct). Usually, if the probability level is smaller than 5% (.05), then we reject the null hypothesis of a zero coefficient though frequently 1% (.01) or even 0.1% (0.001) have been used to reduce the likelihood that a false alternative hypothesis has been selected (called a *Type I error*).

The last two parameters included in the table are the *tolerance* of the coefficient and the *VIF* (or Variance *Inflation Factor*). They are measures of multicollinearity (or one type of overfitting). Basically, they measure the extent to which each independent variable correlates with the other dependent variables in the equation. The traditional tolerance test is a normal

model relating each independent variable to the *other* independent variables (StatSoft, 2010; Berk, 1977). It is defined as:

$$Tol_i = 1 - R^2_{j \neq i} \tag{15.18}$$

where $R^2_{j \neq i}$ is the R-square associated with the prediction of one independent variable with the remaining independent variables in the model using an OLS model. The VIF is simply the reciprocal of tolerance:

$$VIF_i = 1 / Tol_i \tag{15.19}$$

In other words, the tolerance of each independent variable is the unexplained variance of a model that relates the variable to the other independent variables. If an independent variable is highly related (correlated with) to the other independent variables in the equation, then it will have a low tolerance. Conversely, if an independent variable is independent of the other independent variables in the equation, then it will have a high tolerance. In theory, the higher the tolerance, the better since each independent variable should be unrelated to the other independent variables. In practice, there is always some degree of overlap between the independent variables so that a tolerance of 1.0 is rarely, if ever, achieved. However, if the tolerance is low (e.g., 0.70 or below), this suggests that there is too much overlap in the independent variables and that the interpretation will be unclear. In Chapter 17, we will discuss multicollinearity and the general problem of overfitting in more detail.

Note that the statistic is labeled as *pseudo-tolerance* in the CrimeStat output. The reason is that this statistic is only approximate when the independent variable is skewed, a situation that we will discuss shortly. For a normally-distributed independent variable (or approximately normally-distributed), however, the tolerance test is exact.

Looking at the output in Table 15.1, we see that the number of burglaries is positively associated with the intercept and the number of households and negatively associated with the median household income. The relationship to the number of jobs is also negative, but not significant. Essentially, zones with larger numbers of households but lower household incomes are associated with more residential burglaries. Because the model is linear, each of the coefficients contributes to the prediction in an additive manner. The intercept is 12.93 and indicates that, on average, each zone had 12.93 burglaries. For every household in the zone, there was a contribution of 0.0256 burglaries. For every job in the zone, there was a contribution of -0.0002 burglaries. For every dollar increase in median household income, there is a decrease of -0.0002 burglaries. Thus, to predict the number of burglaries with the full model in any one zone, i, we would take the intercept – 12.93, and add in each of these components:

$$(BURGLARIES)_i = 12.93 + 0.0256(HOUSEHOLDS)_i - 0.0002(JOBS)_i - 0.0002(MEDIAN\ HOUSEHOLD\ INCOME)_i \quad (15.20)$$

To illustrate, TAZ 833 had 1762 households in 2006, 2,698 jobs in 2006, and had a median household income of $27,500 in 2000. The model's prediction for the number of burglaries in TAZ 833 is:

Number of burglaries (TAZ833) = 12.93 + 0.0256*1762 – 0.0002*2,698 – 0.0002*27,500
= 52.0

The actual number of burglaries that occurred in TAZ 833 was 78.

Estimated Error in the Model for Individual Coefficients

In *CrimeStat*, and in most statistical packages, there is additional information that can be output as a file. There is the *predicted* value for each observation. Essentially, this is the linear prediction from the model. There is also the *residual* error, which is the difference between the actual (observed) value for each observation, i, and that predicted by the model. It is defined as:

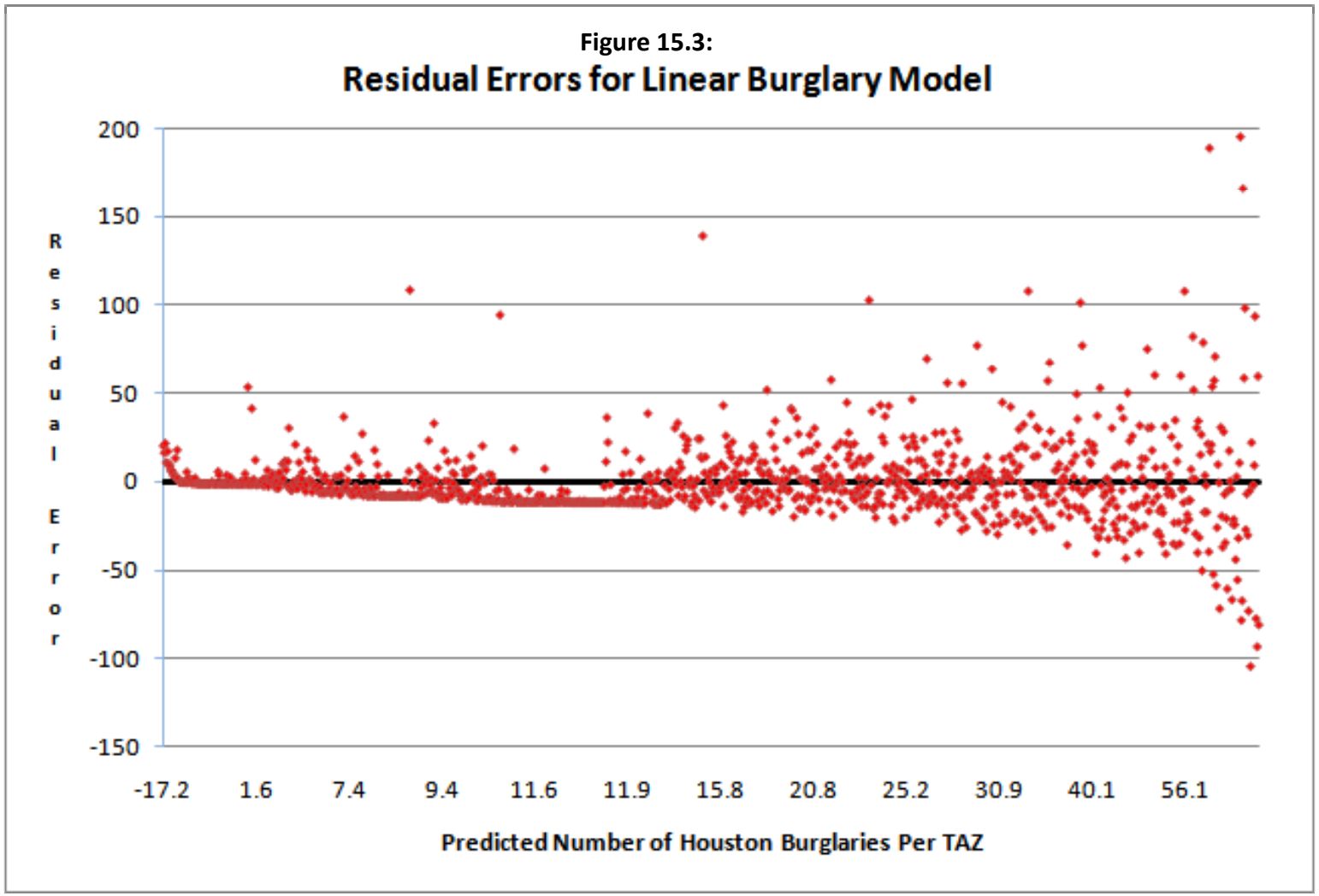
$$\text{Residual error}_i = \text{Observed Value}_i - \text{Predicted value}_i \quad (15.21)$$

Table 15.2 below gives predicted values and residual errors for five of the observations from the Houston burglary data set. Analysis of the residual errors is one of the best tools for diagnosing problems with the model. A plot of residual errors against predicted values indicate whether the prediction is consistent across all values of the dependent variable and whether the underlying assumptions of the normal model are valid (see below). Figure 15.3 show a graph of the residual errors of the full model against the predicted values for the model estimated in table 1. As can be seen, the model fits quite well for zones with few burglaries, up to about 12 burglaries per zone.

Table 15.2:
Predicted Values and Residual Error for Houston Burglaries: 2006
(5 Traffic Analysis Zones)

Zone (TAZ)	Actual value	Predicted value	Residual error
833	78	52.0	26.0
831	46	35.9	10.1
911	89	67.6	21.4
2173	30	42.3	-12.3
2940	3	10.2	-7.2

Figure 15.3:


Residual Errors for Linear Burglary Model
Residual Error
200
150
100
50
0
-50
-100
-150
-17.2
1.6
7.4
9.4
11.6
11.9
15.8
20.8
25.2
30.9
40.1
56.1
Predicted Number of Houston Burglaries Per TAZ

However, for the zones with many predicted burglaries (the ones that we are most likely interested in), the model does quite poorly. First, the errors increase the greater the number of predicted burglaries. Sometimes the errors are positive, meaning that the actual number of burglaries is much higher than predicted and sometimes the errors are negative, meaning that we are predicting more burglaries than actually occurred. More importantly, the residual errors indicate that the model has violated one of the basic assumptions of the normal model, namely that the errors are independent, constant, and identically-distributed. It is clear that they are not.

Because there are errors in predicting the zones with the highest number of burglaries and because the zones with the highest number of burglaries were somewhat concentrated, there are spatial distortions from the prediction. Figure 15.4 show a map of the residual errors of the normal model. As can be seen by comparing this map with the map of burglaries (figure 15.1), typically the zones with the highest number of burglaries (mostly in southwest Houston) were under-estimated by the normal model (shown in red) whereas some zones with few burglaries ended up being over-estimated by the normal model (e.g., in far southeast Houston).

In other words, the normal linear model is not necessarily good for predicting Houston burglaries. It tends to underestimate zones with a large number of burglaries but overestimates zones with few.

Violations of Assumptions for Normal Linear Regression

There are several deficiencies with the normal (OLS) model. First, normal models are not good at describing skewed dependent variables, as we have shown. Since crime distributions are usually skewed, this is a serious deficiency for multivariate crime analysis. Second, a normal model can have negative predictions. With a count variable, such as the number of burglaries committed in a zone, the minimum number is zero. That is, the count variable is always *positive*, being bounded by 0 on the lower limit and some large number on the upper limit. The normal model, on the other hand, can produce negative predicted values since it is additive in the independent variables. This clearly is illogical and is a major problem with data that are highly skewed. If most records have values close to zero, it is very possible for a normal model to predict a negative value.

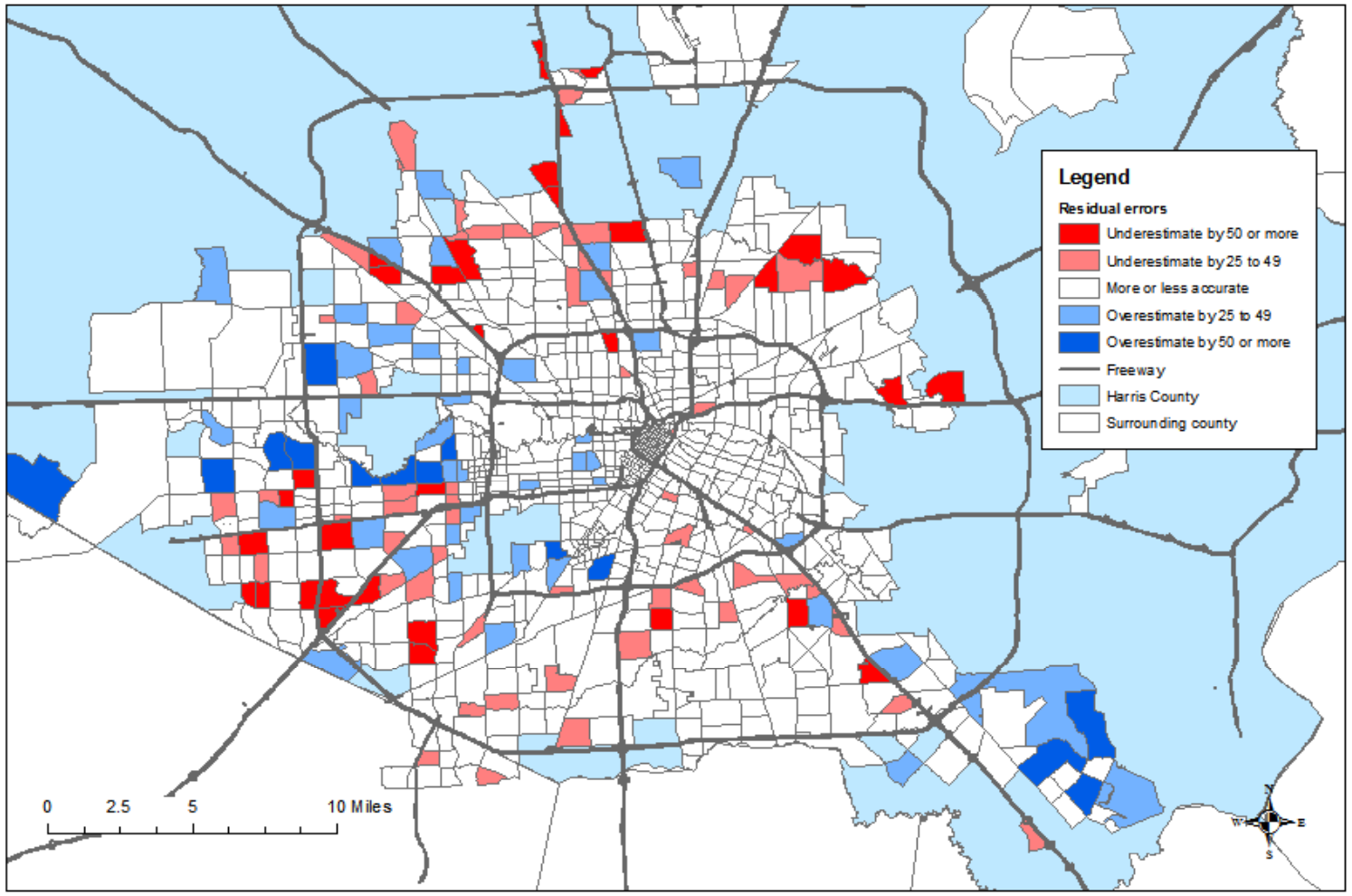
Non-consistent Summation

A third problem with the normal model is that the sum of the observed values does not necessarily equal the sum of the predicted values. Since the estimates of the intercept and coefficients are obtained by minimizing the sum of the squared residual errors (or maximizing the joint probability distribution, which leads to the same result), there is no balancing mechanism to require that they add up to the same as the input values. In calibrating the model,

Figure 15.4:

Predicting Burglaries in the City of Houston: 2006

Residual Errors from Linear Model



adjustments can be made to the intercept term to force the sum of the predicted values to be equal to the sum of the input values. But in applying that intercept and coefficients to another data set, there is no guarantee that the consistency of summation will hold. In other words, the normal method cannot guarantee a consistent set of predicted values.

Non-linear Effects

A fourth problem with the normal model is that it assumes the independent variables are normal in their effect. If the dependent variable was normal or relatively balanced, then a normal model would be appropriate. But, when the dependent variable is highly skewed, as is seen with these data, typically the additive effects of each component cannot usually account for the non-linearity. Independent variables have to be transformed to account for the non-linearity and the result is often a complex equation with non-intuitive relationships.[2] It is far better to use a non-linear model for a highly skewed dependent variable.

Greater Residual Errors

The final problem with a normal model and a skewed dependent variable is that the model tends to over- or under-predict the correct values, but rarely comes up with the correct estimate. As we saw with the example above, typically a normal equation produces non-constant residual errors with skewed data. In theory, errors in prediction should be uncorrelated with the predicted value of the dependent variable. Violation of this condition is called *heteroscedasticity* because it indicates that the residual variance is not constant. The most common type is an increase in the residual errors with higher values of the predicted dependent variable. That is, the residual errors are greater at the higher values of the predicted dependent variable than at lower values (Draper and Smith, 1981, 147).

A highly skewed distribution tends to exacerbate this. Because the least squares procedure minimizes the sum of the squared residuals, the regression line balances the lower residuals with the higher residuals. The result is a regression line that neither fits the low values nor the high values. For example, motor vehicle crashes tend to concentrate at a few locations (crash hot spots). In estimating the relationship between traffic volume and crashes, the hot spots tend to unduly influence the regression line. The result is a line that neither fits the number of expected crashes at most locations (which is low) nor the number of expected crashes at the hot spot locations (which are high).

---

[2] For example, to account for the skewed dependent variable, one or more of the independent variables have to be transformed with a non-linear operator (e.g., log or exponential term). When more than one independent variable is non-linear in an equation, the model is no longer easily understood. It may end up making reasonable predictions for the dependent variable, but is not intuitive nor easily explained to non-specialists.

Corrections to Violated Assumptions in Normal Linear Regression

Some of the violations in the assumptions of an OLS normal model can be corrected.

Eliminating Unimportant Variables

One good way to improve a normal model is to eliminate variables that are not important. Including variables in the equation that do not contribute very much adds 'noise' (variability) to the estimate. In the above example, the variable, JOBS, was not statistically significant and, hence, did not contribute any real effect to the final prediction. This is an example of overfitting a model. Whether we use the criteria of statistical significance to eliminate non-essential variables or simply drop those with a very small effect is less important than the need to reduce the model to only those variables that truly predict the dependent variable. We will discuss the 'pros' and 'cons' of dropping variables in Chapter 17, but for now we argue that a good model - one that will be good not just for description but for prediction, is usually a simple model with only the strongest variables included.

To illustrate, we reduce the burglary model further by dropping the non-significant variable (JOBS). Table 15.3 show the results. Comparing the results with those from Table 15.1, we can see that the overall fit of the model is actually slightly better (an F-value of 536.0 compared to 357.2). The R^2 values are the same while the mean squared predictive error is slightly worse while the mean absolute deviation is slightly better. The coefficients for the two common independent variables are almost identical while that for the intercept is slightly less (which is good since it contributes less to the overall result).

In other words, dropping the non-significant variable has led to a slightly better fit. One will usually find that dropping non-significant or unimportant variables makes models more stable without much loss of predictability, and conceptually they become simpler to understand.

Eliminating Multicollinearity

Another way to improve the stability of a normal model is to eliminate variables that are substantially correlated with other independent variables in the equation. This is the *multicollinearity* problem that we discussed above. Even if a variable is statistically significant in a model, if it is also correlated with one or more of the other variables in the equation, then it is capturing some of the variance associated with those other variables. The results are ambiguity in the interpretation of the coefficients as well as error in trying to use the model for

Table 15.3:

Predicting Burglaries in the City of Houston: 2006
Ordinary Least Squares: Reduced Model

(N= 1,179 Traffic Analysis Zones)

DepVar:	**2006 BURGLARIES**
N:	1,179
Df:	1,175
Type of regression model:	Ordinary Least Squares
F-test of model:	536.0 $p \leq .0001$
R-square:	0.48
Adjusted r-square:	0.48
Mean absolute deviation:	13.5
1st (highest) quartile:	26.5
2nd quartile:	10.6
3rd quartile:	8.3
4th (lowest) quartile:	8.8
Mean squared predictive error:	505.1
1st (highest) quartile:	1498.8
2nd quartile:	269.5
3rd quartile:	135.1
4th (lowest) quartile:	120.2

Predictor	DF	Coefficient	Stand Error	Tolerance	VIF	t-value	p
INTERCEPT	1	12.8099	1.240	-	-	10.33	0.001
HOUSEHOLDS	1	0.0255	0.0008	0.994	1.006	33.44	0.001
MEDIAN HOUSEHOLD INCOME	1	-0.0002	0.00003	0.994	1.006	-7.03	0.001

prediction. Multicollinearity means that essentially there is overlap in the independent variables; they are measuring the same thing. It is better to drop a multicollinear variable even if it results in a loss in fit since it will usually result in a simpler and more stable model.

For the Houston burglary example, the two remaining independent variables in Table 15.3 are relatively independent; their tolerances are 0.994 respectively, which points to little overlap in the variance that they account for in the dependent variable. Therefore, we will keep

these variables. However, in the next chapter, we will present an example of how multicollinearity can lead to ambiguous coefficients.

Transforming the Dependent Variable

It may be possible to correct the normal model by transforming the dependent variable (in another program since *CrimeStat* does not currently do this). Typically, with a skewed dependent variable and one that has a large range in values, a natural log transformation of the dependent variable can be used to reduce the amount of skewness. The problem will occur for zones with 0 since the natural log of 0 cannot be calculated. Consequently, one takes:

$$\ln y_i = \log_e (y_i + 1) \tag{15.22}$$

where *e* is the base of the natural logarithm (2.718…) and regresses the transformed dependent variable against the linear predictors,

$$\ln y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \tag{15.23}$$

This is equivalent to the equation

$$y_i = e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i} \tag{15.24}$$

with, again, *e* being the base of the natural logarithm.

In doing this, it is assumed that the log transformed dependent variable is consistent with the assumptions of the normal model, namely that it is normally distributed with an independent and constant error term, ε, that is also normally distributed.

One must be careful about transforming values that are zero since the natural log of 0 is unsolvable. Usually researchers will set the value of the log-transformed dependent variable to 0 or the value of the dependent variable to a small number (e.g., 1) for cases where the raw dependent variable actually has a value of 0 (e.g., equation 15.22 above). But, one must be careful that it does not distort relationships if there are many zeros in the data. For example, in the burglary data, there were 250 zones (out of 1,179, or 21%) that had zero burglaries!

Example of Transforming Dependent Variable on Houston Burglaries

Using the Houston burglary example from above, we transformed the dependent variable– number of 2006 burglaries per TAZ, by taking the natural logarithm of it. All zones with zero burglaries were automatically given the value of 0 for the transformed variable.

The transformed variable was then regressed against the two independent variables in the reduced form model (from Table 15.3 above). Table 15.4 present the results: The coefficients are similar in sign. The R^2 value is smaller than the untransformed model (0.42 compared to 0.48). Further, the mean squared predictive error is now much lower than the original raw values (1.47 compared to 505.14) and the mean absolute deviation is also much lower (1.05 compared to 13.50).[3] In other words, transforming the dependent variable into a logarithm has improved the fit of the estimate substantially.

Another type of transformation that is sometimes used is to convert the independent variables and, occasionally, the dependent variable into Z-scores. The Z-score of a variable is defined as:

$$z_k = \frac{x_k - \bar{x}_k}{std(x_k)} \tag{15.25}$$

But all this will do is to standardize the scale of the variable as standard deviations around an expected value of zero, but not alter the shape. If the dependent variable is skewed, taking the Z-score of it will not alter its skewness.

A third type of transformation takes the square root of the dependent variable and regress it in an OLS model. When we did this with the Houston burglary data, however, the fit was not as good as the log transformation (model not shown). The mean absolute deviation was more than 50% higher and the mean squared predictive error was three times higher. Again, the basic reason is that a count, such as the number of burglaries, is typically Poisson-distributed, meaning that it is exponential in form. Essentially, skewness is a fundamental property of a distribution and the normal model is poorly suited for modeling it.

Example of Modeling Skewed Variable with OLS

A simple example can illustrate this theoretically. Figure 15.5 shows an exponential distribution that relates a dependent variable, Y, to an independent variable, X. Think of these as any two variables that are positively related (e.g., crime & poverty; crime & unemployment). The data were created in a spreadsheet by the function $Y_i = e^X$ with a random error added to simulate randomness. However, the underlying curve is still exponential. In Figure 15.6, we fit a linear model to the data using the *CrimeStat* module. The result show that the model tended

[3] The errors were calculated by, first, transforming the dependent variable by taking its natural log; second, the natural log was then regressed against the independent variables; third, the predicted values were then calculated; and, fourth, the predicted values were then converted back into raw scores by taking them as the exponents of *e*, the base of the natural logarithm. The residual errors were calculated from the re-transformed predicted values.

Table 15.4:

Predicting Burglaries in the City of Houston: 2006

Log Transformed Dependent Variable

(N= 1,179 Traffic Analysis Zones)

DepVar:	**Natural log of 2006 BURGLARIES**
N:	1,179
Df:	1,175
Type of regression model:	Ordinary Least Squares
F-test of model:	417.4 $p \leq .0001$
R-square:	0.42
Adjusted r-square:	0.42
Mean absolute deviation:	1.05
1st (highest) quartile:	1.23
2nd quartile:	0.94
3rd quartile:	0.56
4th (lowest) quartile:	1.46
Mean squared predictive error:	1.47
1st (highest) quartile:	2.02
2nd quartile:	1.14
3rd quartile:	0.47
4th (lowest) quartile:	2.24

Predictor	DF	Coefficient	Stand Error	Tolerance	VIF	t-value	p
INTERCEPT	1	1.5674	0.067	-	-	23.44	0.001
HOUSEHOLDS	1	0.0012	0.00004	0.994	1.006	28.84	0.001
MEDIAN HOUSEHOLD INCOME	1	-0.000006	0.000001	0.994	1.006	-4.09	0.001

to underestimate both the upper- and lower-ends of the distribution of X, especially the high end while overestimating the middle range.

Transforming the dependent variable into a natural log (i.e., Ln[X]) creates a better fit (Figure 15.6). Similarly, transforming the dependent variable into a square root (i.e., Sqrt[X]) is better than the linear though not as good as the log transformation (Figure 15.7). However, neither transformation are as good as fitting a true Poisson function (Figure 15.8). This can be

Figure 15.5:

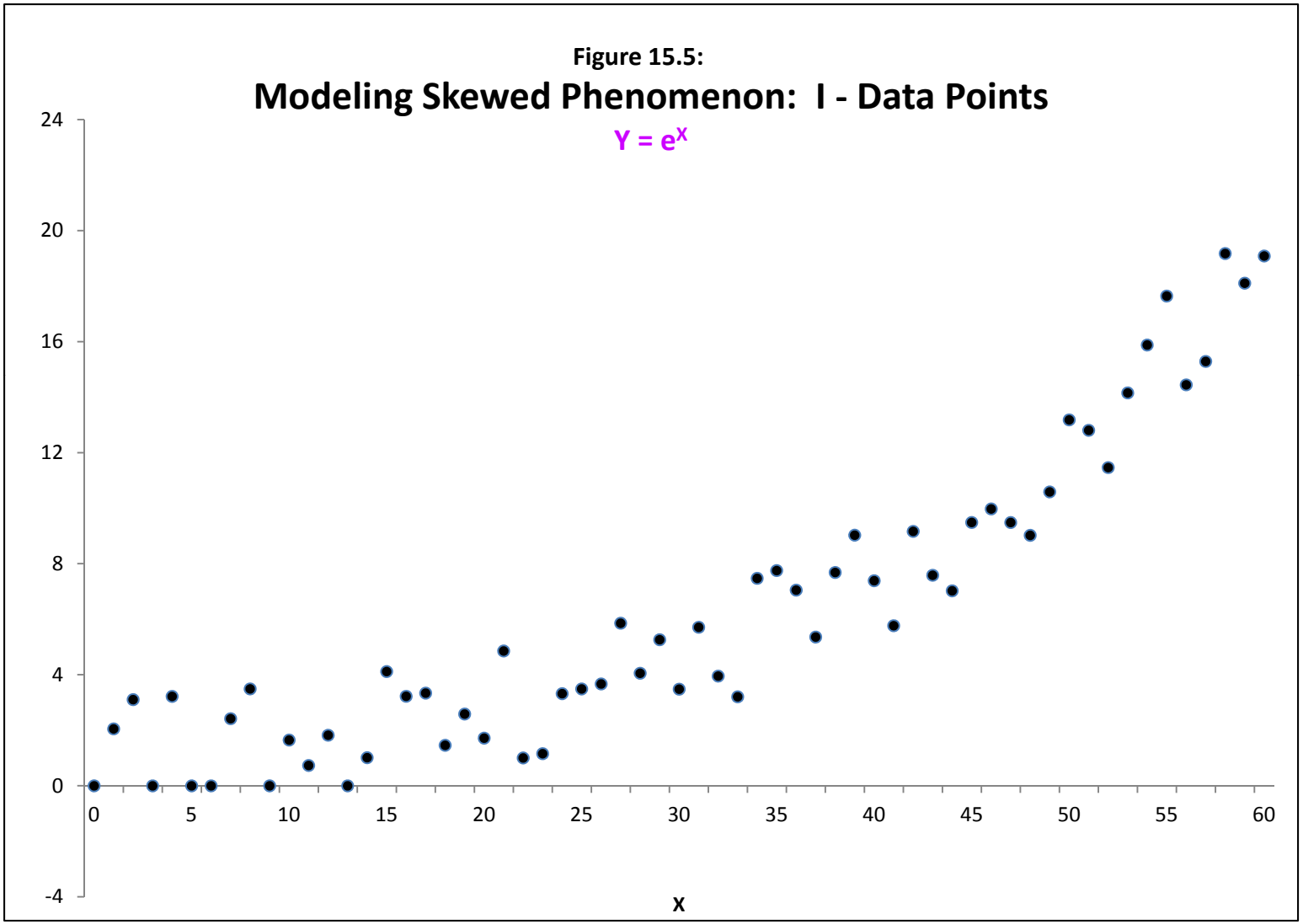
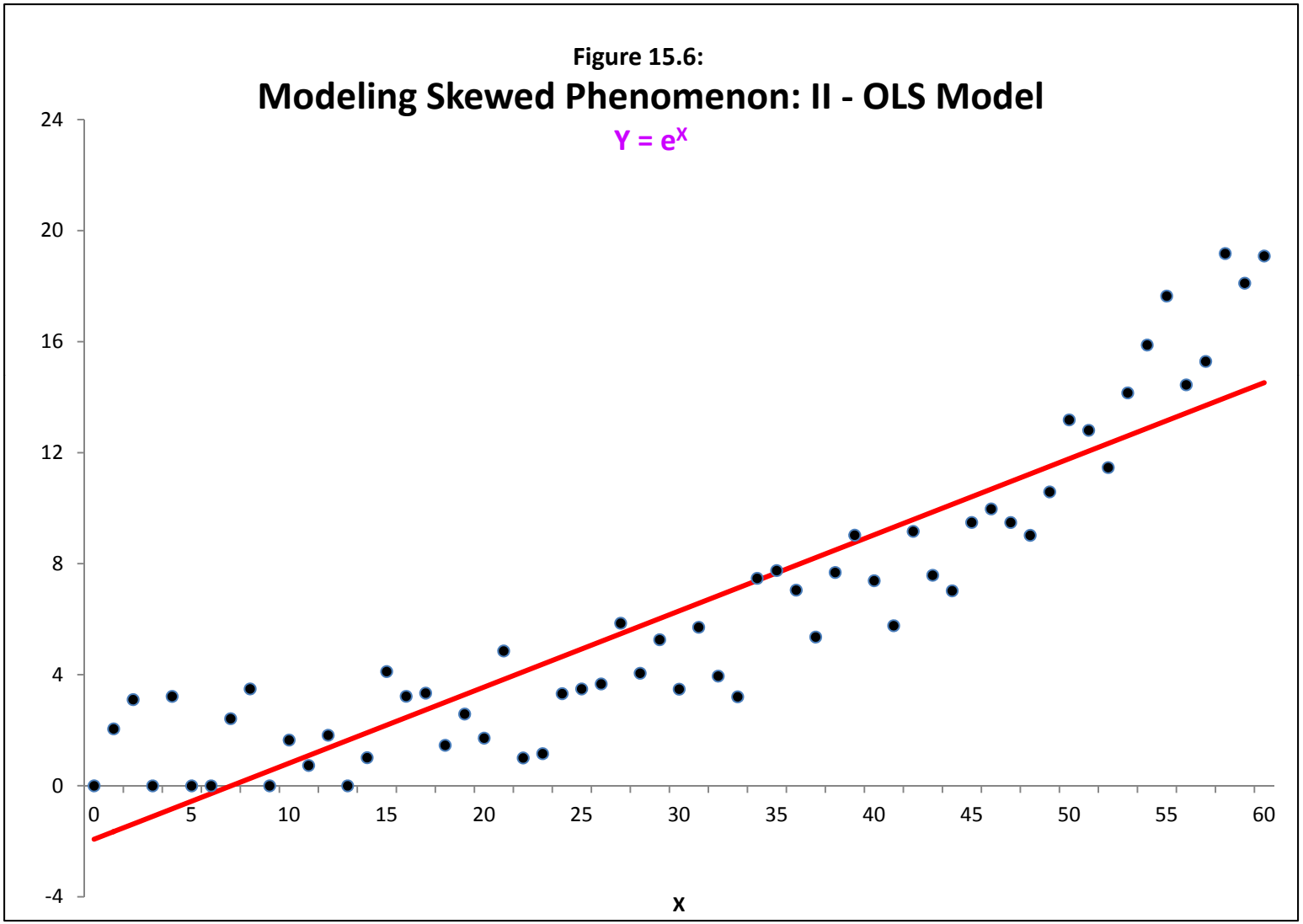

Modeling Skewed Phenomenon: I - Data Points
Y = e^X
24
20
16
12
8
4
0
-4
0
5
10
15
20
25
30
35
40
45
50
55
60
X

Figure 15.6:


Modeling Skewed Phenomenon: II - OLS Model
Y = e^X
24
20
16
12
8
4
0
-4
0
5
10
15
20
25
30
35
40
45
50
55
60
X

seen by comparing the Mean Square Predictive Error (MSPE) and the Mean Absolute Deviation (MAD) statistics including the quartiles for the MAD (Table 15.5).

Table 15.5:
Comparing Errors for Models Estimating Exponential Function
Mean Squared Predictive Error and Mean Absolute Deviation

	Model			
Error Statistic	**OLS**	**OLS w. Ln(Y)**	**OLS w. Sqrt(Y)**	**Poisson**
MSPE:	**4.96**	**1.94**	**2.57**	**1.80**
MAD:	**1.79**	**1.19**	**1.31**	**1.15**
1st quartile:	2.15	1.15	1.48	0.96
2nd quartile:	2.15	1.35	1.28	1.36
3rd quartile:	2.16	1.01	1.50	1.06
4th quartile:	1.50	1.21	0.93	1.21

As seen, the Poisson provides the best overall fit with both the MSPE and the MAD. While the OLS using the log-transformed dependent variable produces a reasonably good fit, certainly better than the OLS on the untransformed dependent variable, it still provides a poorer fit than a non-linear Poisson function, which is an exponential function. Further, the MAD for the first quartile (i.e., the data points with the highest actual values) is much worse for the OLS of the transformed dependent variable compared to the Poisson. Where the transformed dependent variable does as well if not better than the Poisson is in the last two quartiles, the low end of the X distribution.

With either the log transformation or the square root transformation, the fit is better for the low end of the dependent variable (i.e., those observations with fewer counts of the dependent variable) than for the high end. The reason is because the OLS minimizes the sum of the squared deviations of the predictions from the dependent variable. Since it assumes homoscedasticity in the residual errors across the ranges of independent variables, it cannot adjust the errors at the high end. In other words, no matter what transformation is used with an OLS, the result will always be worse than a Poisson-based model. Since we are usually interested in the high end of the dependent variable (i.e., those observations with many counts), that is a substantial deficiency of the OLS model.

Figure 15.7:

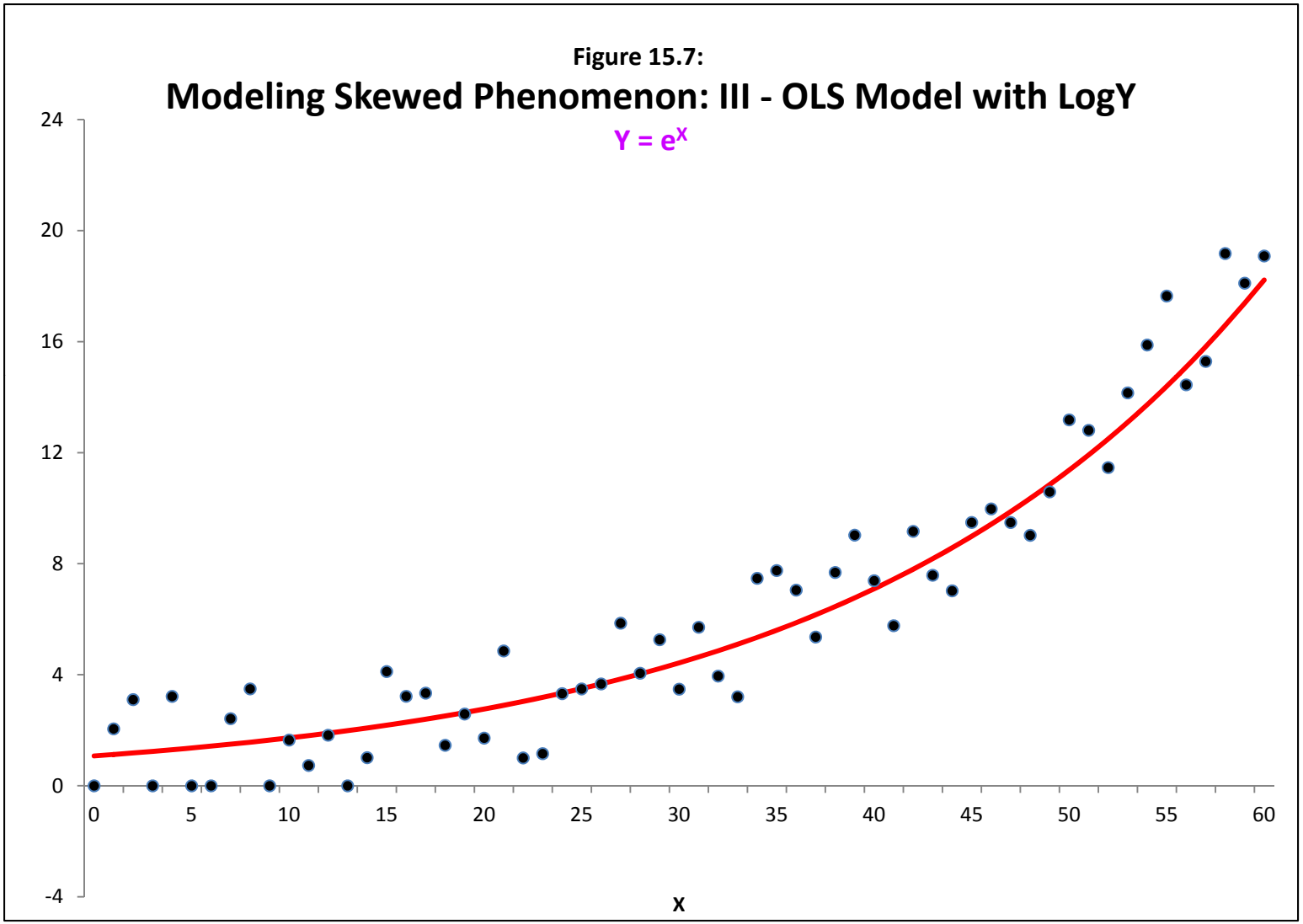

Modeling Skewed Phenomenon: III - OLS Model with LogY
Y = e^X
24
20
16
12
8
4
0
-4
0
5
10
15
20
25
30
35
40
45
50
55
60
X

Figure 15.8:

Modeling Skewed Phenomenon: IV - OLS Model with Square RootY

$Y = e^X$

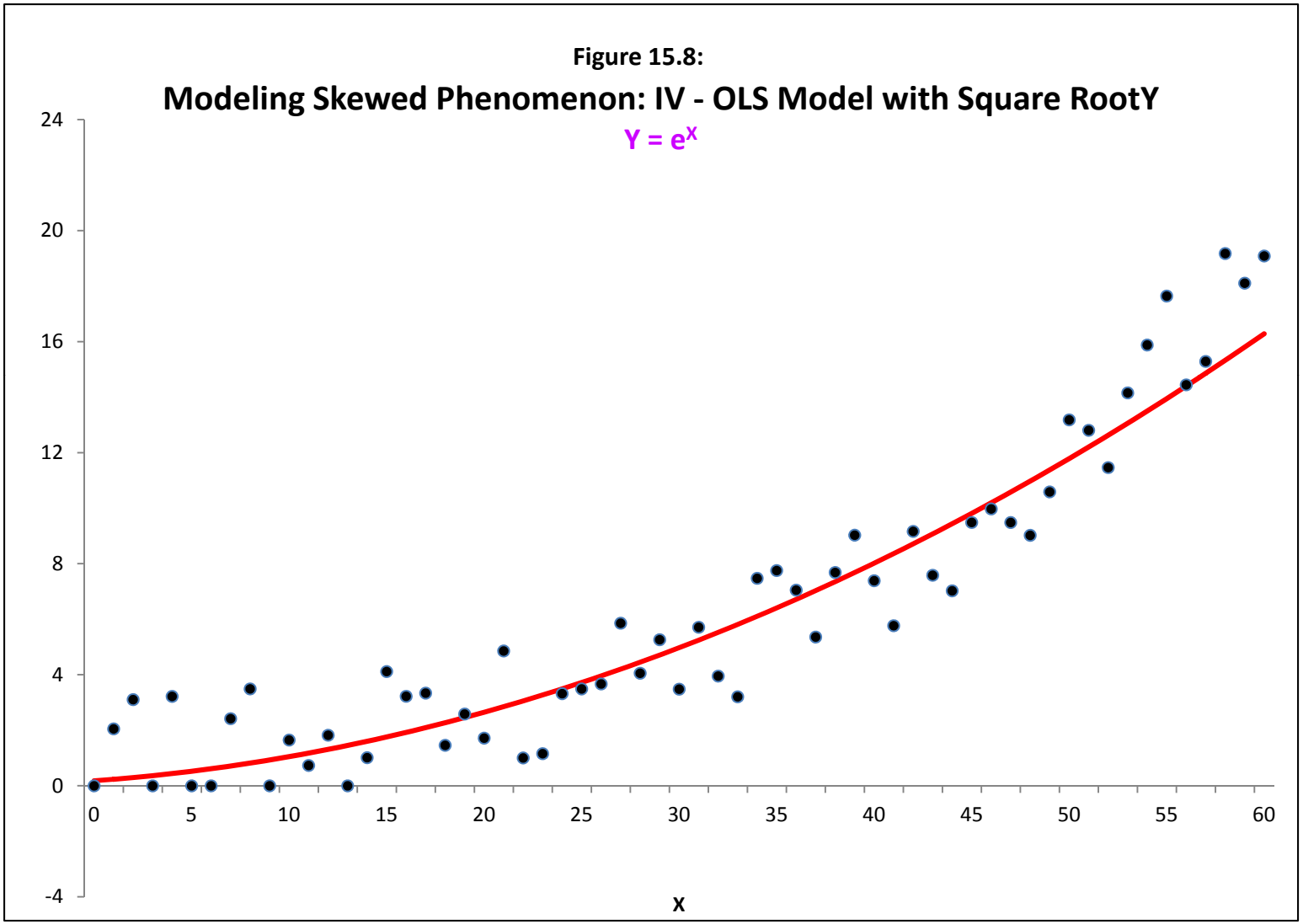
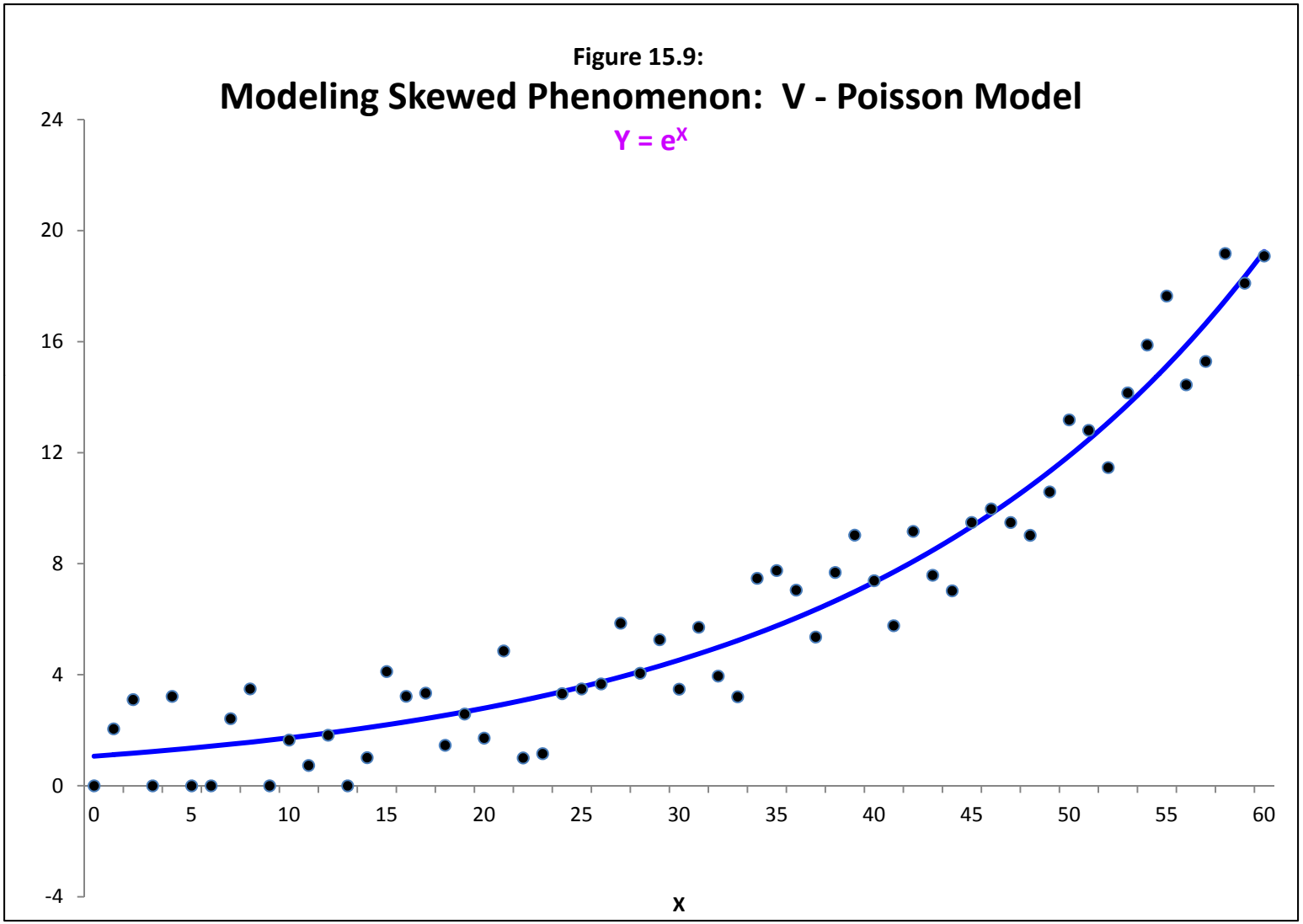


Figure 15.9:


Modeling Skewed Phenomenon: V - Poisson Model
Y = e^X
24
20
16
12
8
4
0
-4
0
5
10
15
20
25
30
35
40
45
50
55
60
X

Keep in mind that this was a created distribution where the data points were distributed equally across the X spectrum and where the errors were constant throughout (homosceadstic). With real data, a count variable (e.g., number of crimes, income) will usually be highly skewed with most observations having low values with a small percentage having high values (or vice versa such as with distance traveled) and the errors will typically increase with the value of the dependent variable.

Diagnostic Tests and OLS

To evaluate skewness and other violations of assumptions of a linear model, it is essential to examine various diagnostics about the dependent variable. The regression module has a set of diagnostic tests for evaluating the characteristics of the data and the most appropriate model to use. There is a diagnostics box on the Regression I page (see Figure 20.1 in chapter 20). Diagnostics are provided on:

1. The minimum and maximum values for the dependent and independent variables
2. Skewness in the dependent variable
3. Spatial autocorrelation in the dependent variable
4. Estimated values for the distance decay parameter – alpha, for use in the Poisson-Gamma-CAR model
5. Multicolinarity among the independent variables

Minimum and Maximum Values for the Variables

The minimum and maximum values of both the dependent and independent variables are listed. A user should look for ineligible values (e.g., -1) as well as variables that have a very high range. The MLE routines are sensitive to variables with very large ranges. To minimize the effect, variables are internally scaled when being run (by dividing by their mean) and then re-scaled for output. Nevertheless, variables with extreme ranges in values and especially variables where there are a few observations with extreme values can distort the results for models.[4] A user would be better choosing a more balanced variable than using one where one or two observations determines the relationship with the dependent variable.

[4] For example, in Excel, two columns of random numbers from 1 to 10 were listed in 99 rows to represent two variables X1 and X2. The correlation between these two variables over the 99 rows (observations) was -0.03. An additional row was added and the two variables given a value of 100 each for this row. Now, the correlation between these two variables increased to 0.89! The point is, one or two extreme values can distort a statistical relationship.

Skewness Tests

As we have discussed, skewness in a variable can distort a normal model by allowing high values to be underestimated while allowing low or middle-range values to be overestimated. For this reason, a Poisson-type model is preferred over the normal for highly skewed variables.

The diagnostics utility tests for skewness using two different measures. First, the utility outputs the "***g"*** statistic (Microsoft, 2003):

$$g = \frac{n}{(n-1)(n-2)} \sum_i [(X_i - \bar{X}) / s]^3 \tag{15.26}$$

where n is the sample size, X_i is observation i, $\bar{X}$ is the mean of X, and s is the sample standard deviation (corrected for degrees of freedom). The sample standard deviation is defined as:

$$s = \sqrt{\sum_i \frac{(X_i - \bar{X})^2}{(n-1)}} \tag{15.27}$$

The standard error of skewness (SES) can be approximated by (Tabachnick and Fidell, 1996):

$$SES = \sqrt{\frac{6}{n}} \tag{15.28}$$

An approximate Z-test can be obtained from:

$$Z(g) = \frac{g}{SES} \tag{15.29}$$

Thus, if Z is greater than +1.96 or smaller than -1.96, then the skewness is significant at the $p \leq .05$ level.

An example is the number of crimes originating in each traffic analysis zone within Baltimore County in 1996. The summary statistics were:

$\bar{X}$ = 75.108
s = 96.017
n = 325

$\sum_i [(X_i - \bar{X}) / s]^3 = 898.391$

Therefore,

$$g = \frac{325}{324 * 323} * 898.391 = 2.79$$

$$SES = \sqrt{\frac{6}{325}} = 0.136$$

$$Z(g) = \frac{2.79}{0.136} = 20.51$$

The Z of the g value shows the data are highly skewed.

The second skewness measure is a ratio of the simple variance to the simple mean. While this ratio had not been adjusted for any predictor variables, it is usually a good indicator of skewness. Ratios greater than about 2:1 should make the user cautious about using a normal model.

If either measure indicates skewness, *CrimeStat* prints out a message indicating the dependent variable appears to be skewed and that a Poisson-type model should be used.

Testing for Spatial Autocorrelation in the Dependent Variable

A fourth test that is available is a test for spatial autocorrelation in the dependent variable. It will be discussed in the spatial regression section (Chapter 19).

Multicollinearity Tests

The fifth type of diagnostic test is for multicollinearity among the independent predictors. As we have discussed in this chapter, one of the major problems with many regression models, whether MLE or MCMC, is multicollinearity among the independent variables.

To assess multicollinearity, the pseudo-tolerance test is presented for each independent variable. This was discussed above in the chapter (see equation 15.18).

MCMC Version of Normal (OLS)

There is also a Markov Chain Monte Carlo (MCMC) version of the OLS model which assumes the dependent variable is normally distributed. This will be discussed in chapter 17 on Markov Chain Monte Carlo estimation and in chapter 19 on spatial regression modeling.

References

Abraham, B. & Ledolter, J. (2006). *Introduction to Regression Modeling*. Thompson Brooks/Cole: Belmont, CA.

Berk, K. N. (1977). "Tolerance and condition in regression computations", *Journal of the American Statistical Association*, 72 (360), 863-866.

Draper, N. & Smith, H. (1981). *Applied Regression Analysis, Second Edition*. John Wiley & Sons: New York.

H-GAC (2010). Transportation and air quality program, *Houston-Galveston Area Council*. http://www.h-gac.com/taq/.

Hilbe, J. M. (2008). *Negative Binomial Regression (with corrections)*. Cambridge University Press: Cambridge.

Kanji, G. K. (1993). *100 Statistical Tests*. Sage Publications: Thousand Oaks, CA.

Miaou, S. P. (1996). *Measuring the Goodness-of-Fit of Accident Prediction Models*. FHWA-RD-96-040. Federal Highway Administration, U.S. Department of Transportation: Washington, DC.

Microsoft (2003). "SKEW - skewness function", *Microsoft Office Excel 2003, Microsoft: Redmond, WA*.

Myers, R. H. (1990) Classical and Modern Regression with Applications, 2nd edition, Duxbury Press, Belmont, CA.

Oh, J., Lyon, C., Washington, S., Persaud, B., & Bared, J. (2003). "Validation of FHWA crash models for rural intersections: lessons learned". *Transportation Research Record 1840*, 41-49.

StatSoft (2010). "Tolerance", *StatSoft Electronic StatisticsTextbook,* StatSoft:Tulsa, OK. http://www.statsoft.com/textbook/statistics-glossary/t/button/t/

Tabachnick, B. G. & Fidell, L. S. (1996). *Using Multivariate Statistics* (3rd ed). Harper Collins: New York.

Train, K. (2009). *Discrete Choice Methods with Simulation* (2nd edition). Cambridge University Press: Cambridge.

References (continued)

Venables, W.N. & Ripley, B. D. (1997). *Modern Applied Statistics with S-Plus (second edition)*. Springer-Verlag: New York.

Wikipedia (2010b). "Maximum likelihood", *Wikipedia*. http://en.wikipedia.org/wiki/Maximum_likelihood. Accessed March 12, 2010.