6 Modeling Relationships

Regression and Geographically Weighted Regression

THEORY

Learning Objectives

This chapter deals with

- Linear regression notions and mathematical formations
- The assumptions of linear regression
- Identifying linear relationships between a dependent variable and independent variables
- Residual plots, leverage and influential points
- Evaluating the degree (effect) to which one variable influences the positive or negative change of another variable
- The geometrical interpretation of the intercept, the slope, and metrics and tests used to assess the quality of the results
- The interpretation of coefficients
- Studying casual relationships (under specific conditions)
- Building predictive models by fitting a regression line to the data
- Evaluating the correctness of the model
- Model overfitting
- Ordinary least squares (OLS)
- Exploratory OLS
- Geographically weighted regression (GWR)

After a thorough study of the theory and lab sections, you will be able to

- Distinguish which regression method is more appropriate for the problem at hand and the data available
- Interpret statistics and diagnostics used to evaluate a regression model
- Interpret the outcomes of a regression model such as the coefficient estimates, the scatter plots, the statistics and the maps created
- Identify if relationships, associations or linkages among dependent and independent variables exist
- Analyze data through regression analysis in Matlab

- Conduct exploratory regression analysis in ArcGIS
- Conduct geographically weighted regression in ArcGIS

6.1 Simple Linear Regression

Definition

Simple linear regression (also called bivariate regression) is a statistical analysis that identifies the linear relationships between a dependent variable, *y*, also called response, and a single independent variable, *x*, also called explanatory variable (for multiple independent variables, see Section 6.2). Independent *x* is any variable that is used to explain/predict a single variable that is called dependent *y*.

Linear regression analysis is based on the fitting of a straight line to the dataset in order to produce a single equation that describes the dataset. The equation of the regression line is (6.1) (see Figure 6.1A and B):

$$\hat{y} = a + bx \tag{6.1}$$

where

$\hat{y}$ is the predicted value (also called fitted value) of the dependent variable (it is pronounced "y-hat"; the observed [real] value of the dependent variable is denoted as *y* – without a hat)
x is the independent variable
a is the intercept
b is the slope of the trend line

Slope (*b*) and intercept (*a*) are also called coefficients and are the parameters that have to be estimated. Methods to estimate the parameters of a linear regression include ordinary least squares (see Section 6.1.2), Bayesian inference, least absolute deviations and nonparametric regression.

The difference of the observed value with the predicted (fitted) value is the model error called residual *e* (6.2).

$$e = y - \hat{y} \tag{6.2}$$

When residuals are positive, there is underprediction implying that the predicted (fitted) value is lower than the observed (see Figure 6.1.B). On the other hand, overprediction means that the fitted value is higher than the observed value. In this case, residuals are negative.

In regression analysis, we build a probabilistic model (6.3) consisting of a deterministic component ($\hat{y}$) and a random error component (*e*). In particular, each observation of the dependent variable is the sum of the predicted value with the residual (6.3) (see Figure 6.1B):

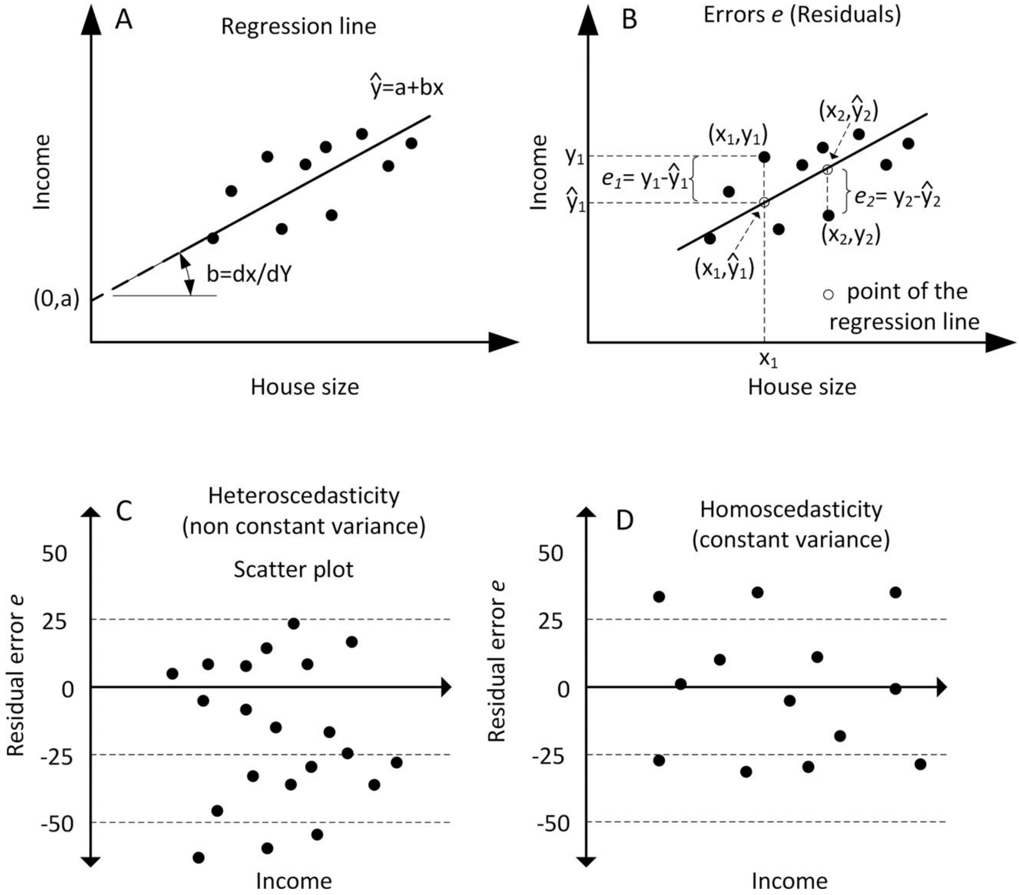


Figure 6.1 Regression basics. (A) A regression linear fit. (B) Calculating residuals. (C) Depicting heteroscedasticity with residual plot. (D) Depicting homoscedasticity with residual plot. These residual plots are not standardized. The *y*-axis has the same units as the dependent variable.

$$y = \hat{y} + e = a + bx + e \tag{6.3}$$

In deterministic models, outcomes (a) are precisely determined by well-defined equations, (b) no random variation exists and (c) a given input always produces the same output. In contrast, a probabilistic (stochastic) model handles complex problems where uncertainty exists. Errors are inevitable and should be included in the model. If under- or overpredictions occur in a systematic way, then the regression model is biased.

Interpretation

The intercept (*a*) is the value of the dependent variable when the independent variable is zero. The slope (*b*) is the expected change of the dependent variable for a one-unit change in the independent variable (Lee & Wong 2001).

Why Use

Simple linear regression is used to (Rogerson 2001 p. 104)

- Identify relationships between a dependent variable and an independent variable
- Evaluate the impact (importance) of the independent variable on the dependent variable
- Build a predictive model by fitting a regression line to the data
- Evaluate the correctness of the model

It should be emphasized that simple regression does not study casual relationships. Any hypothesized causal relationship can be tested under specific conditions within the multiple linear regression context presented in Section 6.2.

Discussion and Practical Guidelines

Suppose we study if “Income” is related to “House size.” We can calculate the correlation coefficient between “Income” and “House size,” which measures the direction and the strength of a relationship. For example, a strong positive correlation coefficient would reveal that when income increases, house size increases as well. What correlation cannot estimate is the change in “Income” for a given increase in “House size” or the expected (predicted) “Income” for a given “House size” value. By regression modeling, we can analyze more thoroughly the relationship between a dependent (Income) and an independent variable (House size) by fitting a straight line among the data points (see Figure 6.1B). The equation of the line reveals the trend and can be used for predicting purposes. For example, if we use satellite data to measure the size of a house, we can calculate the expected income of that specific household by plugging in the ”House size” value to the regression function (which has been constructed already based on other data). Such models and approaches provide quick and economic ways to estimate socioeconomic data based on free, easily acquired imagery.

Regression works by fitting a line to the data points by determining parameters *a* and *b*. There are many ways to fit a line in a dataset. We can simply draw the mean value of *y* as a parallel line of the *z*-axis. We can connect the leftmost to the rightmost data point or draw a line that has the same number of points above and below the line. In fact, regression is much more than a fitting technique. It is a statistical inference approach, estimating the unknown parameters (coefficients).

The following subsections present the assumptions that a simple linear regression model is based on, the geometrical interpretation of the intercept and the slope and various metrics and tests used to assess the quality of the results.

6.1.1 Simple Linear Regression Assumptions

Simple linear regression builds a probabilistic model based on four assumptions (use the word LINE as mnemonic, standing for the first letter of each assumption):

1. **Linearity**. The relationship between *y* and *x* is linear. The linearity assumption can be tested by using scatter plots. Outliers and influential observations should be considered for exclusion, prior to any other analysis (more on this in Section 6.3).
2. **Independence – No autocorrelation**. The residuals are independent from each other, meaning that an error *e* in a data point *x* does not affect by any means another error *e* of some different value of the same variable *x*. The Durbin–Watson test statistic can be used to detect the presence of autocorrelation in the residuals of regression analysis.
3. **Normality.** The distribution of errors *e* at any particular *x* value is normal having zero mean value. This means that for fixed values of *x*, *y* has a normal distribution as well. Assumption can be checked by using a histogram and a fitted normal curve, a Q-Q plot, Jarque–Berra test, etc. (see multiple linear regression, Section 6.2).
4. **Equality – Homoscedasticity.** Errors *e* have a zero mean, equal variance and a constant standard deviation at any particular value of *x*. This is also called homoscedasticity and exists when the variance of residuals does not change (increase or decrease) with the fitted values of the dependent variable. Heteroscedasticity, on the other hand, exists when the variance is not constant. Residual plot inspection is a straightforward way to trace if heteroscedasticity exists (Figure 6.1C and D).

6.1.2 Ordinary Least Squares (Intercept and Slope by OLS)

Definition

Ordinary least squares (OLS) is a statistical method for estimating the unknown parameters (coefficients) of a linear regression model. The regression using this method is also called OLS linear regression, or just linear regression. Other methods used to estimate these parameters include the Bayesian inference, the least absolute deviations and the nonparametric regression. The parameters *a* and *b* are called intercept and slope, respectively (*b* is called the coefficient in multiple linear regression; see Section 6.2). OLS regression determines these values by minimizing the sum of the squared vertical distances from the observed points to the line (sum of the squared residuals – Eqs. [6.4,6.5,6.6], Figure 6.1B, Rogerson 2001 p. 107). The line produced is the least-squares line.

$$min_{a,b} = \sum_{i=1}^{n} (y - \hat{y})^2 \tag{6.4}$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6.5}$$

$$a = \bar{y} - b\bar{x} \tag{6.6}$$

where

$\bar{x}$ is the mean of the independent variable
$\bar{y}$ is the mean of the observed values of the dependent variable
n is the number of data points

Interpretation

The regression line is the line with a slope that equals *b*, passing through the point (*0,a*) and point $(\bar{x}, \bar{y})$ (Figure 6.1B, Rogerson 2001 p. 109) The slope denotes that for each unit of change in the independent variable, the mean change in the dependent variable is *b*. We should highlight that the regression line of *y* on *x* should not be used to predict *x* on *y*. In other words, we cannot use this equation to predict *x* values if we have *y* values available because the regression line has not been constructed to minimize the sum of squared residuals in the *x* direction (Peck et al. 2012).

6.2 Multiple Linear Regression (MLR)

6.2.1 Multiple Regression Basics

Definition

Multiple linear regression (MLR) analysis identifies the linear relationships between a dependent variable *y* and a set of independent variables, also called explanatory variables *x* (Rogerson 2001 p. 124). Regression analysis fits a straight line to the dataset to produce a single equation that describes the data. For *m* independent variables and *n* observations, the regression equation (deterministic function) is

$$\hat{y}_\imath = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_m x_{im} \tag{6.7}$$

where

$\hat{y}_\imath$ is the fitted value for the *i*-th observation.
$i = 1, \ldots n$.
n is the total number of observations.
b_0, b_1, $b_2 \ldots b_m$ are the coefficients.
b_0 is the intercept. It is the value of the equation if all variables or coefficients have zero value.
m is the total number of the independent variables.
x_{im} is the *i*-th observation on the *m*-th independent variable.

The regression probabilistic model is (6.8) (O'Sullivan & Unwin 2010 p. 226):

$$y_i = \hat{y}_i + e = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_m x_{im} + e_i \tag{6.8}$$

where

y_i is the *i*-th value of the dependent variable

e_i is the *i*-th error of the model (random deviation from the deterministic function, Eq. 6.7).

In matrix terms, the model can also be expressed as (6.9, 6.10) (O'Sullivan & Unwin 2010 p. 226):

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} b_0 \\ \vdots \\ b_m \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \tag{6.9}$$

$$Y = Xb + e \tag{6.10}$$

where

Y is the vector containing the *n* observations for the dependent variable
b is the vector of the estimated regression model coefficients
X is the data matrix containing the *n* observations for the *m* independent variables (along with a column of ones)
e is the error term

Interpretation

Each coefficient b_i can be interpreted as the change (impact) in mean *y*, for a one unit change of x_i, holding the remaining independent variables constant. Regression coefficients express how much impact an independent variable has on the dependent variable (de Vaus 2002 p. 279). The sign of the coefficient indicates the direction of the impact.

Why Use

MLR is used to (Rogerson 2001 p. 104)

- Identify linear relationships between a dependent variable and independent variables
- Evaluate the degree (effect) to which one variable influences the positive or negative change of another variable
- Build a predictive model by fitting a regression line to the data
- Study casual relationships (under specific conditions)

Discussion and Practical Guidelines

For most real-world problems, a plethora of variables are available, and simple linear regression is inadequate. Most of the concepts and terms used in simple linear regression can be applied in MLR with no or only a slight modification. The calculations though in MLR are extremely tedious in comparison to simple linear regression. MLR is not preferred solely due to variables' abundance. Explaining a dependent variable by a single independent variable, most of the time, is insufficient, as their relationship may be weak. In such case, we should add extra independent variables to increase the explained variation of the dependent variable. For example, estimating the value of a house is not just a matter of its size, although size is strongly related to price. Other variables that influence the value of a house include the year of construction (if it's new or old), the type of house (apartment, detached house, etc.) and other spatial variables such as the distance from city center, the distance from transportation network (e.g., subway, highway) or the location (i.e., neighborhood, city, county, country in which the house lies). Still, integrating many variables does not necessarily lead to a better model. On the contrary, including many independent variables is likely to lead to model overfitting, which is undesirable (see Section 6.2.2).

MLR can be also used to identify causal relationships (causes and effects) among the dependent and the independent variables. Still, not all types of MLR can establish causal relationships. Identifying causes and effects is a wider analysis lying in the explanatory analysis filed. Explanatory analysis attempts to identify the factors or mechanisms that produce a certain state of some phenomenon (Blaikie 2003). From the social analysis perspective, explanatory analysis attempts to trace the causes that create an effect, by carrying out controlled experiments. Such experiments establish a linking relationship between a variable (cause) and one or more variables (effects) by controlling the order in which causes and effects (and related population's groups) are analyzed (Blaikie 2003). Regression analysis may be used for identifying causes and effects, but not all regression models can be used for this purpose (see Section 6.2.6). In other words, although some regression models may be robust and capable to identify linear relationships among variables, this does not straightforwardly lead to the conclusion of a cause-and-effect relationship (see also Section 2.3.4). In addition, it has to be emphasized that any **hypothesized causal relationship** is one way, meaning that the dependent variable is responsive to the independent variable and not the other way around (Longley et al. 2011 p. 357). That is why the dependent variable is also called the response variable, and the independent variable is called the predictor. It should also be noted that in case of spatial analysis and due to spatial heterogeneity and spatial dependence, the discovered casual relationships through regression modeling might change from place to place, and that is why a spatial aspect of regression should be considered. For more details on spatial regression, see Section 6.5 and Chapter 7.

Finally, prior to any regression analysis, there is a set of decisions that should be determined (explained in the following subsections):

- Choose the variables to be used to avoid overfitting (Section 6.2.2)
- Check the dataset for missing values (Section 6.2.3)
- Check the dataset for outliers and leverage points (Section 6.2.4)
- Check if any dummy variable is needed (Section 6.2.5)
- Select the method of entering variables in the model (Section 6.2.6)

6.2.2 Model Overfit: Selecting the Number of Variables by Defining a Functional Relationship

Definition

An **overfit model** is a model that performs well on a specific dataset but cannot generalize new solutions (to newly presented data) with accuracy.

Discussion and Practical Guidelines

In general, overfitting occurs when a model has too many parameters relatively to the number of observations. The number of variables to be included in the model depends on the problem. The first step is to identify if any functional relationship (or else a meaningful relationship or causation) among the variables under consideration exists through an established theoretical background. Solid comprehension of such background leads to the identification of potential independent variables. The regression model will then carry out the burden to convert this functional relationship into a meaningful mathematical equation.

We should avoid the temptation to use as many variables as possible. The more variables included, the more likely for an overfitted model. Model overfitting is not desirable, as it leads to a model with poor predictive performance. One way to avoid overfitting is to have considerably more observations than the number of variables. The following rules of thumb can be used for selecting the number of variables in relation to the number of observations available (de Vaus 2002 p. 357):

- When using standard MLR the ratio of observations to variables should be larger than 20/1 (Tabachnick et al. 2012).
- The sample size should be at least 100 + *k*, where *k* is the number of independent variables (Newton & Rudestam 1999)
- In case of a non-normal distributed dependent variable, samples should be increased to more than 100.

These rules of thumb are for general guidance. The final decision on the choice of variables should be made in relation to the problem at hand and the data available. A smooth approach is to initially build a model that is simple, explaining a large proportion of variation, that is also backed up by a solid theoretical background. By using the *R*-squared adjusted, we can determine which variables are useful and which are not by observing the fluctuations in the *R*-squared adjusted value (see Section 6.3.3). Exploratory regression explained in Section 6.7 provides such approach. Section 6.2.6 also presents the effect that the order of variables' entry infers to the model.

6.2.3 Missing Values

Before we begin our analysis, we should inspect if missing values exist. Three of the most common ways to confront missing values are

- Deleting observations that contain missing values.
- Replacing the missing values of a variable with the mean value of this specific variable. Although in this approach, we retain all observations, it is not guaranteed that the final results will be meaningful. For instance, if we miss age values and we use the mean value of age for the entire variable for those missing, then we will probably get wrong estimations.
- Estimate the missing values by using simple OLS, or MLR , of other variables with no missing values. For example, if the "House size" variable has no missing values and some respondents refused to reveal their income, we can build a simple linear regression of "Income" and "House size." If the model is statistically significant and all related diagnostics are checked, we can estimate/predict the income of the missing observations by using the known size of a house.

6.2.4 Outliers and Leverage Points

Outliers and high **leverage points** change a lot the results of any analysis (see Section 6.3.11). There are mainly three approaches to deal with such observations in regression analysis.

- First, trace and remove outliers before regression analysis takes place. To do so, one should typically follow the procedures for tracing outliers as explained in Section 2.2.7. As long as outliers are removed, then regression analysis is carried out along with the calculation of high leverage values (see Section 6.3.11).
- Second, perform MLR including all data, and create the standardized residual plots where both outliers and high leverage points can be traced. After we trace influential points, we may remove them and test the model again. If the model is significantly improved, we can drop these influential observations (see how this is done in detail in Section 6.3.11). Any observation that leads to a large residual should be scrutinized.
- Apply robust regression that is specifically designed to handle outliers and heteroscedasticity (not further explained in this book).

In general, we should check if an outlier reflects an error or if it is an actual value revealing unusual behavior/status. Although outliers are usually removed, it depends on the studied problem and the researcher's choice of how outliers will be handled.

6.2.5 Dummy Variables

Definition

A **dummy variable** is a binary variable getting values either 1 or 0, indicating the presence (1) or absence (0) of some categorical effect (in the case of a dependent binary variable, we consider logistic regression). If a categorical variable is decomposed to its categories, then each single category can be a new variable, named a dummy variable.

Why Use

Apart from ratio variables, categorical independent variables may be integrated to a regression model. For example, *Income* is often classified into a category (i.e., *Low*, *Average* or *High*). A person with average income is assigned a value of 1 for the category Average and 0 for the two others. *Place of living* may also be reported as a nominal variable such as *Urban*, *Suburban* and *Rural*. A person living in rural area is assigned a value of 1 for *Rural* and 0 for the other two categories (*Urban* and *Suburban*). Other examples include *Gender* (i.e., *Male* or *Female*) and *Land Use* (i.e., *Forest*, *Water*, *Urban* and *Shrubland*; Grekousis et al. 2015a).

Discussion and Practical Guidelines

To handle dummy variables with k categories, we create $k - 1$ variables by omitting one category to avoid multicollinearity (Rogerson 2001 p. 128). For example, for the variable *Place of living*, suppose that X_1 is the *Urban* category, X_2 is the *Suburban* category and X_3 is the *Rural* category. The following regression equation would be inappropriate, violating the non-multicollinearity assumption for MLR (see Section 6.4), since the sum of all k columns should always equal 1 (Rogerson 2001 p. 129) (6.11):

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3. \tag{6.11}$$

We can arbitrarily omit the category "Rural" (6.12):

$$Y = b_0 + b_1X_1 + b_2X_2. \tag{6.12}$$

A person living in a rural area gets 1 for *Rural* and 0 for *Urban* and *Suburban*. In other words, we can determine from only two out of three variables where a person lives. Once dummy variables are defined, then regression analysis proceeds at the usual fashion.

Dummy variables can be combined to ratio variables within the same regression model. For example, to model $Y = Income$, the ratio independent variables X_1 = *Years of education* and X_2 = *House size* can be combined to the categorical variable *Place of living* (three categories: *Urban*, *Suburban* and *Rural*). From the three categories, we create two dummy variables, X_3 = *Urban* and X_4 = *Suburban* by arbitrarily omitting *Rural*.

The regression equation is (6.13):

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 \tag{6.13}$$

$$Income = b_0 + b_1 YearsEdu + b_2 HouseSize + b_3 Urban + b_4 Suburban \tag{6.14}$$

The model would lead to different coefficients but precisely the same conclusions if we dropped any other category.

Interpreting Coefficients of Dummy Variables

The regression coefficients for the dummy variables can be interpreted in relation to the omitted category (Rogerson 2001 p. 129). For Eq. (6.14), if a person lives in an urban area, then X_3 gets 1 and X_4 gets 0. Keeping constant the other variables, then *Income* is, by b_3 units, higher for a person living in an urban area, relatively to the omitted category (it would be less if b_3 were negative).

Being located in *Suburban* means that X_4 gets 1, X_3 gets 0 and *Income* is by b_4 units different, relatively to *Rural*, keeping constant the other variables. Finally, being located in *Rural* implies that both X_3 and X_4 are 0, and by keeping constant the other variable, *Income* is b_0, which is the intercept.

Suppose we add two more categorical variables, namely *Gender* and *Age*. *Gender* consists of two categories, *Male* and *Female*, and *Age* consists of three, *Young*, *Middle* and *Old*. We can add these variables in the previous model to test whether *Income* is further explained by these variables (6.15). We choose to omit the *Male* and *Old* categories:

$$\begin{aligned} Income = b_0 + b_1 YearsEdu + b_2 HouseSize + b_3 Urban \\ + b_4 Suburban + b_5 Female + b_6 Young + b_7 Middle \end{aligned} \tag{6.15}$$

Suppose that the coefficients of the model are (6.16):

$$\begin{aligned} Income = 2{,}500 + 10 YearsEdu + 0.5 HouseSize + 1{,}500 Urban \\ +750 Suburban - 500 Female - 1{,}000 Young + 2{,}200 Middle \end{aligned} \tag{6.16}$$

The coefficients will be interpreted relatively to the omitted categories. Thus, a person living in a city earns on average \$1,500 more than a person living in rural areas. A woman earns \$500 less than a man. A young person earns \$1,000 less than an old person.

From the policy perspective, this type of analysis is appropriate for evaluating various scenarios. Suppose we build a regression model for land use changes, based on population, income, GDP and other socioeconomic variables. Based on the coefficients produced from the model, various scenarios can be tested on projected pressures on land use changes – for example, of a 20% urban population increase or a 10% GDP increase within the next decade (Grekousis et al. 2016).

6.2.6 Methods for Entering Variables in MLR: Explanatory Analysis; Identifying Causes and Effects

Although for some regression models, controlling the order in which variables are entered does not matter, for others, it offers the choice to conduct statistical experiments and identify causes and effects through explanatory analysis (de Vaus 2002 p. 363). There are four main methods of MLR based on the sequence that the independent variables are entered in the regression analysis:

- **Standard or single-step regression**, when all variables are analyzed in a single step. This method (a) identifies which independent variables are linked to the dependent variable and (b) quantifies how much of the total variation of the dependent variable is explained (through *R*-square adjusted). In this method, we explore linkages, not causal relationships (de Vaus 2002 p. 360).

  When to use: When we are dealing with a small set of variables and we do not have a clear view of which independent variables are appropriate, when the main task is the identification of linkage and not causation.
- **Stepwise regression** is used when we have more than one independent variable and we want to use an automated procedure to select only those significant to the model (those that explain as much variation as possible of the dependent variable). Instead of calculating statistics and diagnostics for the whole dataset in a single step (as in the standard method), variables are added or removed to the model sequentially. When data are added to the model, the method is called forward stepwise regression, starting from a simple regression model having a few variables or just the constant. On the other hand, when variables are removed, the method is called backward stepwise, and the initial model contains all variables. The order of entry plays a crucial role in stepwise regression. In cases where variables are uncorrelated, the order is not an issue and does not affect results. Still, it is hard to find completely uncorrelated variables in a dataset. In cases of correlated variables, order matters. Although the final adjusted *R*-squared value will be the same, no matter what the order is, the relative contribution of each variable to the model will change, resulting in different coefficients. Getting different models from the same set of potential terms is a

disadvantage of the method; by selecting different criteria/order for entering or removing variables, stepwise yields different results.

When to use: When we want to use an automated method to select the appropriate variables, when we want to know how much additional variance each variable explains, when we want to maximize the prediction capabilities of the model with as few variables as possible (de Vaus 2002 p. 365).

- The **hierarchical method** is used when the order of entering or removing the independent variables is determined by the researcher (also called hierarchical regression). Controlling the order of entry allows for statistical experimentation and the identification of casual relationships by testing how much the adjusted *R*-squared increases each time a new variable is added/removed. By applying hierarchical regression, we use the same statistics and diagnostics as in MLR. The difference lies in the way we built the sequence of the independent variables. The focus is on the change in the predictability of the model associated with independent variables entered later in the analysis in comparison to those entered earlier. The order is based on the scope of the analysis, the associated theory and the intuition of the analyst.

  More specifically, applying the hierarchical method allows for

  (a) Testing theories and assumptions by conducting experiments and controlling the order of entry.
  (b) Calculating the extra exploratory power of each variable or block of variables.
  (c) Controlling for confounding variables. There are cases in which two variables (i.e., dependent and independent) exhibit correlation but with no causal relationship between them. A confounding variable is a third variable that is associated with both the dependent and the independent variables and causes extraneous changes. It is the confounding variable's impact that explains the variance on both the dependent and independent variables. In this sense, hierarchical regression allows for the identification of causal relationships (de Vaus 2002 p. 365).

  When to use: To have more control over the analysis as well as to test causal relationships.

- **Exploratory regression** runs all models' combinations to find the optimal one. It applies OLS regression to many different models to select those that pass all necessary OLS diagnostic tests. A close inspection of the detailed results and models can lead to a robust analysis similar to hierarchical regression. In other words, exploratory regression, if used wisely, allows for assessing the exploratory power of each variable as well for controlling of confounding variables by cross-comparing the models that better reflect the tested theories and assumptions. Still, exploratory regression is a data-mining approach, more similar to stepwise regression and not to hierarchical regression, as it does not allow for much experimentation (ESRI 2016a).

When to use: When we want to test all combinations in an automated way, when we want to identify causal relationships.

6.3 Evaluating Linear Regression Results: Metrics, Tests and Plots

There are various metrics, tests and graphs to evaluate the performance of a regression model. In the next subsections, the following methods are presented:

- Multiple *R*
- Coefficient of determination *R*-squared
- Adjusted *R*-squared
- Predicted *R*-squared
- F-test
- T-statistic
- Wald test
- Standardized coefficients (beta)
- Residual plots and standardized residual plots
- Influential points (outliers and leverages)

6.3.1 Multiple *r*

Definition

Multiple *r* is the absolute value of the correlation between the observed y (response) and the estimated $\hat{y}$ predicted by the regression equation.

Interpretation

Multiple *r* measures correlation and can be interpreted accordingly. In simple linear regression, the sign of *b* coefficient (slope) reveals the positive or negative correlation. Multiple *r* is calculated by the same way for multiple regression as well; that is why it is called "multiple." Still, multiple *r* is not very indicative of assessing the results of regression analysis, and it is not commonly analyzed further.

6.3.2 Variation and Coefficient of Determination *R*-Squared

Definition

Before we define *R*-squared, let's define the following quantities (see Figure 6.2):

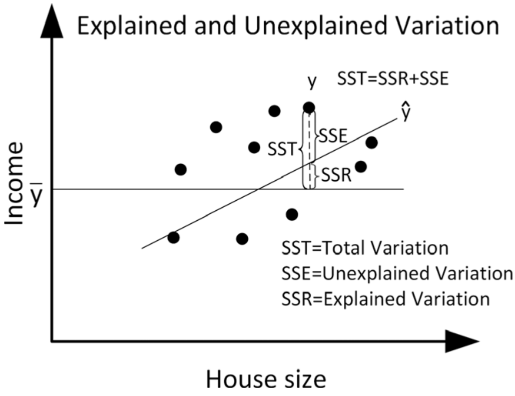


Figure 6.2 Partitioning the variation in *y*. The distance of the total variation, SST, is the summation of SSR and SSE distances.

Sum of Squared Regression (6.17), also called explained variation (of the model):

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{6.17}$$

It is a measure of the explained variation or the amount of variation in *y* that can be modeled by a linear relationship to *x*.

Sum of Squared Error (6.18), also called unexplained variation, or residual sum of squares:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2 \tag{6.18}$$

It is a measure of the unexplained variation or the amount of variation in *y* that cannot be modeled by a linear relationship to *x*.

Sum of Squared Total (6.19) is a measure of the total variation of *y*:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{6.19}$$

The total variation in *y* can be calculated as:

Total variation = *Explained Variation by Regression* + *Unexplained variation*,

$$\text{which is equivalent to } SST = SSR + SSE \tag{6.20}$$

In practice, the variation of *y* can be considered as the result of two quantities: (a) the variation explained by the linear model assumed to describe better the data and (b) the unexplained variation that the linear model did not achieve to explain (Linneman 2011 p. 226).

Coefficient of determination denoted as R^2 (*R*-squared) is the *Percent* of the variation explained by the model (Figure 6.2; de Vaus 2002, p. 354). It is calculated as the ratio between the variation of the predicted values of the dependent variable (explained variation *SSM*) to the variation of the observed values of the dependent variable (total variation *SST*; Eq. 6.21). It equals the squared correlation coefficient (see Figures 6.1D and 6.2):

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6.21}$$

Put simply, it is the squared sum of the difference of predicted values $\hat{y}_i$ with the mean value of the observed values $\bar{y}$ (squared sum of regression), divided by the squared sum of the difference of the observed values y_i, with the mean value of the observed values $\bar{y}$. The coefficient of determination is also calculated by the following formula (6.22):

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6.22}$$

As the residual sum of squares is minimized in the ordinary least squares regression, the ratio *SSE/SST* is always smaller than 1.

Why Use

R-squared is used to assess if a linear regression model fits well to the data points. It provides information about the goodness of fit of the model.

Interpretation

The coefficient of determination ranges from 0 to 1 and expresses the percentage of the variation explained by the model (see Figure 6.3). A zero value indicates that no variation is explained and, thus, the model cannot be used, whereas a value of 1 explains all data variability perfectly fitting a line on the data points. For example, if *R*-squared is 82%, this typically means that 82% of the variation in *Y* is due to the changes in values of *X*, meaning that *X* is a good predictor for *Y* having a probabilistic linear relationship. The smaller the unexplained variation, the smaller the ratio (in Eq. 6.22) and the larger the coefficient of determination or the larger the explained variation. The larger the determination of coefficient is, the more robust the model. As the linear regression model fits a line to minimize residuals, an ideal fit would result in zero residuals. When residuals are small, we have a *good fit* of the line. As residuals get larger, the less robust the model becomes; it produces predicted values that deviate a lot from the observed ones.

Let's see how *R*-squared is graphically explained. In general, when variables are not correlated, each one is expected to explain a different amount of variation of the dependent variable. Still, variables are rarely uncorrelated, and to some extent, they explain the same percentage of the dependent variable's variation. In other words, the variation they explain partially overlaps. For example, in Figure 6.3A, X_1 variable explains portion *a* (the shaded overlapping part) of *Y* variation through simple linear regression. In Figure 6.3B, two independent variables are used, each one explaining a different amount of

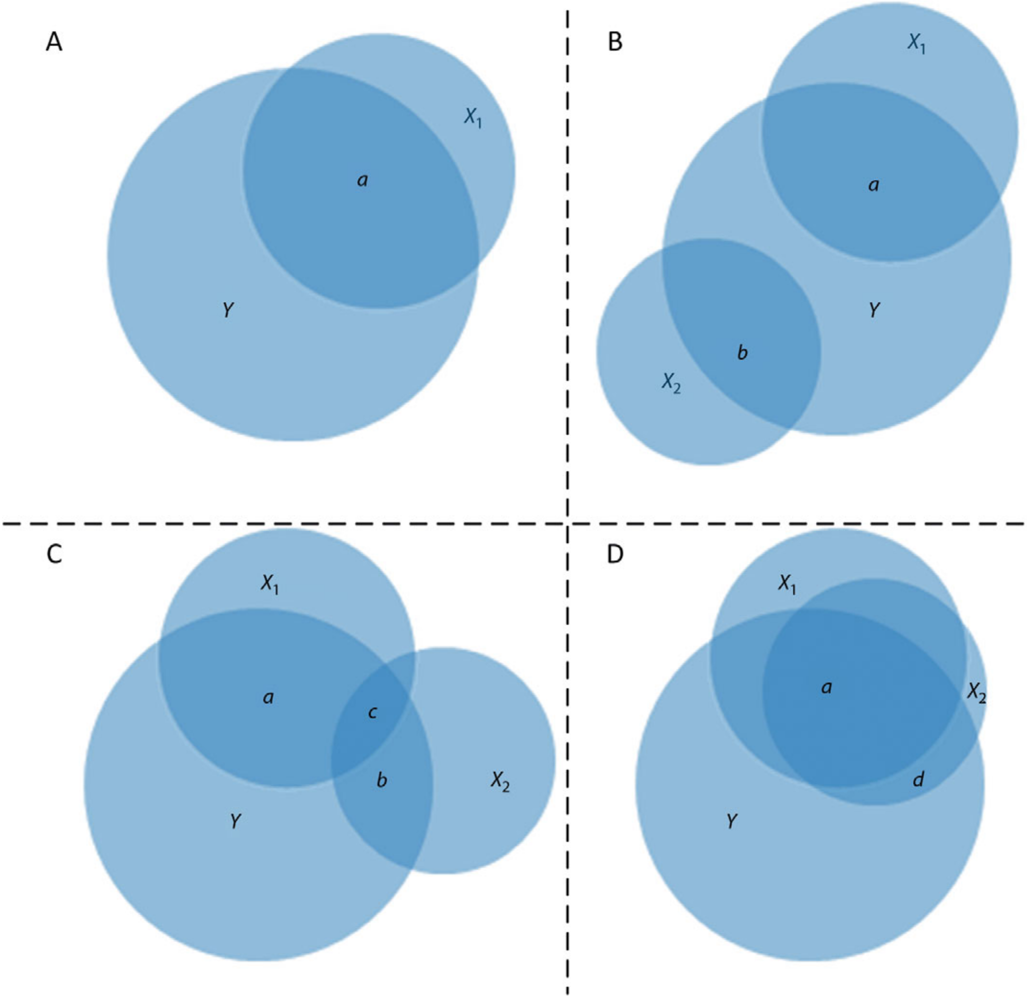


Figure 6.3 Variation explained in a graphical representation. (A) In cases of simple linear regression, variable X_1 explains portion *a* of independent's variable *Y* variation. (B) In MLR, an inclusion of a second variable leads to additional explanation *b* of variation. (C) These variables explain a common (overlapping) variation *c*. (D) X_1 and X_2 are highly correlated and explain almost the same proportion of variation.

variation. The total explained variation *R*-squared is $a + b$. There is no overlap between X_1 and X_2, as these two variables are uncorrelated. In Figure 6.3C, X_1 and X_2 are moderately correlated. There is an overlapping portion *c* of *Y* variation that is explained by both independent variables, so it should be included only once in the total variation explained. It is only the extra non-overlapping proportion $(b - c)$ of variation explained through X_2, which is of great importance. The final *R*-squared is $a + (b - c)$. In Figure 6.3D, X_1 and X_2 are strongly correlated and the extra variation explained, when including X_2 in the model, is portion *d*. The final *R*-squared is $a + d$, where *a* is the variation explained by X_1 and d is the non-overlapping variation between X_1 and X_2. If $X1$ explains 80% of the variation and by adding X_2 the explained variation increases to 82%. This typically means that X_2 assists in explaining an extra 2% of the explained variation and not that X_2 would explain 2% if X_1 was not included in the model. The more correlated the variables are, the less the extra variation explained when adding a new (correlated) variable. The less correlated the variables are, the more extra variation explained when adding a new (uncorrelated) variable.

Discussion and Practical Guidelines

A small *R*-squared is not necessarily a sign of an insufficient regression model. How we interpret *R*-squared depends on the scopes of the study and the research questions. For example, for identifying the relationships between dependent and independent variables, even a low *R*-squared can still be useful. Although there is not much of variation explained (small *R*-squared), the model might still have statistically significant independent variables (low *p*-values). Having statistically significant coefficients means that we cannot reject the null hypothesis that these coefficients are zero (do not affect the dependent variable), thus revealing valid relationships among variables. In other words, even with low variation explained, relationships might still exist, and as such, the model will be beneficial in identifying trends (i.e., increasing *X* will increase *Y*, according to coefficients). For this reason, *R*-squared should always be evaluated under the prism of the residual plot (see example in the following sections).

On the other hand, it would not be wise to use a model with low *R*-squared for predictive purposes, as the large unexplained variation will lead to large predictive intervals for the *Y* values and thus to a model with poor predictive quality. Prediction should be considered when *R*-squared is high. Furthermore, the regression line should not be used for predicting values far out of the range of initial *Y* and *X* values, as we cannot directly assume that relationships extracted inside a specific range stand well outside this range. This is called the danger of extrapolation (Peck et al. 2012).

R-squared provides a meassure of the strength of the relationships among dependent and independent variables. Still, it is not so helpful in cases of multiple regression because for every new variable added, *R*-squared increases. This reveals potential model overfitting, which is undesirable. To overcome this problem, we introduce the adjusted *R*-squared (see Section 6.3.3). Finally, *R*-squared does not provide a measure of how statistically significant the hypothetical relationship is. To do so, we use the F-test of overall significance (Section 6.3.6).

6.3.3 Adjusted *R*-Squared

Definition

Adjusted *R*-squared is the coefficient of determination, adjusted for the number of coefficients. The formula to calculate adjusted *R*-squared is (6.23):

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST} \tag{6.23}$$

where

SSE is the sum of the squared error (residuals; see Eq. 6.18)
SST is the sum of the squared total (see Eq. 6.19)
n is the number of observations
p is the number of parameters – regression coefficients (with intercept included)

Why Use

In MLR, for every new variable added, *R*-squared increases, revealing potential model overfitting. To account for this problem, we use adjusted *R*-squared, which does not necessarily increase with the addition of extra variables. In MLR, adjusted *R*-squared should be used instead of *R*-squared, which is mainly used in simple linear regression.

Interpretation

Adjusted *R*-squared reveals the additional variation explained if an extra independent variable is included in the model. If we add a new variable and *R*-squared adjusted increases, then this is a useful variable. If *R*-squared adjusted decreases, then this variable might not be useful for the model.

Discussion and Practical Guidelines

Both *R*-squared and adjusted *R*-squared are used to evaluate the percentage of total variation explained. Still, there is a significant difference between them. While *R*-squared assumes that every single variable explains the variation of *Y*, adjusted *R*-squared explains the percentage of variation of only those independent variables that have an impact on *Y*. *R*-squared tends to increase with each additional variable, leading us to the wrong conclusion that we create a better model. In fact, by adding more and more variables, we confront the problem of overfitting data. If we use the adjusted *R*-squared instead, we can determine which variables are useful and which are not by inspecting the adjusted *R*-squared values. We may consider keeping only those variables that increase the adjusted *R*-squared. The adjusted *R*-squared is a downward adjustment of *R*-squared and, as such, is always smaller from the *R*-squared value. Adjusted *R*-squared is also used to compare models with different number of predictors (see also predicted *R*-squared in Section 6.3.4).

6.3.4 Predicted *R*-Squared

Definition

Predicted *R*-squared is a statistic that measures the ability of a regression model to predict a set of newly presented data (Ogee et al. 2013). Predicted

R-squared is an iterative process and is calculated by removing each observation successively from the dataset and then estimating the regression equation again. The new model is used to calculate the observation removed, and the result is compared with the actual value.

Why Use

R-squared and adjusted *R*-squared explain how well the dependent variable is explained by the independent variables available. Still, they do not directly provide an estimation of the predictive quality of the model. When there is a big difference between *R*-squared and adjusted *R*-squared, it is a sign that the model contains more predictors than necessary, leading to low predictive quality. As such, the closer to 1 the ratio of adjusted *R*-squared to *R*-squared is, the better the model fits the existing data. A better way to assess the predictive quality of the model is by using the predicted *R*-squared.

Interpretation

Predicted R-squared reveals if overfitting of the original model exists. The higher the predicted *R*-squared is, the better the original model is. The lower the value is, the more unsuitable the model is, indicating overfitting.

Discussion and Practical Guidelines

Adjusted *R*-squared is used to test whether additional independent variables affect the explained variation of the dependent variable. On the other hand, predicted *R*-squared is more valuable than adjusted *R*-squared when it comes to evaluating the predictive capabilities of the model. Predicted *R*-squared is calculated using existing observations with known *Y* values that are not included in the initial model creation (out of sample). For example, suppose we get adjusted *R*-squared = 0.89 and predicted *R*-squared = 0.61. The high adjusted *R*-squared reveals that the original model fits existing data well, but when the model is fed with unseen data, its predictive quality decreases sharply to the value of 0.61 (predicted *R*-squared). A low predicted *R*-squared might be a sign of an overfitted model. A remedy would be to remove independent variables, add observations, or both. Predicted *R*-squared is not widely available in statistical software packages, but it can be calculated if one has basic scripting knowledge following the preceding steps/procedures.

6.3.5 Standard Error (Deviation) of Regression (or Standard Error of the Estimate)

Definition

The **standard error (deviation) of regression** for samples is the square root of the sum of squared errors divided by the degrees of freedom (6.24) (de Smith 2018 p. 507):

$$s_e = \sqrt{\frac{SSE}{DFE}} = \sqrt{\frac{SSE}{n-p}} \tag{6.24}$$

where

SSE is the sum of squared errors (errors; see Eq. 6.18)
DFE = $n - p$ is the degrees of freedom for error
n is the number of observations
p is the number of parameters (coefficients) of the model (including slope)

The quantity $\frac{SSE}{n-p}$ is also known as the mean-squared error (MSE). As a result, the estimate of the standard error s_e is the square root of the MSE (Lacey 1997).

In cases of simple linear regression, the standard error is calculated as (6.25):

$$s_e = \sqrt{\frac{SSE}{n-2}} \tag{6.25}$$

where $p = 2$, as two parameters are used, the slope and the intercept.

Why Use

To calculate the extent (average distance) of the observations' deviation from the regression line. The standard error of regression is also used to compute the confidence intervals for the regression line (de Smith 2018 p. 507).

Interpretation

The magnitude of the standard deviation reveals how close or far, on average, observations lie from the least squares line (average distance that a data point deviates from the fitted line). Approximately 95% of the observations should range within two standard errors from the regression line (Scibila 2017). In this respect, we can use the standard deviation of the regression as a rough estimate of the 95% prediction interval. The standard error of regression has the same units with the dependent variable. Consequently, it gives an estimation of the prediction's precision in the same units to the dependent variable. Lower values of standard errors are preferred, indicating smaller distances between the data points and the fitted values.

Discussion and Practical Guidelines

With the standard error of regression, we assess the precision of prediction measured in the units of *Y*, which is more straightforward compared to the *R*-squared – a percentage. By combining *R*-squared and standard regression error, we better evaluate the validity of our model. The standard error of regression can sometimes assist in defining how high *R*-squared should be in order to consider the model valid. For example, a not very high *R*-squared value with a small standard error of regression might be accepted according to the study and the research questions.

6.3.6 F-Test of the Overall Significance

Definition

The **F-test** statistic of the overall significance of the model, evaluates whether the coefficients in MLR are statistically significant (de Smith 2018 p. 508) (6.26):

$$F = \frac{SSR/DFM}{SSE/DFE} \tag{6.26}$$

where

SSR is the sum of squared regression (see Eq. 6.17)
SSE is the sum of squared errors (see Eq. 6.18)
$DFM = p - 1$ is the degrees of freedom for the model
p is the number of model parameters
$DFE = n - p$ is the degrees of freedom for error
n is the number of observations

The hypotheses for the F-test of the overall significance are as follows:

- **Null hypothesis:** coefficients' values are zero H_0: $b = 0$.
- **Alternative hypothesis:** coefficients' values are not zero H_1: $b \neq 0$.

The F-test is a test on the joint significance of all coefficients except the constant term and is also called joint F-statistic. F-test also produces a *p*-value, which is called F-significance. The F-value of a regression model is compared to the F-critical value resulting from the F-tables and the corresponding degrees of freedom. It is not essential to showcase how these values are extracted, but for those interested more explanation can be found in Peck et al. 2012 (p. 840).

Why Use

F-test is used to test if the regression fit is statistically significant, or else, if the regression has been successful at explaining a significant amount of the variation of Y (Rogerson 2001 p. 110).

Interpretation

In the context of regression analysis, the F-statistic value is not useful by itself. It is used for comparison with the F-critical value. As such, large or low values do not indicate any significance of the model. Instead, we should be directed to the F-significance value (*p*-value) to draw our conclusions. If the F-significance is small (e.g., less than 0.05), then we can reject the null hypothesis and accept that the regression model is statistically significant. In this case, we state that the observed trend is not due to random sampling of the population. If the F-significance is high (e.g., larger than 0.05), then we cannot reject the null hypothesis. In this case, we have to be cautious on how to interpret results, as we cannot definitely state that there is no relationship between *Y* and *X*. Possible interpretations are the following:

- There is not a linear relationship between *X* and *Y*. Still, there might exist a nonlinear relationship such as exponential or logarithmic. A scatter plot of *X* over *Y* is the first step to identify potential curvature.
- The *X* variable might explain a small portion of the variation of *Y*, but it is not sufficient enough to consider the model statistically significant by using only *X*. We might need to add some more variables to explain *Y*. By adding extra variables, we might discover a linear relationship that was not unraveled initially. Additional variables are likely to increase the percentage of variation explained as well.
- There might exist a linear trend, but if our sample size is small, then we are not able to detect any linearity.

Discussion and Practical Guidelines

In general, F-statistic is not very useful as it's quite common that the output is statistically significant (Anselin 2014). Other statistics such as the Wald test and the adjusted *R*-squared can be used to assess the overall performance of the model.

6.3.7 t-Statistic (Coefficients' Test)

Definition

The **t-statistic** in regression analysis is a statistic used to test if coefficient *b* is statistically significant. Coefficients are calculated as point estimates through the *n* data points available. In any point estimate, an indication of the accuracy is needed (Peck et al. 2012). The hypotheses to be tested are

- **Null hypothesis:** The coefficient value is zero H_0: $b = 0$.
- **Alternative hypothesis:** The coefficient value is not zero H_1: $b \neq 0$.

The t-statistic is calculated as (6.27):

$$t = \frac{b}{SE_b} \tag{6.27}$$

where SE_b is the standard error of the estimated coefficient *b*.

Why Use

To assess if the coefficients values are statistically significant and decide which independent variables to include in the model.

Interpretation

For a small *p*-value, we reject the null hypothesis that coefficient *b* is zero, and we accept its value as statistically significant. In this case, the independent variable *X* that the coefficient is assigned is important for the calculation of the dependent *Y*. For a large *p*-value, we cannot accept *b* as statistically significant, and we should consider removing the corresponding independent variable from the model.

Discussion and Practical Guidelines

It is important to highlight that each measurement is not error-free. The ability of each independent variable to contribute to increased explained variation is also relative to the measurement error. When we interpret the t-statistic, we should have in mind that a potential rejection of a predictor in our model might be just due to measurement error and not due to a nonexisting relationship. It is likely (but still not that common) that with another, more accurate dataset, this same variable turns out to be significant.

Before we evaluate the t-statistic, we should check the Koenker (BP) statistic (Koenker & Hallock 2001). When the Koenker (BP) statistic is statistically significant, heteroscedasticity exists, and the relationships of the model are not reliable (see Section 6.4). In this case, a robust t-statistic and robust probabilities (calculated automatically) should be used instead of the t-statistic *p*-value. Robust probabilities can be interpreted as the *p*-values, with values smaller than 0.01 for example, being statistically significant.

6.3.8 Wald Test (Coefficients' Test)

Definition

The **Wald test** is a statistical test that evaluates the significance of the coefficients.

The hypotheses for the Wald test of coefficients' significance are as follows:

- **Null hypothesis:** The coefficient value is zero H_0: $b = 0$.
- **Alternative hypothesis:** The coefficient value is not zero H_1: $b \neq 0$.

Interpretation

If the test rejects the null hypothesis (*p*-value smaller than the significance level), then we accept that the coefficient *b* is not zero, and thus, we can include the related variable in the model. If the null hypothesis is not rejected, then removing the respective variable will not substantially harm the fit of the model.

Discussion and Practical Guidelines

The Wald test may additionally be applied for testing the joint significance of several coefficients, and that is why it is called joint Wald test as well. In this case, it can be used as a test for the general performance of the model, or else as a test for model misspecification. If the *p*-value is small, it is an indication of robust overall model performance. The joint Wald test should be checked instead of the F-significance when the Koenker (BP) statistic is statistically significant.

6.3.9 Standardized Coefficients (Beta)

Definition

Standardized coefficients are the coefficients resulting when variables in the regression model (both dependent and independent) are converted to their z-scores before running the fitting model. The standardized regression coefficients – also called beta coefficients, beta weights or just beta – are expressed in standard deviation units, thus removing the measurement units of the variables. Coefficients that are not standardized are also called unstandardized coefficients.

An alternative way to compute the beta coefficients, is to use the regression coefficients b_i estimated when variables are not standardized, multiplied by the ratio of the standard deviation of the independent variable X_i, to the standard deviation of the dependent variable Y (6.28):

$$Standardized\ b_i = \ b_{i^*} \frac{Standard\ Deviation\ (X_i)}{Stdandard\ \ Deviation\ (Y)} \tag{6.28}$$

where

b_i is the unstandardized coefficient of X_i

Why Use

To decide which independent variable X is more important to model dependent Y. To compare the effect of each independent variable to the model, especially when the units or the scale among the independent variables is different. In fact, in case of unstandardized coefficients, we cannot state which independent variable is more important in determining the value of Y since the value of the regression coefficients depend on the units with which we measure X.

Interpretation

In cases of beta coefficients, a change in X_i by 1 standard deviation of X_i leads, on average, to beta coefficient (b_i) standard deviations (s_y) change, in the dependent variable Y or else: an average change in Y of *Standardized* $b_i \times s_y$.

Discussion and Practical Guidelines

Coefficients can be defined as the numbers by which the variables in the regression equation are multiplied. Unstandardized coefficients reveal the effect on the dependent variable of a one-unit change (increase or decrease depending sign) of an independent variable, if all other independent variables remain stable.

For example, consider the following model for estimating the monthly income of individuals

$$Income = 2{,}000 + 50(Years\ of\ education) + 10(Size\ of\ house)$$

This model suggests that for every additional year of education, holding constant the *Size of house*, income increases by 50 units. Still, as coefficients are estimated from variables measured in different units and scales, their importance in determining the dependent variable cannot be compared. In other words, we cannot infer that *Years of education* is more important because its coefficient value ($b_1 = 50$) is larger than the coefficient of *Size of house* ($b_2 = 10$). The significance level is sometimes used to compare coefficients but that is not what significant levels are meant to be used for (Linneman 2011). A better way to compare coefficients is to standardize them.

Suppose that for the model above, the beta coefficients are

$$Income = 0.32 + 0.18(Years\ of\ education) + 0.23(Size\ of\ house)$$

As *Years of education* increase by one standard deviation, income increases by 0.18 standard deviations. Having the largest effect on income is *Size of house*, as it has a larger beta value (0.23) compared to the beta of *Years of education* (0.18). By standardizing the coefficients, we may assess the importance of each variable better compared to the unstandardized. For example, it is now revealed that the *Size of house* is more important predictor for income, compared to *Years of education*, having a higher beta coefficient. By estimating beta coefficients, we calculate how important each statistically significant variable is to the model. The higher the beta value (ignoring the sign), the more critical the variable is, compared to the other variables of the model.

Standardized coefficients have been criticized, as they express standard deviations (thus removing the scale of the actual unit of measurement from each variable), and as such, they become less meaningful and are not easily interpreted, especially when distributions are skewed. Furthermore, comparisons across groups are difficult as the standardization is different for each group. In this respect, as beta coefficients most of the time are automatically calculated by statistical software and are readily available, they can be used as a way to assess which variable is more important. Still, it would be more reasonable to use unstandardized coefficients to interpret what is the effect of one unit change of X to Y.

To answer the question of which variable is more important in a regression model, strongly depends on the specific context in which any analysis is carried out. Statistical analysis provides measures to evaluate any model, but the researcher should consider the meaning of a one-unit change in an independent variable in relation to the dependent variable in a real-world context. From the policy perspective, some variables might not be feasible to significantly change. If, for example, we build a model where the *Number of cars* has a strong effect on *Pollution*, then applied policies can hardly interfere with *Number of cars*, as the number of cars is unlikely to decrease worldwide. It may be more rational to switch from petrol to another type of fuel that might result in a positive effect on mitigating pollution. From the cost perspective, one might also consider whether a change in one independent variable through applied policies is preferred to a change in another independent variable to obtain an equivalent change in the dependent variable. Depending on the problem studied, it might be more cost-effective to prefer a small change to a variable through specific policy actions instead of a larger change to another variable that is more rigid.

6.3.10 Residuals, Residual Plots and Standardized Residuals

Definition

The **residual** (as shown in Eq. 6.2) is the difference of the observed value with the predicted (fitted) value (Figure 6.1B).

Residual plots are a set of plots used to verify the accuracy of the regression model in relation to the residuals' errors and the regression's assumptions. The most commonly used residual plot is the scatter plot that plots the residuals *e* of a regression in the *y*-axis for each fitted (predicted) value $\hat{Y}$ in the *x*-axis ($\hat{Y}$, e) (Figure 6.4A and B). It is used to test if the residuals' errors have a constant variance (homoscedasticity). Additional residual plot include the normal probability plot of residuals that determine whether they are normally distributed and the histogram of residuals to test if the residuals are skewed or if outliers exist.

Standardized residuals are residuals divided by their estimated standard deviation. The standardized residual for the *i*-th observation is (6.29):

$$standardized\ residual_i = \frac{residual(i)}{estimated\ standard\ deviation\ of\ residual\ (i)} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}} \tag{6.29}$$

where

MSE is the mean squared error
h_{ii} is the leverage value for observation *i* (see more in Section 6.3.11)
e_i is the residual of the *i*-th observation

Standardized residuals have a mean of zero and a standard deviation of 1.

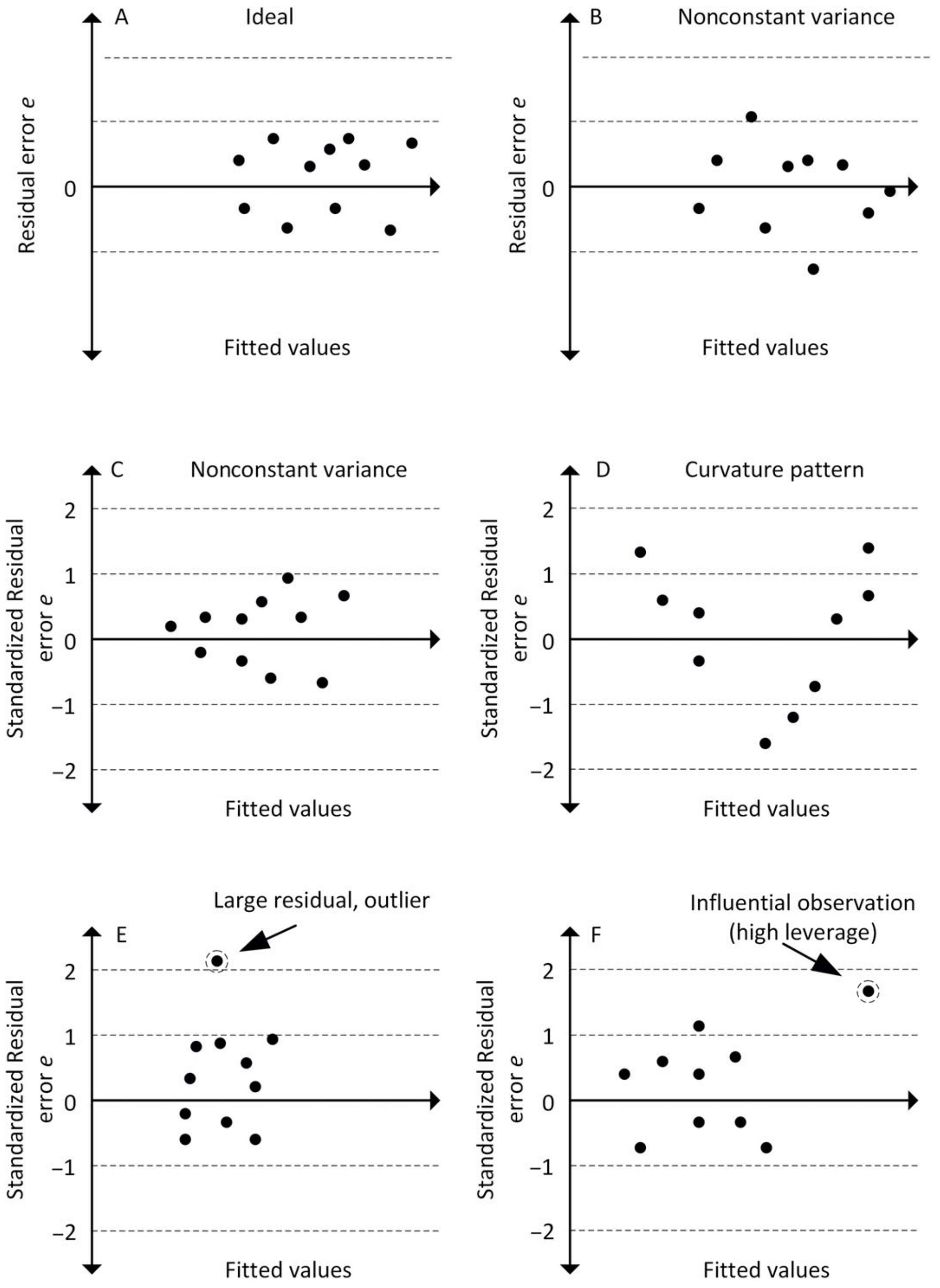


Figure 6.4 (A) Residual plot with randomly dispersed residuals (no pattern). (B) Residual plot with nonconstant variance. (C) Standardized residual plot with nonconstant variance. (D) Standardized residual plot. A curvature pattern in residuals is detected, indicating that a better fit would be a nonlinear model. (E) Standardized residual plot with a large residual (outlier). (F) Standardized residual plot with influential observation.

The **standardized residual plot** is the plot of the fitted value against the standardized residuals (*y*-axis depicts the number of standard deviations, and the *x*-axis the fitted values; see Figure 6.4C and D).

Why Use

The residual plot and standardized residual plot are ways to assess the quality of the regression model and identify if the normality or the homoscedasticity assumptions are violated. A standardized residual plot offers an easy way to locate large discrepancies of the dataset. For example, data points that lie further away in the *y*-axis direction create large residuals that may be potential outliers (see Figure 6.5D). Similarly, data points lying further away from the average fitted values are potential influential points (see Section 6.3.11 and Figure 6.5D).

Interpretation

When residuals are positive, there is underprediction. Underprediction means that the fitted (predicted) value is lower than the observed. On the other hand, overprediction means that the fitted value is higher than the observed. In this case, residuals are negative. If under- or overpredictions occur systematically, then there is bias in the model. To check if bias exists, we use the residual plot.

Ideally, the residual plot should exhibit no particular pattern, and data points should be randomly distributed, have constant variance and be as close to the *x*-axis as possible (see Figure 6.4A). These conditions satisfy the third (normal distribution of errors) and the fourth (homoscedasticity) assumptions of linear regression (see Section 6.4). If clusters, outliers, or trends exist in the data points pattern, it is an indication of one or more assumptions' violation (see Figure 6.4B–E). If assumptions are severely violated, then the model is unreliable and cannot be used although some statistics may be statistically significant.

By standardizing residuals, we calculate how many standard deviations a residual lies from the mean (mean, in our case, is 0; see Chapter 2 for standardization methods). Standardized residual plots are interpreted like the nonstandardized plots in regards to their shape and variance. Additionally, observations that lie vertically more than ±2 standard deviations from the *x*-axis (in the direction of *y*-axis) might be regarded as outliers (see Figure 6.4E). If we locate an outlier, we should check if there is an error in the dataset or if it is a value revealing unusual behavior.

Discussion and Practical Guidelines

Residuals analysis deals with analyzing the residuals (or the standardized residuals) of a regression model by plotting them against the fitted values. Residual plots, standardized residual plots, scatter plots and other graphs are as important as the outputs of the statistical tests and metrics reported in regression analysis. These graphs provide a graphical representation of inspecting data and regression results concurrently, thus increasing our perception about the model performance. For example, inspecting Figure 6.4, an ideal pattern of residuals would look similar to the one in subplot A. No trends, clusters or outliers are observed. The observed pattern resembles a random point pattern. Additionally, the closer to the *x*-axis the points lie, the less magnitude of errors *e*. Subplot B reveals some trend at a diagonal direction (upper left to bottom right). This arrangement indicates a nonconstant variance of errors *e*, violating the homoscedasticity assumption of regression. Subplots C to E depict the standardized residuals. Subplot C indicates an increasing trend in errors. A linear pattern might still be valid, but we may also test a weighted least squares method instead of OLS. By this method, we assign more weights to the observations with high variability and less to those observations of low variability. An opposite pattern (reflection) would be treated similarly. We can also transform data and then apply OLS again. The pattern in subplot D indicates a curvature pattern. Either we use a nonlinear model or we transform data.

6.3.11 Influential Points: Outliers and High-Leverage Observations

Definition

An **influential point** is any data point that substantially influences the geometry of the regression line (Figure 6.5D).

Outliers are data points for which their response *Y* lie far away from the other response values (see Figure 6.5A). In cases of regression analysis, we may use the standardized residuals to trace potential outliers, as those data points exceeding more than 2 (sometimes more than 2.5) standard deviations away from the mean (vertical distance of a point from the *x*-axis; see Figure 6.4E).

Leverage of a data point is a measure of the effect of this point, in respect to its *X* values, on the regression predictions (see Figure 6.5B and D). It is the value of the *i*-th diagonal element of the hat matrix *H* (6.30), where

$$H = X(X^T X)^{-1} X^T \tag{6.30}$$

The diagonal elements satisfy

$$0 \le h_{ii} \le 1 \tag{6.31}$$

$$\sum\nolimits_{i=1}^{n} h_{ii} = p \tag{6.32}$$

where

h_{ii} is the leverage value for observation *i*
p is the number of parameters (coefficients) in the regression model and
n is the number of observations.

Why Use

Outliers are used to identify extreme *Y* values. Leverage is used to identify outliers with regard to their *X* values. An influential point should be identified and potentially removed, as it significantly distorts the geometry of the regression line, leading to inaccurate models.

Interpretation

Those data points that exceed vertically more than 2 (or 2.5) standard deviations away from the *x*-axis of the standardized residual plot can be considered outliers (see Figure 6.4E). Data points with extreme values in the *x*-axis direction (or else those lying far away from the center of the input data space) have greater leverage than those close to the center of the input space (see Figure 6.5B).

As a rule of thumb, leverage is large when it exceeds (6.33):

$$2(p + 1)/n \tag{6.33}$$

where *p* is the number of parameters (intercept including) and *n* is the number of observations.

Discussion and Practical Guidelines

An influential point might be both an outlier point (at the *y*-axis direction of the residual plot or far away from the regression line) and a point with high leverage (extreme value at the *x*-axis direction; see Figures 6.4F and 6.5D). It is quite common that influential points are those with high leverage. The typical meaning of leverage is the exertion of force using a lever. In cases of simple linear regression, a high-leverage point tends to rotate the regression line as a lever unless it lies very close to the regression line (Figure 6.5B). An outlier lying around the mean of the *x*-axis values does not create large changes in the geometry of the regression line (see Figure 6.5C).

Similarly, having a point with high leverage that is close to the regression line infers small changes in the slope. The more it deviates from the regression line, the sharper the changes in the regression are. For example, a point lying far right will produce a lower slope if it also lies away from the regression line and, therefore, will be an influential point (see Figure 6.5D). A point with high leverage that lies on the regression line infers no change in the geometry of the line (see Figure 6.5B). In general, we can remove influential objects and test if the model is improved. If the model is considerably improved, then we have to consider dropping these observations. Another method we may apply to control for outliers is the robust regression.

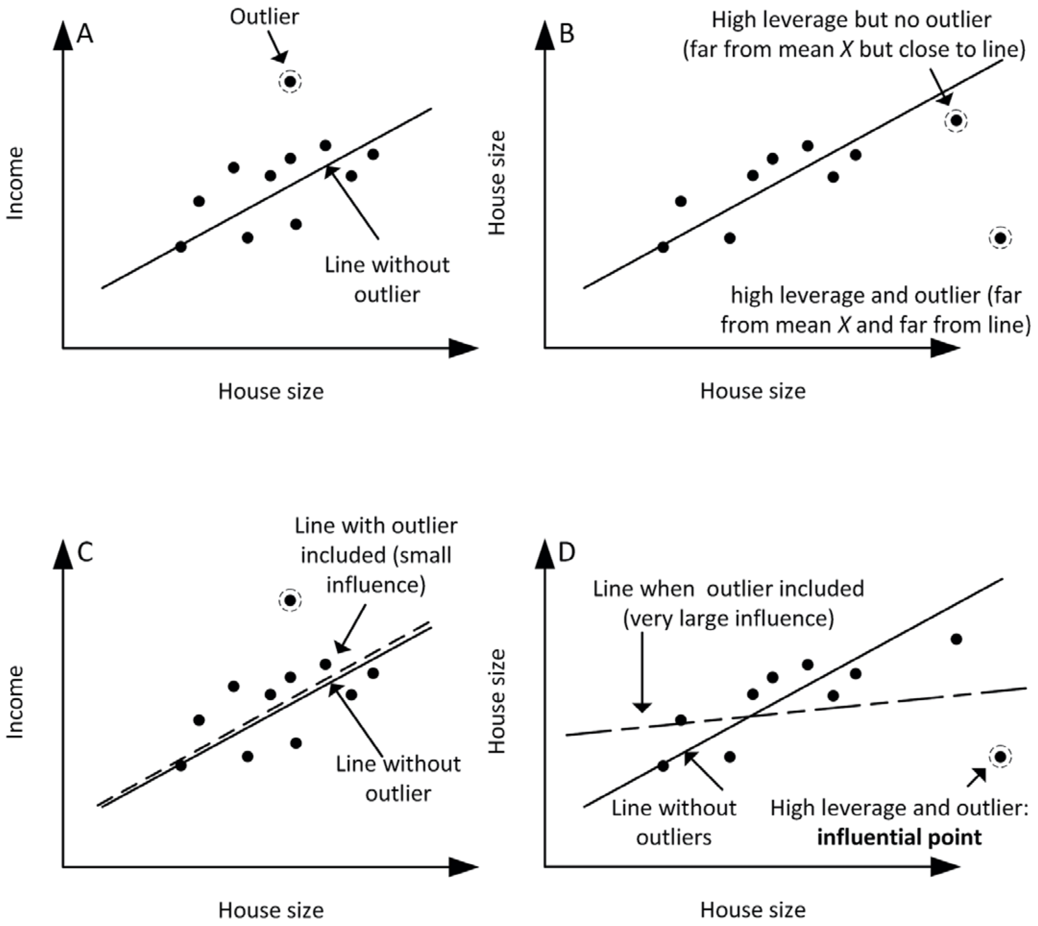

A
Outlier
Income
Line without outlier
House size
B
High leverage but no outlier (far from mean X but close to line)
House size
high leverage and outlier (far from mean X and far from line)
House size
C
Line with outlier included (small influence)
Income
Line without outlier
House size
D
Line when outlier included (very large influence)
House size
Line without outliers
High leverage and outlier: **influential point**
House size

Figure 6.5 Outliers, high-leverage points and influential points. (A) A point far away from the regression line might be an outlier. (B) A point that can change the slope of the regression line is a leverage point. (C) An outlier with small influence on the regression line. (D) A high-leverage point that influences the slope of the regression line.

6.4 Multiple Linear Regression Assumptions: Diagnose and Fix

MLR builds a probabilistic model based on five assumptions (tip to remember: LINE-M):

1. **Linearity**. The relationships between Y and X_i are linear. Regression analysis is based on Pearson's r correlation coefficient, which is appropriate for linear relationships. As such, outliers and influential observations should be considered for exclusion, prior to any other analysis.

   The effect if the linearity assumption is violated (see Table 6.1)*:*

- A large portion of dependent *Y* variation will not be explained (de Vaus 2002 p. 343)
- Important relationships will not be detected

How to diagnose if linearity exists: Linearity assumption can be tested using

- Scatter plots
- Residual plots

What to do if the linearity assumption is violated (fixing nonlinearity):

- Apply transformations
- Use another type of regression not sensitive to linear relationships (e.g., logistic)

2. **Independence – No spatial autocorrelation**. The residuals are independent of each other, meaning that an error *e* in a data point *X* does not affect by any means another error *e* for some different value of the same variable *X*. In other words, errors should be randomly scattered.

The effect if the independence assumption is violated (see Table 6.1)*:*

- Inefficient coefficient estimates
- Biased standard errors
- Unreliable hypothesis testing
- Unreliable predictions

How to diagnose if independence exists:

- For non–time series data or nonspatial data, residual plots may be used to inspect if errors are randomly scattered.
- For spatial data, various spatial autocorrelation indices can be used, including Moran's I index, Geary's C index, General G-Statistic (see Chapter 4).

What to do if independence is violated (fixing error non-independence):

- For non–time series data or nonspatial data, apply transformations.
- For spatial data, if residuals are spatially autocorrelated, then we may improve the model by adding or removing variables. We can also use a spatial regression model that controls for spatial autocorrelation – e.g., spatial filtering with eigenvectors (Thayn & Simanis 2013). Geographically weighted regression (GWR; see Section 6.8) can be also tested.

3. **Normality.** The distribution of residual errors *e* at any particular *X* value is normal, having zero mean value. This means that for fixed values of X_i, *Y* has a normal distribution as well.

The effect if the normality assumption is violated (see Table 6.1)*:*

- Inefficient coefficient estimates
- Biased standard errors
- Unreliable hypothesis testing
- Unreliable predictions

How to diagnose if normality exists: Assumption can be checked by using

- A histogram and a fitted normal curve of residuals.
- Q-Q plot of standardized residuals.
- Standardized predicted values plotted against standardized residual values. No pattern reveals normality.
- Histogram of standardized residuals. For normality, residuals should approximate a normal distribution.
- Jarque–Berra test. If the value of the test is statistically significant (*p*-value smaller than the significance level), then there is an indication for non-normality. The hypotheses for the Jarque–Berra test are

  Null hypothesis: normal distribution of regression errors

  Alternative hypothesis: non-normal distribution of regression errors (when *p*-value is smaller than the significance level, then we reject the null hypothesis).

What to do if the normality assumption is violated (fixing non-normality):

- Apply transformations

4. **Equality – Homoscedasticity.** Residuals (errors *e*) have a zero mean, equal variance and constant standard deviation at any particular value of *X*. This is also called homoscedasticity and exists when the variance of residuals does not change (increase or decrease) with the fitted values of the dependent variable. Heteroscedasticity, on the other hand, exists when the variance is not constant.

The effect if the homoscedasticity assumption is violated (see Table 6.1)*:*

- Inefficient coefficient estimates
- Biased standard errors
- Unreliable hypothesis testing
- Unreliable predictions

How to diagnose if heteroscedasticity exists: Assumption can be checked through

- Residual plot inspection. It is a straightforward way to diagnose if heteroscedasticity exists.
- Statistical tests such as the White test (White 1980), Breusch–Pagan test (Breusch & Pagan 1979) and Koenker (KB) test. If the value of the test is statistically significant (*p*-value smaller than the significance level), then we reject the null hypothesis, and there is an indication for heteroscedasticity. The hypotheses for the aforementioned tests for diagnosing heteroscedasticity are:

  Null hypothesis: constant variance of regression error (homoscedasticity)

Alternative hypothesis: nonconstant variance of regression error (heteroscedasticity)

- If the difference between the Breusch–Pagan test and the Koenker (KB) test values (not the *p*-values) is large, it is an indication of potential non-normality of the error, as, in the case of normality, the values should be similar. For the implementation of these and additional diagnostic tests, consider LeSage et al. (1998).

What to do if heteroscedasticity exists (fixing non-homoscedasticity):

- Heteroscedasticity often exists due to the skewness of one or more variables. Apply transformations to fix normality, and heteroscedasticity will be fixed to some extent. Often a log transformation is appropriate.
- Use robust standard errors, also called Huber–White sandwich estimators. It is a regression with robust standard errors that estimates the standard errors using the Huber–White sandwich estimators (and is preferred over weighted least squares regression). Regression with robust standard errors is one of the available approaches to perform robust regression. In most statistical software, robust standard errors are automatically calculated. Although coefficients remain the same (with non-robust standard errors), *p*-values change as standard errors change. In this respect, we should consider the new *p*-values, as some variables might not be statistically significant anymore.
- Apply a weighted least squares regression (if we know the form of heteroscedasticity).
- Use quantile regression.
- Consider using geographically weighted regression in cases of spatial data.

Although heteroscedasticity is undesirable for regression analysis leading to unreliable models, it reveals interesting patterns for the data. Heteroscedasticity usually occurs when subpopulations exhibit substantial differences in their values. For example, suppose we study *Income* versus *Cost of house*. To some extent, it is expected that people with lower incomes will buy houses with relatively less variability in their price, as their budgets are more rigid. On the other hand, for people with considerably higher income, the price of a house may vary a lot depending on extra criteria (extra variables to the model). To exaggerate a little, a preference (variable) might be the size of the indoor pool; that would never trouble the low-income group. A regression fit line based only on income to model the house cost would most likely yield a residual plot similar to the one in Figure 6.4C. Although it might have a good fit for the low-income values, as income increases, residual errors are expected to increase as well, simply because extra variables should be included to reflect reality better.

Residuals' heteroscedasticity is a typical problem of model misspecification. In practice, we should eliminate heteroscedasticity by adding variables, using robust standard errors or applying more sophisticated methods, such as geographical regression. Still, we have to point out that in a geographical context, heterogeneity reveals valuable information for the data. The variation of a phenomenon from place to place indicates that there are underlying processes at play, and geographical analysis should trace and explain them – not hiding them just because a statistical model should work. Instead of using robust standard errors, which is another mathematical operation applied on data (by weighting them with weights not easily interpreted), it is worth considering clustering data to homogeneous groups or using local regression models. As a general advice, it is better to select methods that are easier to interpret when dealing with geographical analysis instead of using statistical tests that are not easily comprehended in relation to the distortion they infer to the dataset.

5. **Multicollinearity** absence. Variables X_i should be independent of each other, exhibiting no multicollinearity (see Section 6.5).

The preceding assumptions reveal that a model should be appropriately designed to avoid errors. Errors occurring due to poor model design are called misspecification errors, leading to **model misspecification** and to a weak unreliable model. Misspecification may arise when

- Proper variables are omitted
- Wrong variables are included (irrelevant variables' presence)
- Data from included proper variables are not well selected (samples not appropriate or some measurements are wrong)
- Considering a relationship linear when it is not
- Violating the assumptions of linear regression (LINE-M)
- Independent variable is a function of the dependent (e.g., using population as the dependent variable and population density as one of the independent variables)

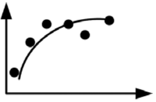
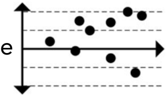
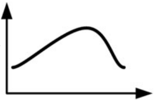
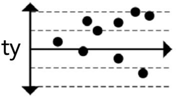
Table 6.1 presents a list with the assumptions violated, their definition, the effects of these violations, the diagnostic tests and the remedies.

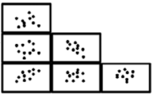
6.5 Multicollinearity

Definition

Multicollinearity exists among two or more variables X_i when they are highly or moderately correlated. No multicollinearity exists when their correlation is absent or very small.

Table 6.1 Assumptions violations, diagnostics and remedies.

Violations	Definition	Effect	Diagnose	Fixing violations
Nonlinearity 	The relationships between Y (dependent) and X_i (independent variable) are not linear	• A large portion of the variation of dependent Y will not be explained • Significant relationships remain undetected	• Scatter plots • Residual plots	• Apply transformations • Other types of regression, not sensitive to linear relationships, may be used (e.g., logistic) • Outliers and influential data should be removed
Non independence 	Residuals are not independent of one other	• Inefficient coefficient estimates • Biased standard errors • Unreliable hypothesis testing • Unreliable predictions	• Residual plots • Standardized residual plots • For spatial data, various spatial autocorrelation indices such as Moran's I index, Geary's C index, General G-Statistic	• Apply transformations • Improve the model by adding or removing variables • Use a spatial regression model that controls for spatial autocorrelation – e.g., spatial filtering with eigenvectors
Non-normality 	The distribution of errors e at any particular X value does not follow the normal distribution	• Inefficient coefficient estimates • Biased standard errors • Unreliable hypothesis testing • Unreliable predictions	• Histogram and a fitted normal curve • Q-Q plot • Jarque-Bera test • Standardized predicted values plotted against standardized residual values • Histogram of standardized residuals	• Apply transformations
Heteroscedasticity 	Nonconstant errors variance	• Inefficient coefficient	• Residual plots • White test	• Apply transformations

		estimates Biased standard errors Unreliable hypothesis testing • Unreliable predictions	• Breusch–Pagan test • Koenker (KB) test	• Weighted least squares regression • Robust standard errors (Huber-White sandwich estimators) • Use quantile regression • Use geographically weighted regression when dealing with spatial data
Multicollinearity 	When variables X_i in a dataset have high correlations among them	• Sensitive coefficient estimates • Coefficients are inflated • Not reliable confidence intervals and predictions • Insignificant independent variables might appear as significant	• Bivariate or pairwise correlation • Diagnostic statistics such as singular value decomposition, condition index (CI) • Variance decomposition proportions, variance inflator factor (VIF)	• Remove one of the variables that are highly correlated • Reduce the total number of variables by using principal component analysis • Use stepwise regression or exploratory regression
Model misspecification/ Overall performance low	Misspecification arises when the model is not well designed	• Not accurate model	Adjusted *R*-squared F-statisticJoint Wald's test	• Add omitted variables • Remove redundant variables • Apply transformations • Change model

Why Use

The absence of multicollinearity among independent variables is one of the five assumptions that should not be violated in order for the regression model to be reliable. If multicollinearity exists, then (Rogerson 2001 p. 126, de Vaus 2002 p. 343):

- Estimates of the coefficients are sensitive to individual observations, meaning that if we add or delete some observations, coefficients' values may change significantly.
- The variance of the coefficients estimates is inflated (increases).
- Significance levels, confidence intervals and prediction intervals are not reliable and are also wider.
- Insignificant independent variables might appear to be significant due to the large variability in coefficients.

Eliminating or reducing the multicollinearity in a dataset yields more accurate results, and more reliable (and less wide) confidence intervals for the coefficients.

Interpretation

Diagnostic statistics for detecting multicollinearity include the variance inflator factor (VIF), the condition index (CI) and the variance decomposition proportions (Belsley et al. 1980). Table 6.2 presents the values that multicollinearity is evident based on these diagnostics.

Discussion and Practical Guidelines

- **Variance inflation factor (VIF)** is a measure that estimates how much the variance of a coefficient is inflated due to multicollinearity existence (Rawlings et al. 1998). VIF values between 4–10 indicate increased multicollinearity, while high values VIF (>10) are a sign of severe collinearity (Table 6.2). For example, if VIF for an independent variable is 7.2, it means that the variance of the estimated coefficient is inflated by a factor of 7.2 as this independent variable is highly correlated with at least one of the other independent variables in the model.
- **Condition index (CI) and variance decomposition proportions** are examined based on the following conditions (Belsley et al. 1980):
  1. A large condition index associated with

Table 6.2 Multicollinearity diagnostics.

VIF	Collinearity	Condition index (CI)	Collinearity	Variance decomposition	Collinearity
1–4	No	$5 < \text{CI} < 30$	Weak	<0.5	Weak
4–10	Further investigation needed	$30 < \text{CI} < 100$	Moderate	>0.5	Severe
>10	Severe	$\text{CI} > 100$	Severe		

2. A large variance decomposition proportion for two or more covariates

   If both conditions are true, then multicollinearity might be present in the dataset. The variance decomposition index is checked when the condition index is large. Put simply, when the condition index is larger than 30 (condition 1), then there may be multicollinearity issues. In this case, we inspect the variance decomposition index. If a large CI is associated with two or more variables with a large variance decomposition index (over 0.5) (condition 2), then it is a sign of multicollinearity in the corresponding variables.

- **Bivariate or pairwise correlation** between the independent variables. Very high correlations (larger than 0.90) will produce collinearity problems. Correlation analysis has been presented in Section 2.3.

In case we detect multicollinearity, we may consider (see Table 6.1):

- Removing one of the highly correlated variables. Deciding which variable to delete strongly depends on the problem at hand. We should keep the variable that is conceptually more evidently linked to the dependent variable. In this respect, the choice should be based primarily on the assumed underlying processes and less on the magnitude of the multicollinearity metric.
- Reducing the total number of variables by using principal component analysis or by creating a composite index (variables combined in a new single variable).
- Applying exploratory regression, stepwise regression or hierarchical regression, which allows for controlling the order of variables entry.

6.6 Worked Example: Simple and Multiple Linear Regression

Suppose we want to model Income (dependent) to Medical expenses (independent) for a set of 64 postcodes (see Box 6.1). Both variables are measured in euros, and values refer to the average income and the average medical expenses of the people living in each postcode. The results of the OLS simple linear regression are:

Metrics and Statistics for Simple Linear Regression

Box 6.1 Matlab. You can find data, Matalab commands and related files to reproduce the following graphs and results in I:\BookLabs\Lab6\Matlab\

```
Run SimpleOLS.m
```

```
****************************RESULTS****************************
Linear regression model:

    Income ~ 1 + MedExpenses
Estimated Coefficients:
                      Estimate       SE       tStat        pValue
                      ________     ______     _____     __________

      (Intercept)     7446.8       365.84     20.355     3.823e-29
      MedExpenses     46.569       3.0092     15.476    5.7356e-23
Number of observations: 64, Error degrees of freedom: 62
Standard Error: Root Mean Squared Error: 897
Tip:    Calculated   based   on   eq.6.25   as:   Square   Root
(SumSqResidual/(n-2))=>(4.9866e+07/(64-2))^-0.5=896.82    {(n-2)
is  the Degrees  of  freedom  (DF)}
R-squared: 0.794, Adjusted R-Squared 0.791
F-statistic vs. constant model: 239, p-value = 5.74e-23
Linear regression model:
      Income = 7446.8 + 46.569*MedExpenses
ANOVA Results
               SumSq          DF      MeanSq         F         pValue
             __________       __    __________     _____     __________

 Total        2.4249e+08      63    3.8491e+06
 Model       1.9263e+08       1     1.9263e+08     239.5   5.7356e-23
Residual     4.9866e+07       62    8.0429e+05

 ***************************************************************
```

Interpreting Results

The adjusted *R*-squared is high, indicating goodness of fit (0.791) and that 79% of Income’s variation is explained by MedExpenses variable (Figure 6.6). As the F-significance is less than 0.05 (F-statistic *p*-value = 5.74e-23), we reject the null hypothesis and accept that the regression model is statistically significant, and the identified trend is a real effect, not due to random sampling of the population. In addition, MedExpenses coefficient’s *p*-value, calculated through t-statistic, is 5.7356e-23 (less than 0.05). As such, we accept the coefficient value as statistically significant for the model. In other words, MedExpenses independent variable is statistically significant and can be included in the model as a good predictor of Income. The regression of income on medical expenses is Income = 7446.8 + 46.569*MedExpenses (Figure 6.6). It practically means that for a one-unit increase in medical expenses (i.e., 1 euro), income increases by *b* = 46.569 euros. The standard error of the regression is 897 euros, expressing the typical distance that the data points fall from the

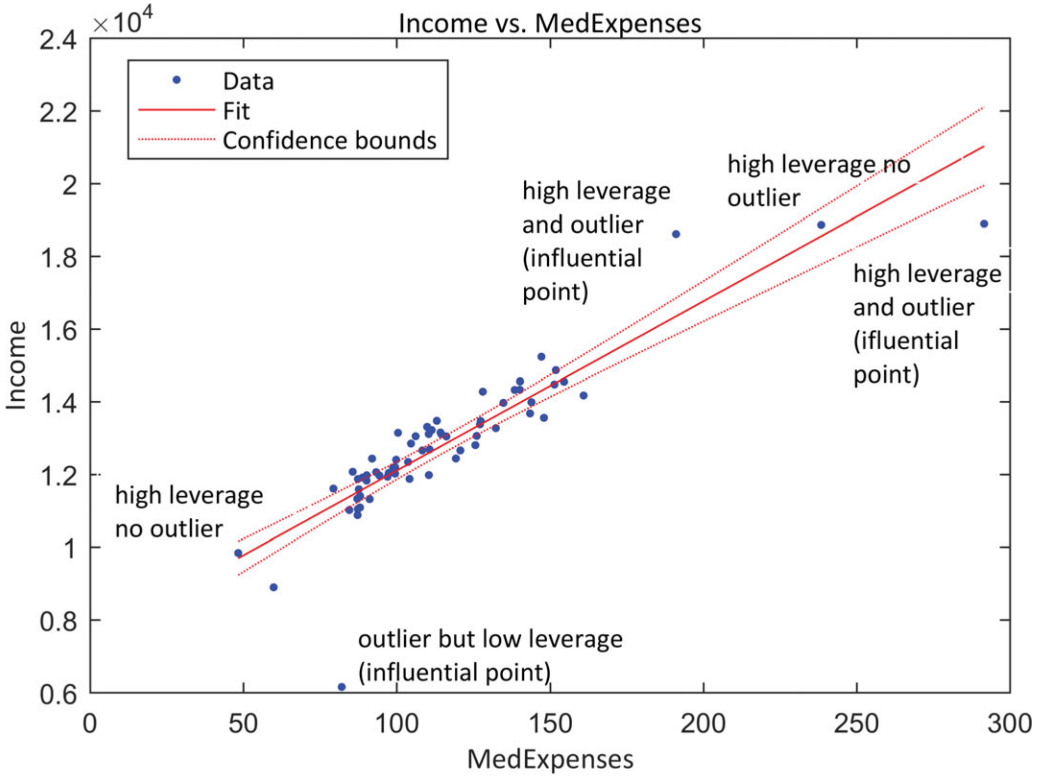


Figure 6.6 Plot observations, regression line and confidence interval for mean (95%). Leverage points and outliers are also mapped.

regression line on average. It also expresses how precise the model's predictions are using the units of the dependent variable. As the income values in the dataset range between 10,000 and 22,000 (Figure 6.6), we consider this error relatively low (although it could be lower; potential outliers might have a negative effect). We should also test (later) our model for homoscedasticity, errors normality and influential points.

Standardized Residual Plots

The standardized residual plot is depicted in Figure 6.7.

Interpretation

The residual plot (Figure 6.7) shows constant variance in errors indicating that there is not a serious heteroscedasticity issue. High leverage points (points with leverage value larger than 0.0625; see Eq. 6.33), along with outliers and influential points, are depicted in Figures 6.6 and 6.7. We should eliminate those points one by one and run the model again. If the model is improved, we may consider dropping these observations and keeping the new model.

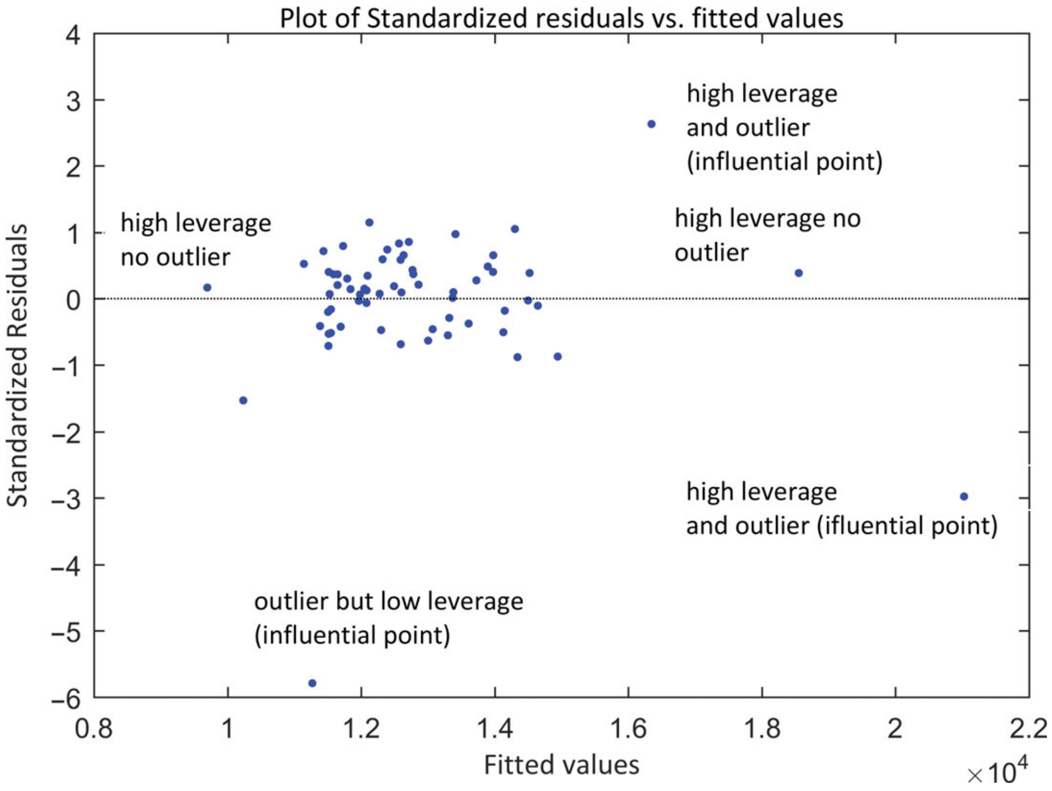


Figure 6.7 Standardized residuals vs. fitted values along with influential points. The lower-left point has low leverage because it is relatively close to the mean fitted value. Still, it is an outlier as it lies vertically from the *x*-axis more than 2.5 standard deviations. The lower-right point, on the other hand, is both outlier and high leverage point, as it is away from the mean fitted value and lies vertically from the *x*-axis more than 2.5 standard deviations. The upper-right point is for the same reasons both outlier and high leverage. Finally, the upper-left point is of high leverage but not an outlier.

Multiple Linear Regression

Continuing on the same example, suppose we want to model `Income` (dependent variable) as a function of the following independent variables (see Box 6.2):

- `Sec`: percentage of people obtained secondary education
- `Unv`: percentage of people that graduated from university
- `Med`: medical expenses per month in euros
- `Ins`: money spent on monthly insurance in euros
- `Ren`: monthly rent in euros

Multicollinearity

Before we apply regression modeling, we test if multicollinearity exists using the Belsley collinearity diagnostics.

Box 6.2 Matlab. You can find data, Matalab commands and related files to reproduce the following graphs and results in I:\BookLabs\Lab6\Matlab\

```
Run MLR.m
```

```
****************************RESULTS******************************

                      |     Variance  Decomposition            |
sValue     condIdx      Sec       Unv       Med       Ins        Ren
-----------------------------------------------------------------
2.2000       1        0.0013    0.0029    0.0004    0.0003    0.0014
0.3118     7.0561    0.0186    0.4484    0.0144    0.0101    0.0253
0.1886    11.6618    0.4080    0.2966    0.0666    0.0078    0.1281
0.1510    14.5684    0.3764    0.1439    0.0174    0.0232    0.8339
0.0677    32.4967    0.1956    0.1083    0.9012    0.9586    0.0113
*****************************************************************
```

Results show that only the last row has condition index larger than the default tolerance value 30. Inspecting this row, the third (Med) and fourth (Ins) independent variables have variance-decomposition proportions exceeding the default tolerance 0.5, suggesting that these variables exhibit multicollinearity.

The pairwise correlation for all variables and the related correlation matrix are as follows (see Figure 6.8):

```
*****************************RESULTS*****************************

  R =      Sec       Unv       Med       Ins       Ren
  Sec     1.0000    0.7630    0.3655    0.3560    0.4371
  Unv     0.7630    1.0000    0.3281    0.2563    0.2609
  Med     0.3655    0.3281    1.0000    0.9372    0.7151
  Ins     0.3560    0.2563    0.9372    1.0000    0.7399
  Ren     0.4371    0.2609    0.7151    0.7399    1.0000
  Pvalue =
      1.0000    0.0000    0.0030    0.0039    0.0003
      0.0000    1.0000    0.0081    0.0409    0.0373
      0.0030    0.0081    1.0000    0.0000    0.0000
      0.0039    0.0409    0.0000    1.0000    0.0000
      0.0003    0.0373    0.0000    0.0000    1.0000
*****************************************************************
```

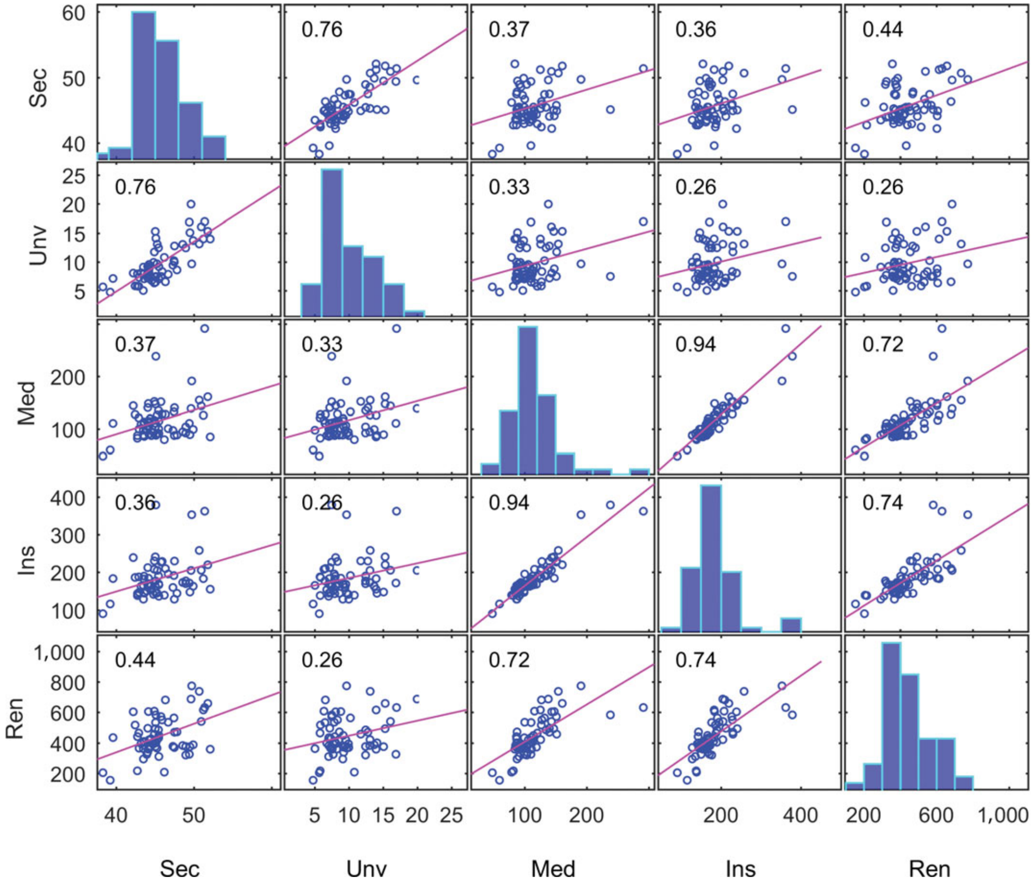


Figure 6.8 Correlation matrix.

Although correlations around 0.70 exist, we consider variables exhibiting serious collinearity issues as those exceeding correlation of 0.90. `Med` and `Ins` have a high pairwise correlation (0.94), indicating multicollinearity. We also calculate the variance inflation factor.

```
******************************RESULTS******************************

     VIF =
         2.8929

         2.6859

         9.1455

         9.4201

         2.4803

*******************************************************************
```

A high VIF value (close to 10) for the third and fourth variable (Med, Ins) is identified, a sign of multicollinearity existence. VIF shows how much the variance of the coefficient of an independent variable is inflated, but it does not point directly to which variable it exhibits multicollinearity. We should turn to the pairwise correlation and the variance decomposition index calculated earlier, to conclude that Med and Ins exhibit multicollinearity. We should continue by running the MLR but without including Ins.

```
******************************RESULTS***************************
Linear regression model:
     Income ~ 1 + Sec + Unv + Med + Ren
Estimated Coefficients:
                  Estimate         SE        tStat          pValue
                  ________       ______     ________      __________

   (Intercept)     9034.2        2299.7       3.9284      0.00022677
   Sec             -56.92        60.785     -0.93642         0.35287
   Unv              34.14        50.989      0.66957         0.50575
   Med             38.072        4.2493       8.9594      1.3349e-12
   Ren             3.7421        1.2923       2.8958       0.0052968

Number of observations: 64, Error degrees of freedom: 59
Root Mean Squared Error: 860
Tip: Calculated based on eq.6.24 as: Square Root (SumSqResidual/
(n-p))=> (4.3653e+07/(64-5))^-0.5=860.16    {(n-p) is the Degrees
of freedom (DF)}
R-squared: 0.82,  Adjusted R-Squared 0.808
F-statistic vs. constant model: 67.2, p-value = 2.71e-21
              SumSq       DF      MeanSq        F          pValue
            ________      __    __________    ______     _________

Total       2.4249e+08    63    3.8491e+06
Model       1.9884e+08     4     4.971e+07    67.187     2.709e-21
Residual    4.3653e+07    59    7.3988e+05
****************************************************************
```

Coefficient's *p*-values for variables Sec (*p*-value = 0.35287) and Unv (*p*-value = 0.50575) are not statistically significant, indicating that we should remove these variables from the model. Only Ren and Med are statistically significant and can be included in the model, with *p*-values 0.0052968 and 1.3349e-12, respectively.

We run the model again with only Ren and Med, and we calculate the unstandardized and the standardized (beta) coefficients.

```
*****************************RESULTS*****************************
Linear regression model:
    Income ~ 1 + Med + Ren
Estimated Coefficients:
                    Estimate       SE        tStat        pValue

                    ________     ______    ______       __________


    (Intercept)     6912.6       397.63     17.385     2.5899e-25
    Med             38.473       4.0907      9.405     1.7656e-13
    Ren              3.312       1.1967     2.7676      0.0074669

Number of observations: 64, Error degrees of freedom: 61
Root Mean Squared Error: 852
R-squared: 0.817, Adjusted R-Squared 0.811
F-statistic vs. constant model: 136, p-value = 3.04e-23

Standardized coefficients (beta coefficients):
Linear regression model:
    Income ~ 1 + Med + Ren

Estimated Standardized Coefficients:
              Estimate         SE          tStat           pValue

            __________     ________     __________     __________


(Intercept)  4.6221e-16   0.054298     8.5124e-15               1
Med             0.73632   0.078291          9.405      1.7656e-13
Ren             0.21668   0.078291         2.7676       0.0074669

Number of observations: 64, Error degrees of freedom: 61
Root Mean Squared Error: 0.434
R-squared: 0.817, Adjusted R-Squared 0.811
F-statistic vs. constant model: 136, p-value = 3.04e-23
*****************************************************************
```

Interpretation

Adjusted *R*-squared is 0.811, meaning that 81.1% of the total variation of Income is explained by Med and Ren independent variables. It slightly improves the simple OLS model with MedExpenses (adjusted *R*-squared = 0.791). The higher adjusted *R*-squared typically means that adding Ren to the model (on top of Med) contributed by 2% in explaining the total variation if Income.

As the F-significance is small (3.04e-23 less than 0.05), we reject the null hypothesis. We accept that the regression model is statistically significant and that the trend is a real effect, not due to random sampling of the population. Coefficient's *p*-values are statistical significant (Med:1.7656e-13

Ren:0.007466). The Medical expenses variable is more important (has more effect on income) compared to Ren as its beta ($b_1 = 0.73632$) is more than triple than the Ren beta ($b_2 = 0.21668$).

The purpose of the preceding model is to identify which variables are linked to income and which are not. Education attainment seems to be irrelevant to the income, as neither related variables (Sec, Unv) are statistically significant. We also quantified how much of the total variation of income is explained by including medical expenses and rent along with the effect of each statistically significant variable to the model. The higher the beta value, the more critical the variable is. With this single-step approach (all data analyzed in one step), we explore linkages, not causal relationships. The main conclusion is that a large portion of income's variation (81.1%) is linked to/explained by/dependent on first medical expenses (higher beta) and then monthly rent expenses. Looking at the unstandardized coefficients, keeping rent stable, a 100-unit increase in medical expenses leads to 3,847.3 euros extra annual income. Be careful here. As previously mentioned, MLR at one step does not explore causations. The preceding finding does not mean that by spending more money on medical expenses, one will increase his income (i.e., get a raise in salary). It is not a cause-and-effect relationship. It just links medical expenses to income. It is more rational to assume that because one has more money, he/she can afford to pay more for medical expenses.

Such type of models can be used when we need to estimate a variable, but we cannot accurately measure it. In a survey, for example, asking about personal income might be inappropriate. The question is unlikely to receive an honest answer, as people seldomly reveal their actual income. Having the preceding model at hand, it would be more rational to ask about medical expenses, as we anticipate to get an accurate response. Plugging this value into the model would reveal a good estimate of the actual income. Another example is establishing the relation of the house size to the owner's income through the linear regression model. By using satellite images, we can efficiently estimate the size of a house and then the potential income of its owner without having to conduct a costly survey. Similar regression models are also used to assess the population in areas that censuses are rare or not accurately updated. By utilizing satellite images to measure the size of settlements, estimations can be made for the total living population.

We should also mention that the preceding findings are statistically significant only for the specific dataset. Conclusions drawn are valid only for the specific case study and do not necessarily hold true elsewhere. In this respect, when we report results, we should avoid generalization as well as the temptation to refer to causations when we have not carried out such analysis. It is always advisable to seek other similar studies for the same case study to explore if common findings exist. Comparisons with other geographical regions can be made mostly for discussing the outcomes but not for approving or rejecting the results of these studies.

6.7 Exploratory Regression

Definition

Exploratory regression is a process assisting in the appropriate selection of an OLS model when the number of independent variables is large.

Why Use

Exploratory regression is used to automatically test all possible combinations of different OLS models when many independent variables are available. In the previous sections, we handled a handful of variables to showcase how regression works. As the number of available independent variables increases, the process of selecting the most appropriate model becomes tedious. For example, for 10 independent variables to run all possible combinations (all models with just one variable, all models with two variables and so on up to a final single model that will include all variables), it involves creating 1,023 different regression models. It is a combination problem of selecting *r* by *n* with final variables' order not important but with repetition not allowed (a variable cannot participate more than once in the same model). An additional independent variable would increase the models to be tested to 2,047. Exploratory regression is used to take this burden and run all models' combinations to find the most appropriate one.

Interpretation

For a thorough analysis of how exploratory regression outputs are interpreted, see Exercise 6.1.

Discussion and Practical Guidelines

Exploratory regression is a data-mining tool and not a new regression technique, as it just applies OLS regression to many different models, finally to select those that pass all necessary OLS diagnostic tests. By testing all available models:

- Chances of finding the optimal regression modal are increased.
- Hidden independent variables that were not theoretically associated with dependent variable might be traced.
- Independent variables that are statistically significant in the majority of the tested models are more likley to be associated with the dependent variable.

To evaluate the results of exploratory regression and test whether the regression assumptions are violated, we apply the following tests (see Table 6.3):

- The Jarque–Bera test for residuals normality as proof of not biased residuals.
- The Breusch–Pagan test for checking residual errors' heteroscedasticity.
- The Moran's I test to identify if residuals errors are spatially autocorrelated.

Table 6.3 Diagnostic tests used in exploratory regression. Keep in mind that in these tests the desired result is not to reject the null hypothesis. In other words, we prefer *p*-values larger than the significance levels. For example, if the significance level is 0.05, and the result for the Jarque–Bera test is 0.01, then we reject the null hypothesis and we accept that the residual errors do not follow a normal distribution.

Name of diagnostic test	Detects	Hypothesis tested (When *p*-value is smaller than the significance level, we reject the null hypothesis.)
Jarque-Bera test	Non normality	**Null hypothesis:** normal distribution of residuals errors **Alternative hypothesis:** non-normal distribution of regression errors
Breusch-Pagan test	Heteroscedasticity	**Null hypothesis:** constant variance of residuals error (homoscedasticity) **Alternative hypothesis:** nonconstant variance of regression error (heteroscedasticity)
Moran's I	Spatial autocorrelation	**Null hypothesis:** there is no spatial autocorrelation in residuals errors **Alternative hypothesis:** there is spatial autocorrelation in residual errors

Even with modern-day excessive computer power, it is still not wise to run all possible combinations. Not only does computational time increase but results are hard to interpret. For example, models with more than 10 variables would be too complex to interpret. A function of many variables is inappropriate because it is difficult to comprehend how they are meaningfully interconnected with the dependent variable. For this reason, we can set the maximum number of independent variables, no matter how many variables are available, in order to expediate the overall process and reach more straightforward conclusions.

Additionally, we may set the following parameters:

- The threshold of adjusted *R*-squared accepted, as proof of goodness of fit.
- The *p*-values that coefficients are statistically significant as proof of variables' significance (variables with a justifiable relationship with the dependent variable).
- The VIF threshold value to account for multicollinearity as proof of non-redundant variables (no multicollinearity).
- The *p*-value for the Jarque–Bera test for residuals' normality, as proof of unbiased residuals.
- The *p*-value for the Global Moran's I tool on model's residuals, as a proof of no spatial autocorrelation.

From the practical perceptive, exploratory regression is similar to stepwise regression, but its evaluation is not only based on the change of the adjusted *R*-squared. In exploratory regression, all common tests and diagnostics are used to evaluate the performance of the model, adding more reliability and also offering better insight into the model structure and the associated data.

Moreover, a stepwise regression or a regression defined by the researcher is biased on the order that variables are entered or removed from the model. In contrast, exploratory regression is not based on an initial model specification alerting it by each run, but it creates new models that are self-evaluated each time. In this respect, exploratory regression is more powerful than stepwise regression.

Still, there is some controversy in using the exploratory regression, depending on the way we approach a scientific (geographical in our case) problem. There are roughly two approaches, namely the theoretically based approach and the data-mining approach:

- Theoretically based approach: In this approach, researchers build their assumptions based on theoretical evidence before exploring available data. They conceptually form a model by identifying the variables, backed up theoretically, and then adopt a model to solve the problem. The extent of researching additional variables is small. For example, when we model the variable *House price*, the theoretical approach would mainly seek relationships among variables – such as location, size or income – that have been reported to have links to a house's value. In the case of exploratory regression analysis, though, models may be overfitted, as the more variables included, the more unstable the model becomes. A second aspect that the theoretically based approach, as opposed to the data-mining approach, is that regression statistics are heavily based on probabilities expressed by the resulting *p*-values. A 95% confidence interval means, for example, that a coefficient may be statistically significant, but there is a 5% chance that it is not significant. In other words, we may reject the null hypothesis when, in fact, we shouldn't (Type I error, see Section 2.5.5). In case of exploratory regression, where we test hundred or even thousand models, Type I error is more common. A smaller *p*-value (e.g., 0.01) is a potential solution, but, again, the problem remains.
- Data-mining approach: Those adhering the data-mining approach advocate that in many complex problems, theories do not provide all necessary background to shape a rigid model. In this respect, experimenting with additional variables might lead to hidden patterns and relations in the dataset that should be further investigated.

Although the pitfalls of exploratory regression are acknowledged, a fusion of both approaches is suggested in this book. It is more rational to move on solid theoretical foundations. Still, analysts/researchers should think outside the box, as doubting is the key to discovering knowledge. Exploratory regression is a good starting point to analyze data and discover potentially hidden patterns. With the knowledge gained regarding variables' associations, we may move forward to deploy other, more advanced models if necessary, including spatial econometric models, spatial regime models and geographically weighted regression.

It is always advised to select variables that are supported by theory, experts' knowledge or common sense. Statistically significant variables that are not supported by any of these approaches are either useless or a significant discovery. However, if it is a significant discovery, it has to be excessively studied before findings announced. For example, suppose we model human health as a function of variables related to monthly groceries purchases. Through exploratory regression analysis, we might discover that the *variable Canned dog food* (i.e., number of cans containing food for our dog) is statistically significant to the model with a positive coefficient. A first interpretation of this model is, "The more canned dog food you buy, the better your health is." A more exaggerated statement would be, "Eating dog food improves your health." Although this is obviously misleading, it's quite common that research results are misinterpreted or lack some theoretical grounds. Common sense suggests that you buy this type of food if you have a dog. If you have a dog, you probably walk the dog daily. You exercise a lot, so it makes more sense why your health is better (see more on cause and effects in Sections 2.3.4 and 6.2.1).

In more complex problems, where common sense or knowledge cannot straightforwardly guide us on whether to reject an "astonishing" finding or not, we might go completely out of the way. It's common for posts on websites and social media or in the news to be quite startling (e.g., "People that read more live longer," "People who eat burgers are wiser," or "People who eat canned dog food are healthier."), as their primary objective is to gain attention. But be aware of the underlying assumptions, methods and data used, as well as the study's overall approach.

6.8 Geographically Weighted Regression

Definition

Geographically weighted regression (GWR) is a local form of linear regression model used to detect heterogeneity and analyze spatially varying relationships (Fotheringham et al. 2002). GWR provides a local model of the dependent variable allowing for coefficients to vary across space by fitting a weighted linear regression model to every spatial object in the dataset (O'Sullivan & Unwin 2010 p. 228). GWR should obey the same assumptions as multiple linear regression (LINE-M).

The model can be expressed as (6.34) (Wang 2014 p. 178):

$$y_i = \beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \cdots + \beta_{mi}x_{mi} + \varepsilon_i \tag{6.34}$$

where i = 1, ..n with n as the total number of locations/observations

y_i is the value of the dependent variable Y at location i
x_{mi} is the observation of the variable X_m at location i
β_{0i} is the intercept of the model at location i
$\beta_{1i}, \beta_{2i}, \ldots \beta_{mi}$ is the coefficient set at location i
m is the total number of independent variables
ε_i is the error at location i

GWR estimates the coefficients β_i at each location i such as (Wheeler & Páez 2010 p. 462) (6.35):

$$\beta_i = (X^T W_i X)^{-1} X^T W_i Y \tag{6.35}$$

where

Y is the n-by-1 vector of the dependent variable
X is the design matrix of the independent variables (containing a leading column of ones for the intercept)
W_i is the n-by-n diagonal matrix with diagonal elements being the weights w_{ij} at location i (6.36).

$$W_i = \begin{bmatrix} w_{i1} & 0 & \ldots & 0 \\ 0 & w_{i2} & \ldots & 0 \\ \ldots & \ldots & w_{ij} & \ldots \\ 0 & \ldots & \ldots & w_{in} \end{bmatrix} \tag{6.36}$$

w_{ij} is the weight between the regression location i and the location j. w_{ij} takes a value between [0, 1]. The weight matrix W_i has to be calculated prior to local regression coefficients' estimation. There are numerous weighting schemes that can be used, such as the binary scheme (Brunsdon et al. 1996) (6.37):

$$w_{ij} = \begin{cases} 1 \; if \; d_{ij} \leq d \\ 0 \; if \; d_{ij} > d \end{cases} \tag{6.37}$$

In this case, if an object j lies less than a distance threshold value d from i, it gets a weight of 1. If it lies further away than this threshold value, it gets a zero weight. Spatial kernels are most commonly used on the calculation of weights, as shown in the next section.

Why Use

GWR is used to better handle spatial heterogeneity by allowing the regression coefficients (and, consequently, the relationships among the variables), to vary from place to place, in a similar way that that variables' values vary in space (Wheeler et el. 2010 p. 461). Instead of having a global intercept and global coefficients, local values are estimated to better reflect the spatial heterogeneity of the phenomenon studied.

6.8.1 Spatial Kernel Types

In practice, a spatial kernel is used to provide the geographic weighting for the GWR method and should be selected before calibrating a GWR model. Objects closer to the location *i* are weighted more than those further away, based on the assumption of spatial autocorrelation. There are two kernel types (functions) that are most commonly used in the context of GWR, namely fixed and adaptive.

- Fixed: The Gaussian kernel function (a monotonically decreasing function) is used to calculate local weights and is based on a fixed non-varying bandwidth (*h*) distance (6.38):

$$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2\right) \tag{6.38}$$

where d_{ij} is the distance between locations *i* and *j*, and *h* is a kernel bandwidth parameter that controls the decay. The weighting between locations *i* and *j* will decrease according to the Gaussian curve, as the distance between the two points increases (Brunsdon et al. 1996). While the bandwidth is fixed, a different number of spatial objects might be used for calculating the weights for different locations. For large-bandwidth values the gradient of the kernel becomes less steep, and as such, more locations/observations are included in the local estimations. Bandwidth *h* has to be optimally selected for better GWR performance (as explained in next section).

- Adaptive: The bi-square kernel function used to calculate local weights is based on the number *N* of nearest neighbors. In this case, the same number of spatial objects is used to calculate the weights for each location *i* (6.39) (Wheeler et al. 2010 p. 464).

$$w_{ij} = \begin{cases} \left(1 - \frac{d_{ij}^2}{d_{iN}^2}\right)^2 & \textit{if } j \textit{ is one of the } N\textit{-th nearest points of } i \\ 0 & \textit{otherwise} \end{cases} \tag{6.39}$$

Where d_{iN} is the distance of location *i* from the *N*-th nearest neighbor (for example, if we select five nearest neighbors, then *N* = 5). In this case, we have to find the optimal number *N* of nearest neighbors

The choice of the kernel type depends on the problem and on the spatial distribution of the objects. In cases where objects are evenly distributed following a dispersed spatial point pattern, a fixed kernel type is a reasonable selection. When spatial objects are clustered (also revealing spatial autocorrelation), an adaptive kernel type is more appropriate. Adaptive kernel is most widely used, as spatial autocorrelation is evident in most geographical problems.

6.8.2 Bandwidth

The bandwidth of the kernel type – that is, the unknown parameter *h* in the fixed kernel type – and the number of nearest neighbors *N* in the adoptive kernel type define the size of each kernel. There are three main methods for bandwidth selection/estimation:

- **AICc method.** The corrected AIC estimates the bandwidth, which minimizes the AICc value.
- **Cross-validation (CV) method.** The CV method calculates the bandwidth that minimizes the cross-validation score.
- **User-defined method.** The distance or the number of neighbors that will be used to define the kernels are selected by the researcher.

From the three methods, the last one is not automated and does not guarantee any minimization. As GWR heavily relies on the weights' matrix, it is crucial to define this matrix most appropriately. The optimal bandwidth is calculated to keep a balance between bias and variance. A small bandwidth leads to considerable variance in local estimates, while a large bandwidth leads to a significant bias in the local estimates (Fotheringham et al. 2002). The first two methods automatically esimate the bandwidth value based on well-defined criteria and are preferred to the user-defined bandwidth method. Cross-validation method and AICc method are commonly used (Fotheringham et al. 2002, Wheeler et al. 2010 p. 466).

6.8.3 Interpreting GWR Results and Practical Guidelines

GWR results can be evaluated and interpreted in three ways: (A) by using common statistical diagnostics, (B) by mapping standardized residuals and (C) by mapping local coefficients.

(A) Common Diagnostics

A GWR model is firstly compared to an OLS model. The OLS is also called the global model, as it provides a single model for the entire study area. In this respect, GWR is referred to as the local model. Typically, we start by comparing adjusted *R*-squared, AICc and other typical metrics between the local and the global model. Issues of multicollinearity should be mentioned in cases of condition index values larger than 30 (see Lab Exercise 6.3). The main advantage of GWR is that it provides local estimates that can be mapped and further analyzed through spatial statistics.

(B) Mapping Standardized Residuals (Output Feature Class)

Mapping the standardized residuals allows tracing the following:

1. **High or low residuals**. Residuals reveal how much spatial data variations are not explained by the independent variables (Krivoruchko 2011 p. 485). In case that standardized residuals are relatively high, then we may locate under- or overpredictions. Underprediction means that the fitted value is lower than the observed. In this case, residuals are positive. Overprediction means that the fitted value is higher than the observed. In this case, residuals are negative. If under- or overpredictions occur systematically, then there is bias in the model. From the geographical analysis perspective, it is necessary to delve deeper if we identify extremely low or high residuals, as underlying processes at play might explain their presence. For example, suppose we study humans' respiratory health, and after using GWR, we locate extremely large residuals in a region. These extreme values may be an indication of a pollution source (e.g., factory, or industry not conforming with regulations).
2. **Spatial autocorrelation.** We should run a spatial autocorrelation test to identify if spatial autocorrelation exists in residuals. If there is not, then GWR removed the potential spatial autocorrelation existing in the global model. If spatial autocorrelation remains and clusters of over- and/or underpredicted values are formed, this is a sign of an ineffective GWR model, possibly due to omitting one or more independent variables (model misspecification). On the other hand, if we locate clusters of residuals, we should analyze why this clustering occurs. For example, there might be some specific regulations that apply to these areas and not to the others, and as such, local regression models are misspecified. Keep in mind that our ultimate scope is to perform geographical analysis and trace trends and spatial processes that lead to meaningful results and not always to create models that perform at a near-optimally predictive way. Although a model with high predictive performance is sometimes desired, problem complexity, model misspecification and wrong model adoption might drive us away from building a robust model. Still, conclusions can always be drawn if we interpret findings based on the identified relationships, even if the predictive model is weak.

(C) Mapping Local Coefficients (Coefficient Raster Surfaces)

Mapping local coefficients (through a raster surface for points or by polygon rendering, in cases of polygons) shows the variation in the coefficient estimation for each independent variable. Moreover, it shows how much each independent variable impacts the dependent variable locally. For example, suppose we model *House price*, and one of the independent variables is *House size*. Areas with high coefficients in *House size* are areas where the estimation of *House price* is more influenced by this variable compared to other areas with lower coefficients (of the same variable). Directional trends or spatial clusters might also be spotted when mapping coefficients, leaving room for additional geographical analysis and unraveling hidden spatial processes. The range of the local coefficients might be high or low with the same sign. Still, there are rare cases where the majority of coefficients have a specific sign, and relatively few observations have the opposite one. In other words, a coefficient might have both negative and positive values. Although this seems not accurate, it typically suggests that some coefficient values will be zero or very close to zero. In OLS, the t-statistic is used to test if a coefficient is statistically significant. In GWR, there is no such test carried out, and those coefficients near zero may be regarded insignificant (Fotheringham et al. 2002). From the policy perspective, the coefficient analysis may help to identify the following types of variables:

- Statistically significant global variables that do not vary significantly from place to place. This type of variables can be used for analysis at the regional level. For example, the percentage of illiteracy may not vary significantly among postcodes for a cluster of cities in a remote prefecture of a developing country. A regional policy, such as offering free language lessons in each municipality of the region, might be appropriate to decrease rates of illiteracy.
- Statistically significant global variables that vary significantly from place to place. This type of variables can be used for analysis at the local level. For the illiteracy example, by using GWR coefficient mapping, we may identify specific variables that are related to high illiteracy rates in specific areas – e.g., lack of easy access to a learning center. By applying a local policy through better spatial planning, we can select only those areas scattered in the entire region that lack learning centers and create new ones.
- Statistically significant variables that are spatially autocorrelated. Although we cannot include these variables in the GWR model, still we can use them to make conclusions for applying local policies. For example, high values of illiteracy rates may cluster to a specific part of the study area (e.g., the most deprived one). Local policies can be applied again to those areas suffering from high illiteracy rates.

Practical Guidelines:

- GWR is a linear model and must follow the same assumptions the linear regression model does (LINE-M).
- GWR performs better when applied to large datasets with hundreds of features. As a rule of thumb, spatial objects should exceed 160. In cases of smaller datasets, results might be unreliable (ESRI 2016b).
- Data should be projected using a projected coordinate system rather than a geographic coordinate system, as distance calculations are essential to the creation of the spatial weights matrix.

- Begin with an OLS model to specify a reasonably good model. Test VIF values for multicollinearity. VIF values larger than 7.5 reveal global multicollinearity and will prevent GWR from producing a good solution. Use the exploratory regression tool also if the dataset is large.
- Map each independent variable separately and look if clustering emerges. Keep in mind that as spatial autocorrelation exists in most real-world problems, we expect a degree of spatial clustering. If spatial clustering (in the values of a specific independent variable) is very pronounced, then we should consider removing this variable.
- The dependent variable in GWR cannot be binary.
- Remove any dummy variables representing different regimes (geographical areas).
- If multicollinearity is detected, we can remove variables or combine them to increase value variation. For example, if we analyze E*xpenses for supermarket* and *Expenses for groceries* and we identify multicollinearity between them, we can just add these values and create a new variable labeled *Expenses*.
- Problems of local multicollinearity might exist when the values of an independent variable cluster geographically. The condition number reveals if local multicollinearity exists. As a rule of thumb, condition values larger than 30, equal to null or negative reveal multicollinearity, and results should be examined with caution. When categorical variables have few categories or regime variables exist, we run the risk of spatial clustering and, thus, local multicollinearity.
- Mapping coefficients offers a comprehensive view of the spatial variability of coefficient values. By applying a rendering scheme (e.g., cold to hot) to the raster surface of the coefficients of a specific variable, we trace fluctuations across the study area, locate large or small values and potentially identify spatially clustering.
- Run spatial autocorrelation tests in residuals. If the test is statistically significant, then the model may be misspecified.
- Map residuals to identify any heteroscedasticity. If there is heteroscedasticity, the model may be misspecified.
- In cases of misspecification, variables might have been omitted from the model. Use exploratory regression to identify any missing variables.

6.9 Chapter Concluding Remarks

- Before we begin our analysis, we should inspect if missing values exist.
- Underprediction means that the fitted (predicted) value is lower than the observed.
- Overprediction means that the fitted value is higher than the observed.

- If under or overpredictions occur systematically, then there is bias in our model.
- By regression, we built a probabilistic model that is related to a random amount of error (*e*).
- 82% *R*-squared typically means that 82% of the variation in *Y* is due to the changes in values of *X*, meaning that *X* is a good predictor for *Y* having a probabilistic linear relationship.
- In case of a low *R*-squared, we may add more independent variables or more observations (if available) to check if the *R*-squared will increase.
- Adjusted *R*-squared reveals if an extra variable is useful or not.
- Predicted *R*-squared is more valuable than adjusted *R*-squared because it is calculated based on existing observations with known *Y* values that are not included in the model creation (out of sample).
- By standardizing residuals, we calculate how many standard deviations a residual lies from the mean.
- If a residual plot indicates that some assumptions are severely violated, then we cannot use the model although some statistics may be statistically significant.
- Overfitting occurs when a model has too many parameters relative to the number of observations.
- To diagnose multicollinearity, the following metrics can be used: Singular value decomposition, condition index (CI), variance decomposition proportions, and variance inflator factor (VIF).
- Before we evaluate the t-statistic, we should check the Koenker (BP) statistic.
- When the Koenker (BP) statistic is statistically significant, then heteroscedasticity exists, and the relationships of the model are not reliable. In this case, robust t-statistic and robust probabilities should be used instead of the t-statistic *p*-value.
- The joint Wald test should be checked when the Koenker (BP) statistic is statistically significant.
- In case that the Koenker (BP) statistic is statistically significant, the joint Wald test should be inspected instead of the F-significance.
- An influential point might be both an outlier point (at the *Y* direction of the residual plot or far away from the regression line) and a point with high leverage (far away from the *X* or fitted values mean).
- Beta coefficients are used to compare the importance of each independent variable to the model especially when the units or the scale among the independent variables are different.
- Spatial kernel is used to provide the geographic weighting for the GWR method.
- In case of multicollinearity, AIC and CV methods might not accurately calculate the optimal distance or the optimal number of neighbors for the

bandwidth parameter selection. Detect and remove multicollinearity issues first, and then apply the previous methods.

Questions and Answers

The answers given here are brief. For more thorough answers, refer back to the relevant sections of this chapter.

Q1. What is ordinary least squares regression? How does OLS work? What is a residual?

A1. Ordinary least squares (OLS) is a statistical method for estimating the unknown parameters (coefficients) of a linear regression model. OLS regression works by fitting a line to the data points by determining the parameters of the model that minimize the sum of the squared vertical distances from the observed points to the line (sum of the squared residuals). The difference of the observed value with the predicted (fitted) value is the model error called residual (*e*).

Q2. What is the coefficient of determination *R*-squared, and why it is used?

A2. The coefficient of determination denoted as R^2 (*R*-squared) is the percentage of the variation explained by the model. It is calculated as the ratio between the variation of the predicted values of the dependent variable (explained variation *SSR*) to the variation of the observed values of the dependent variable (total variation *SST*). Its values range from 0 to 1.

Q3. What is the F-test and the F significance level, and why it is used in comparison to *R*-squared?

A3. The F-test and the F-significance level are used to test if the regression line fit is statistically significant, or else if the regression has been successful at explaining a significant amount of the variation of *Y*. *R*-squared assesses the strength of the relationship among predicted and independent variables. Still, it does not provide a measure of how statistically significant the hypothetical relationship is. To do so, we use the F-test of overall significance.

Q4. What is a standardized residual plot, and what is used for?

A4. Standardized residual plot is the plot of the fitted values against the standardized residuals (the *y*-axis depicts the number of standard deviations, and the *x*-axis depicts the fitted values). Using a standardized residual plot assists in assessing the quality of the regression model and identifying if normality or homoscedasticity assumptions are violated. Ideally, the standardized residual plot should exhibit no particular pattern; and data points should be randomly distributed, have constant variance and be as close to the *x*-axis as possible. These conditions satisfy the third (normal distribution of *e*) and the fourth (homoscedasticity) assumptions of linear regression (see Section 6.4). If clusters, outliers or trends exist in the data point pattern, it is an indication of one or more assumptions' violation.

Q5. What are the standardized (beta) coefficients?

A5. Standardized coefficients are the coefficients resulting when variables in the regression model (both dependent and independent) are converted to their z-scores before running the fitting model. The standardized regression coefficients – also called beta coefficients, beta weights or just beta – are expressed in standard deviation units thus removing the measurement units of the variables. They are used to decide which independent variable *X* is more important to model *Y*.

Q6. What is influential point, and why tracing it is important?

A6. Influential point is any data point that substantially influences the geometry of the regression line. Influential points should be identified and potentially removed as they significantly distort the geometry of the regression line leading to inaccurate models.

Q7. Which are the assumptions of multiple linear regression?

A7. MLR builds a probabilistic model based on five assumptions. (1) Linearity: The relationships between dependent and independent variables are linear. (2) Independence: The residuals are independent of each other, meaning that an error *e* in a data point *X* does not affect by any means another error *e* for some different value of the same variable *X*. (3) Normality: The distribution of errors *e* at any particular *X* value is normal. 4) Homoscedasticity: Errors *e* have a zero mean equal variance and a constant standard deviation at any particular value of *X*. (5) Multicollinearity absence: Variables *X* should be independent of each other exhibiting no multicollinearity.

Q8. What is exploratory regression, and why is it used?

A8. Exploratory regression is a process that facilitates the appropriate selection of an OLS model when the number of independent variables is large. Exploratory regression is a data-mining tool and not a new regression technique, as it just applies OLS regression to many different models to select those that pass all necessary OLS diagnostic tests. It is used to automatically test all possible combinations of different OLS models when many independent variables are available.

Q9. What is geographically weighted regression (GWR)? Which assumptions should obey?

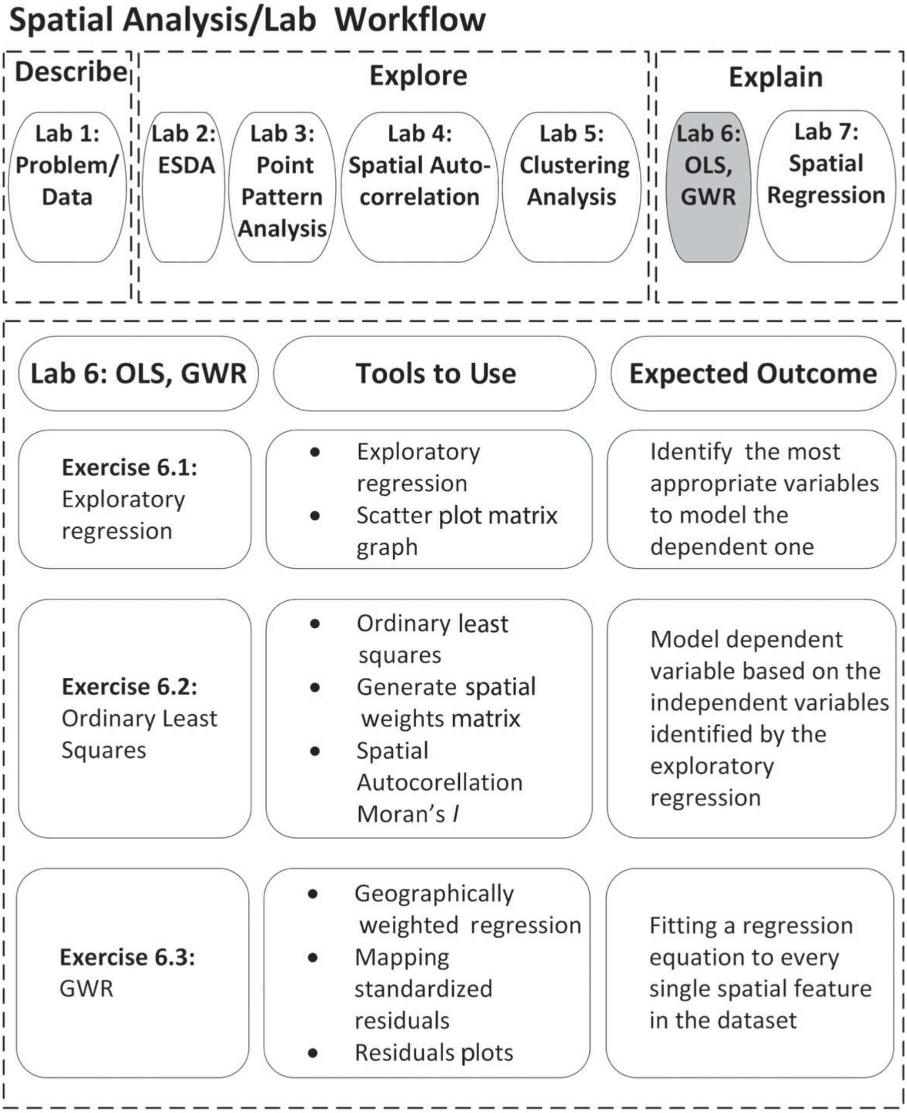
A9. GWR is a local form of linear regression model used to detect heterogeneity and analyze spatially varying relationships. GWR provides a local model of the dependent variable allowing for coefficients to vary across space, by fitting a weighted linear regression model to every spatial object in the dataset. GWR should obey the same assumptions as multiple linear regression (LINE-M).

Q10. Why spatial kernels and bandwidth are used in GWR?

A10. A spatial kernel is used to provide the geographic weighting for the GWR method and should be selected before calibrating a GWR model. Objects closer to the location *i* are weighted more than those further away, based on the assumption of spatial autocorrelation. The bandwidth of the kernel type – that is, the unknown parameter *h* in the fixed kernel type – and the number of nearest neighbors *N* in the adoptive kernel type define the size of each kernel. There are three main methods for bandwidth selection/estimation: AICc method, cross-validation (CV) method and the bandwidth method.

LAB 6
OLS, EXPLANATORY REGRESSION, GWR

Overall Progress


Spatial Analysis/Lab Workflow
Describe
Explore
Explain
Lab 1: Problem/ Data
Lab 2: ESDA
Lab 3: Point Pattern Analysis
Lab 4: Spatial Auto-correlation
Lab 5: Clustering Analysis
Lab 6: OLS, GWR
Lab 7: Spatial Regression

Lab 6: OLS, GWR	Tools to Use	Expected Outcome
Exercise 6.1: Exploratory regression	• Exploratory regression • Scatter plot matrix graph	Identify the most appropriate variables to model the dependent one
Exercise 6.2: Ordinary Least Squares	• Ordinary least squares • Generate spatial weights matrix • Spatial Autocorellation Moran’s *I*	Model dependent variable based on the independent variables identified by the exploratory regression
Exercise 6.3: GWR	• Geographically weighted regression • Mapping standardized residuals • Residuals plots	Fitting a regression equation to every single spatial feature in the dataset

Figure 6.9 Lab 6 workflow and overall progress.

Scope of the Analysis

This lab deals with

- **Objective 4:** Modeling. Identifying socioeconomic drivers behind people’s monthly expenses (including those for coffee-related services) (see Table 1.2).

In this lab, we model Expenses for those people living in the city, based on various socioeconomic variables (see Figure 6.9). The overall scope of the analysis is to identify the socioeconomic drivers behind monthly expenses (including coffee-related expenses). From the investor perspective, this type of analysis can be used, first, to better trace areas with the higher potential of expenses and, second, to identify what makes people spend more in coffee-related purchases. This type of analysis is not linked directly to finding an optimal location (as we did in previous labs), but it focuses on modeling relationships that can be used for better market penetration and customer analysis. A complete analysis should include much more detailed variables such as consumer preferences, everyday habits, type of job and money spent on coffee shops. For educational reasons and to keep analysis short, we only focus on 10 basic socioeconomic variables. We start with exploratory regression first, and then we analyze the most appropriate model resulted, with additional OLS diagnostics. Finally, we apply GWR to handle spatial heterogeneity.

Exercise 6.1 Exploratory Regression

Before we run the tool, we show how ArcGIS presents the results of exploratory regression analysis, and we also highlight how we should interpret outputs (for more details on this section, you can visit ESRI's website, ESRI 2016a). The primary output is a report file summarizing the most successful regressions. Keep in mind that there might be cases where no model passes all tests. The report can still be valuable, as we can analyze which coefficients (and thus associated variables) are consistently statistically significant and which diagnostics most commonly yield non-statistical significant results. In cases of no model passing the tests, we may relax the parameter values and rerun exploratory regression. Results can be summarized in the following five sections:

- **Section 1: Best models summary.** It provides the three best models with the highest adjusted *R*-squared. It also provides the following diagnostics:
  1. AICc (corrected Akaike Information Criterion), as a general measure of goodness of fit, and as a measure to compare different models. The one with the smallest value provides best goodness of fit among those tested (for more on AICc, see Section 6.8.3).
  2. JB (Jarque–Bera) *p*-value, as test of residuals normality.
  3. K(BP) (Koenker's studentized Breusch–Pagan) *p*-value, as a measure of heteroscedasticity.
  4. VIF (variance inflation factor). In order to assess multicollinearity, the largest variance inflation factor is reported.
  5. SA is the *p*-value of the Global Moran's I test for detecting spatial autocorrelation.

Exercise 6.1 (*cont.*)

The models (if any) passing all tests are labeled as "Passing Models." The results are reported as many times as the *Maximum Number of Explanatory Variables* parameter has been set. As mentioned before, we might have many variables, but for simplicity, we consider only those models having no more than a specific number of independent variables. For instance, for twenty independent variables if we set the *Maximum Number of Explanatory Variables* parameter to 5, then the exploratory regression will test the combinations of all models having from one independent variable up to a maximum of five. The output includes the best three models (along with the five diagnostics) with one independent variable, the three best models with two independent variables and so on up to the three best models with five independent variables. Furthermore, all passing models will be reported. In a nutshell, five different subsections will be included in this section, summarizing the best models and the passing ones. If there are no passing models, then the sections that follow reveal a lot regarding the data, assisting in determining which direction the analysis should follow.

- **Section 2: Global summary.** This section lists the five diagnostics and the percentage (as well as the absolute values) of the models that passed each of these tests. In case that no model passes, this section assists in the identification of potential reasons that models do not perform well. For example, if the normality test is not statistically significant for 90% of the models, then the normality assumption is at stake. We might consider transforming some of the variables, as nonlinear relationships might exist in the data (to detect nonlinear relationships we can create a scatter plot matrix, as shown earlier). Potential outliers' existence might also be a cause of non-normality of errors. Furthermore, we might encounter spatially autocorrelated residuals for all models. Spatially autocorrelated residuals is an indication of a missing variable. It is quite common for this omitted variable to be a geographical variable such as the distance from a landmark (e.g., distance from city center). Spatial regimes may also assist on the elimination of spatially autocorrelated residuals. Another approach is to select a good model that passes all or the most of the rest diagnostics and run OLS model to inspect potential problems in the residuals scatter plot.
- **Section 3: Variable significance summary.** Each variable is listed along with three percentage values related to a specific coefficient:
  1. The percentage that a coefficient (and thus the associated variable) is significant. It is the ratio of the number of models that a specific variable is statistically significant to the number of all models tested.

Exercise 6.1 (*cont.*)

2. The percentage that a coefficient is positive.
3. The percentage that a coefficient is negative.

Variables with high percentage of being significant are strong predictors and are expected to have consistently the same sign, either positive or negative. Variables with small percentage of being significant are not explaining the dependent variable. In case we want to drop variables from our model, these variables are the ones we should drop first. In case no passing models emerge, we can use the variables with the highest percentages of being significant and add omitted variables, in anticipation of building models that pass the tests.

- **Section 4: Multicollinearity summary.** It reports how many times each variable with a high VIF (multicollinearity) is included in a model along with all independent variables (in the same model). When two or more variables coexist in many models, with high multicollinearity, they practically explain the same percentage of variation. If we keep only one, we avoid multicollinearity and build a better model. Among those variables that are highly correlated, we may keep the one with the highest percentage of being significant, reported in the previous section, or the one that is more appropriate conceptually.
- **Section 5: Diagnostics summary.** The three models with the highest Jarque–Berra *p*-values (errors' normality) and the three models with the highest Moran's I *p*-values (spatial autocorrelation) are reported in this section along with all other associated diagnostics and independent variables. Models reported here do not necessarily pass all diagnostics. This section is quite helpful when no model passes the tests. By inspecting the *p*-values of the Jarque–Berra test and the Moran's I test, we estimate how far the model lies from fulfilling the assumptions of having normally distributed residuals and nonspatially autocorrelated residuals. For example, if the maximum Moran's I *p*-value is less than 0.001, then spatial autocorrelation of residuals is a major issue. However, if this value is 0.095 with a significance level of 0.10, then we are close to considering that residuals are not spatially autocorrelated. We can also check which independent variables are those that result in higher values (preferred) on these tests, leading to passing or nearly passing models.

The second output of the tool is a table summarizing all models that meet the maximum coefficient cutoff value and the maximum VIF cutoff value. Each row in the table describes a model that meets these two criteria regardless of if it passes the tests overall. All additional diagnostics and independent variables are also included in the same line. This table is supportive when we consider using a model although one or more

Exercise 6.1 (*cont.*)

assumptions are not fulfilled. For example, the most common violated assumption is the errors normality. In such cases, we can sort the table by the AICc values and use the model with the minimum AICc value that also meets as many of the diagnostics as possible except for Jarque–Berra.

Let's proceed to our exercise. In total, 10 independent variables are tested to model the dependent variable (Expenses). We do not define which of the variables will be finally included in the final model. Instead, we lean on the data-mining capabilities that the exploratory regression offers. The 10 independent variables are as follows:

- Population: population per postcode
- Density: people per square meter
- Foreigners: percentage of people of different nationality
- SecondaryE: percentage of people obtained secondary education
- University: percentage of people that graduated from university
- PhD Master: percentage of people obtained a master's degree or PhD
- Income: annual income per capita (in euros)
- Insurance: annual money spent on insurance policy (in euros)
- Rent: monthly rent (in euros)
- Owners: percentage of people living in their own house

If we attempt to test every possible combination of the independent variables, we should build 1,023 different models. Exploratory regression builds these models automatically. Before we perform exploratory regression, we should create a scatter plot matrix to inspect the presence of potential multicollinearity among the independent variables.

ArcGIS Tools to be used: Scatter Plot Matrix Graph, Exploratory Regression

ACTION: Scatter Plot Matrix Graph

We create this plot to trace multicollinearity, outliers or trends among the variables.

Navigate to the location you have stored the book dataset and click Lab6_Regression.mxd

Main Menu > File > Save As > My_Lab6_Regression.mxd

In I:\BookLabs\Lab6\Output

Main Menu > View> Graphs > Create Scatter Plot Matrix >

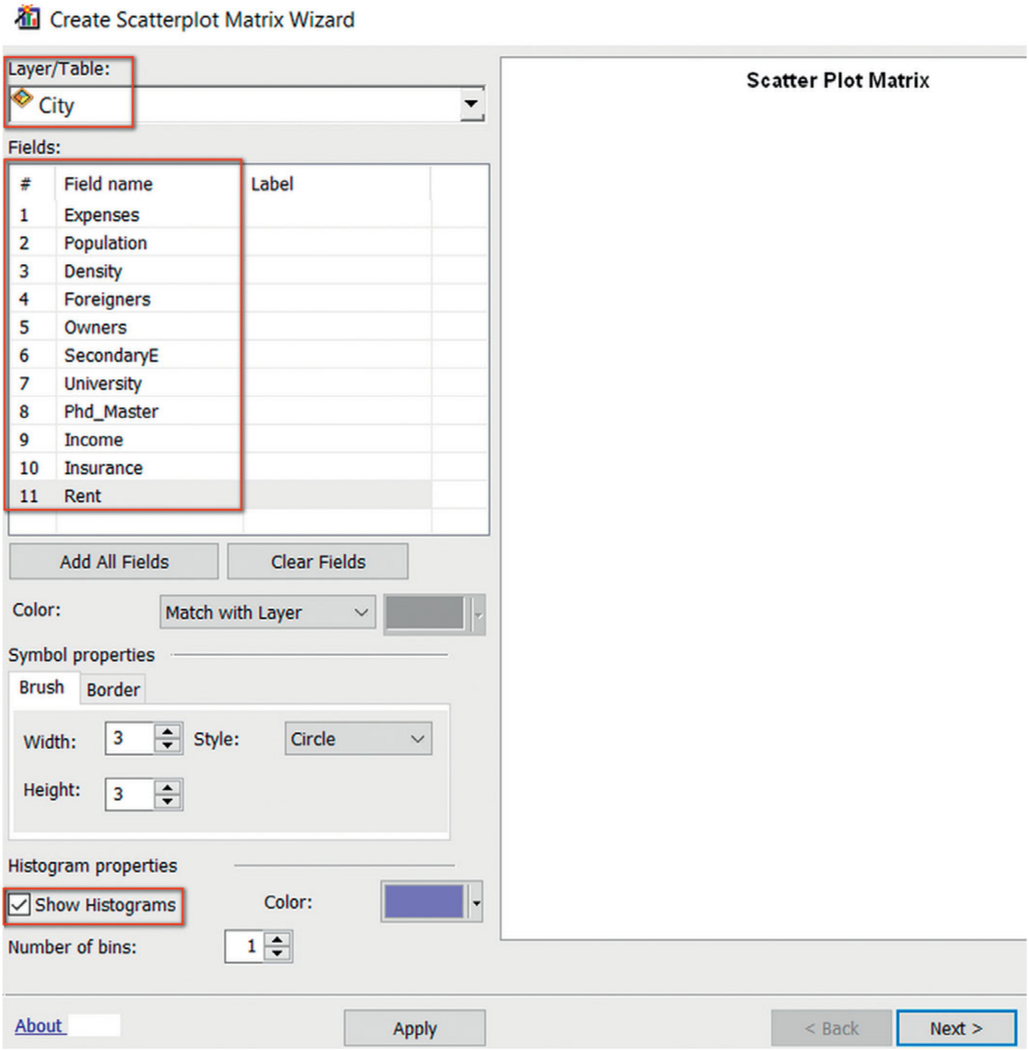
Layer/Table: City

Fields > Click in line 1, under the Field name. Select variables one by one, starting with Expenses and adding the 10 independent variables and the dependent variable, as shown in Figure 6.10.

Exercise 6.1 (*cont.*)

```
Check Show Histograms

Number of bins = 3

Apply > Next > Finish

RC on graph > Add to layout > Return to Data View > Close graph
> Save
```

You should inspect one by one these plots for potential multicollinearity issues or outliers presence (see Figure 6.11).



Create Scatterplot Matrix Wizard

Layer/Table:
City

Fields:

#	Field name	Label
1	Expenses	
2	Population	
3	Density	
4	Foreigners	
5	Owners	
6	SecondaryE	
7	University	
8	Phd_Master	
9	Income	
10	Insurance	
11	Rent	

Color: Match with Layer

Symbol properties

Brush | Border

Width: 3 Style: Circle

Height: 3

Histogram properties

☑ Show Histograms Color:

Number of bins: 1

Scatter Plot Matrix

Figure 6.10 Creating scatter plot matrix.

Exercise 6.1 *(cont.)*

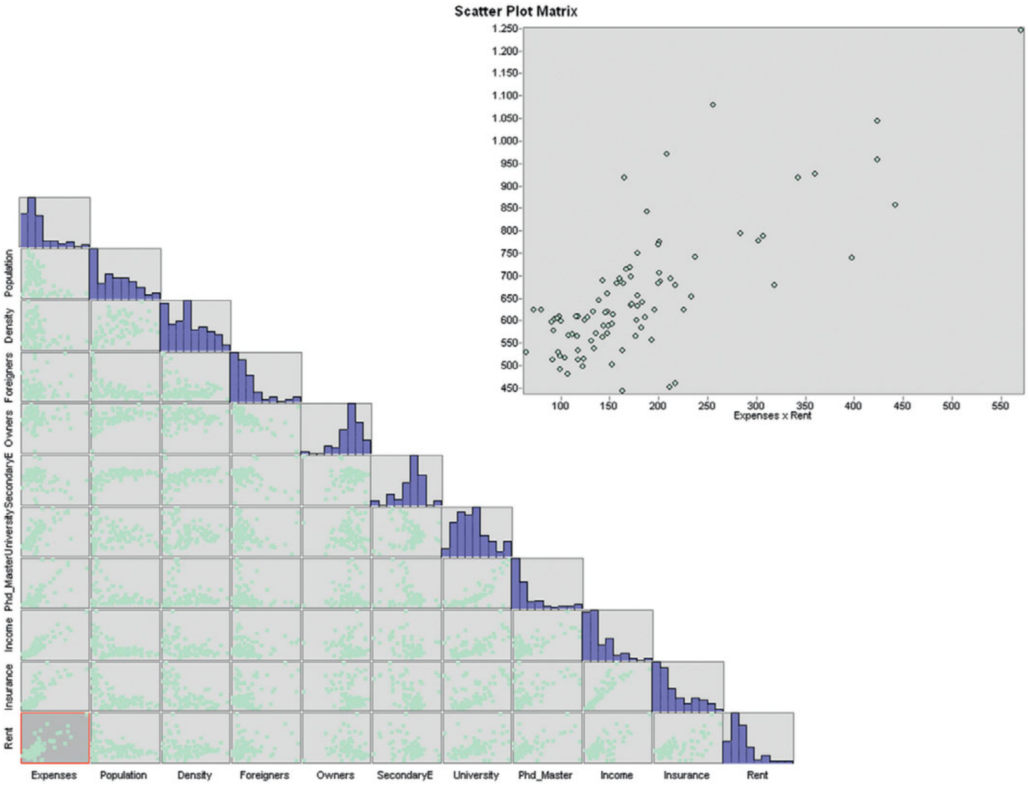


Figure 6.11 Scatter plot matrix of all variables. When clicking a scatter plot in the matrix, a larger version is presented in the upper-right corner of the graph.

ACTION: Exploratory Regression

ArcToolbox > Spatial Statistic Tools > Modeling Spatial Relationships > Exploratory Regression (see Figure 6.12)

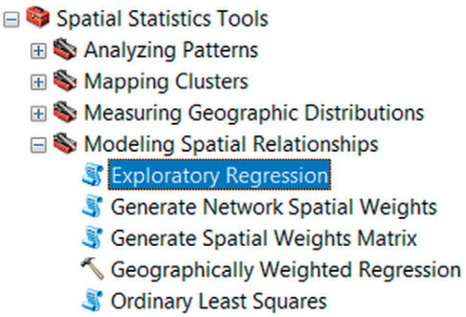

⊟ Spatial Statistics Tools
⊞ Analyzing Patterns
⊞ Mapping Clusters
⊞ Measuring Geographic Distributions
⊟ Modeling Spatial Relationships
Exploratory Regression
Generate Network Spatial Weights
Generate Spatial Weights Matrix
Geographically Weighted Regression
Ordinary Least Squares

Figure 6.12 Exploratory regression tool.

Exercise 6.1 (*cont.*)

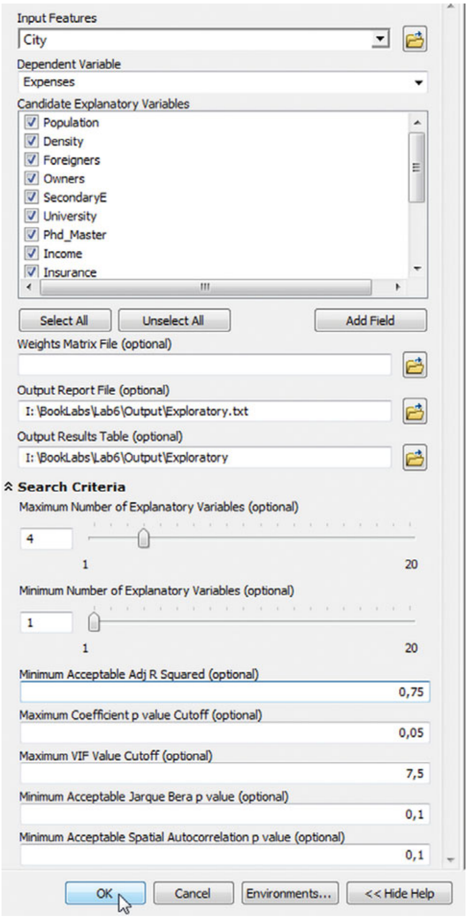
Input Features = City (see Figure 6.13)

Dependent Variable = Expenses

Candidate Exploratory Variables = Population, Density, Foreigners, Owners, SecondaryE, University, PhD Master, Income, Insurance, Rent,

Weights Matrix File = Leave empty. It is used to calculate the spatial autocorrelation of the residuals. By leaving this field empty, the weights are calculated based on the eight nearest neighbors. This weight matrix is not used for OLS calculations.

Output Report File = I:\BookLabs\Lab6\Output\Exploratory.txt



Input Features
City

Dependent Variable
Expenses

Candidate Explanatory Variables
☑ Population
☑ Density
☑ Foreigners
☑ Owners
☑ SecondaryE
☑ University
☑ Phd_Master
☑ Income
☑ Insurance

Weights Matrix File (optional)

Output Report File (optional)
I:\BookLabs\Lab6\Output\Exploratory.txt

Output Results Table (optional)
I:\BookLabs\Lab6\Output\Exploratory

Search Criteria

Maximum Number of Explanatory Variables (optional)
4

Minimum Number of Explanatory Variables (optional)
1

Minimum Acceptable Adj R Squared (optional)
0,75

Maximum Coefficient p value Cutoff (optional)
0,05

Maximum VIF Value Cutoff (optional)
7,5

Minimum Acceptable Jarque Bera p value (optional)
0,1

Minimum Acceptable Spatial Autocorrelation p value (optional)
0,1

Figure 6.13 Exploratory regression dialog box.

Exercise 6.1 (*cont.*)

```
Output Results Table = I:\BookLabs\Lab6\Output\Exploratory

Maximum Number of Explanatory Variable = 4 (We select to build
models with a maximum of four independent variables)

Minimum Number of Explanatory Variable = 1

Minimum Acceptable Adj R Squared = 0.75

Maximum Coefficient p value Cutoff = 0.05

Maximum VIF Cutoff = 7.5

Minimum Acceptable Jargue Bera p value = 0.1

Minimum Acceptable Spatial Autocorrelation p value = 0.1

OK
```

```
Main Menu > Geoprocessing > Results > Current Session >
Exploratory Regression > DC on Output Report File: Exploratory.txt
```

The output report (`Exploratory.txt`) is now presented in detail.

- **Section 1: Best models summary**

```
**************************************************************
Choose 1 of 10 Summary
    Highest Adjusted R-Squared Results
AdjR2     AICc    JB  K(BP)  VIF    SA   Model
 0.81   918.02  0.02   0.00  1.00  0.00  +INCOME***
 0.60   985.34  0.00   0.00  1.00  0.00  +INSURANCE***
 0.57   992.01  0.01   0.00  1.00  0.00  +RENT***
            Passing Models
AdjR2 AICc JB K(BP) VIF SA Model
**************************************************************
Choose 2 of 10 Summary
           Highest Adjusted R-Squared Results
AdjR2   AICc   JB K(BP)  VIF   SA   Model
0.85  900.89 0.00  0.01 1.91 0.00  +INCOME***      +RENT***
0.84  905.78 0.00  0.00 1.66 0.00  +UNIVERSITY**   +INCOME***
0.83  912.56 0.00  0.00 1.47 0.00  +POPULATION*** +INCOME***
       Passing Models
  AdjR2 AICc JB K(BP) VIF SA   Model
**************************************************************
```

Exercise 6.1 (*cont.*)

```
          Choose 3 of 10 Summary
                          Highest Adjusted R-Squared Results
AdjR2   AICc   JB K(BP)  VIF   SA   Model
0.87 884.45 0.00  0.00 2.59  0.26  +UNIVERSITY***  +INCOME***  +RENT***
0.86 895.71 0.00  0.01 2.49  0.60  +POPULATION***  +INCOME***  +RENT***
0.86 896.43 0.00  0.01 1.92  0.11  +OWNERS**       +INCOME***  +RENT***
                    Passing Models
          AdjR2 AICc JB K(BP) VIF SA   Model
          **************************************************************
          Choose 4 of 10 Summary
                                Highest Adjusted R-Squared Results
AdjR2   AICc   JB K(BP)  VIF   SA   Model
0.88 882.52 0.00  0.01 2.64 0.56 +OWNERS**     +UNIVERSITY*** +INCOME***  +RENT***
0.88 883.87 0.00  0.01 3.70 0.86 +POPULATION*  +UNIVERSITY**  +INCOME***  +RENT***
0.88 884.32 0.00  0.00 3.81 0.18 +UNIVERSITY** -PHD_MASTER    +INCOME***  +RENT***
                 Passing Models
          AdjR2 AICc JB K(BP) VIF SA   Model
          **************************************************************
```

Interpreting results: This section provides the three best models that have between one and four independent variables (we set this as one of our parameters). Although many models have high adjusted *R*-squared (>0.80), there is no passing model (meeting all criteria). Multicollinearity is not an issue in the aforementioned models, as VIF is smaller than 4 in all cases. A closer inspection of the table reveals that models do not pass errors normality (JB) and heteroscedasticity K(BP) tests, but some of them pass the spatial autocorrelation (SA) test. The fact the heteroscedasticity in residuals exists is a potential indication that GWR should be applied. We have to go through the following sections to better understand why these assumptions are violated.

- **Section 2: Global summary**

```
********** Exploratory Regression Global Summary (EXPENSES) ************
                 Percentage of Search Criteria Passed
               Search Criterion Cutoff Trials # Passed % Passed
              Min Adjusted R-Squared      > 0.75    385       180     46.75
             Max Coefficient p-value      < 0.05    385        60     15.58
                       Max VIF Value      < 7.50    385       367     95.32
             Min Jarque-Bera p-value      > 0.10    385        16      4.16
Min Spatial Autocorrelation p-value      > 0.10     15         6     40.00
          ______________________________________________________________
```

Exercise 6.1 (*cont.*)

Interpreting results: From the results, we identify one major problem. Only 4.16% of the models passed the errors normality test (Jarque–Bera). Regarding the violation of normality assumption, we might consider transforming some of the variables, as nonlinear relationships might exist in the data. Potential outliers' existence might also be a cause of non-normality of errors. To detect nonlinear relationships or outliers, we can check the scatter plot matrix as produced in Figure 6.11. For example, comparing Expenses to all other variables indicates that there might be a need for transformation in Population or PhD_Master (see Figure 6.11). In addition, by close inspection of the scatter plot matrix, we locate outliers – for example, in SecondaryE with Expenses or Insurance with Expenses. Removing independent variables is an alternative way that may improve normality issues.

We will not deal with outliers or transformations, as this example is intended to showcase the exploratory regression method and not to produce a complete analysis for the city. In real case studies, though, we should test potential transformations like logarithmic or Box–Cox and evaluate if results are improved. In fact, applying spatial regression methods may mitigate the previous problems by including spatial variables and handling spatial autocorrelation.

Another problem we encounter in this example is that residuals are spatially autocorrelated for nearly 60% of the models tested. Spatial autocorrelation of residuals mostly occurs when a variable is missing from the model. Potentially, one or more geographical variables are omitted, such as the distance from a point (e.g., distance from the city center). To identify which variable is missing, we can also select a good model that passes most of the diagnostics and run the OLS model to map residuals. Inspecting residuals' distribution along with knowledge of the study area assists in identifying potential missing variables. The missing variable might also be the spatially lagged Income, but this is a topic of spatial econometrics covered in Chapter 7 (a spatial lag or spatial error model might be the solution to removing spatial autocorrelation).

- **Section 3: Variable significance summary**

```
--------------------------------------------------------------
Summary of Variable Significance
Variable   % Significant % Negative % Positive
INCOME            100.00       0.00     100.00
RENT              100.00       0.00     100.00
UNIVERSITY         84.62       0.00     100.00
POPULATION         74.62      49.23      50.77
INSURANCE          71.54      26.92      73.08
DENSITY            60.77      50.77      49.23
PHD_MASTER         52.31       6.15      93.85
SECONDARYE         18.46      72.31      27.69
FOREIGNERS         14.62      32.31      67.69
OWNERS             12.31      12.31      87.69
--------------------------------------------------------------
```

Exercise 6.1 (*cont.*)

Interpreting results: Coefficients of Income and Rent are consistently significant and positive and are regarded as strong predictors of Expenses. University is by 84.62% significant and can be considered to be included in the models. In all tests (100%), the coefficient of University has a positive sign. On the other hand, Population and Insurance, which have significant coefficients for 74.62% and 71.54% of the models tested, respectively, do not have a constant sign and, as such, are not reliable predictors of expense. The rest of the variables are not stable and should not be included in the final model. In conclusion, in our model, we should use Income, Rent and University

- **Section 4: Multicollinearity summary**

```
---------------------------------------------------------------
Summary of Multicollinearity
Variable      VIF Violations Covariates
POPULATION   2.49      0       ------
DENSITY      1.89      0       ------
FOREIGNERS   2.08      0       ------
OWNERS       1.92      0       ------
SECONDARYE   1.72      0       ------
UNIVERSITY   3.55      0       ------
PHD_MASTER   4.49      0       ------
INCOME      10.97      16      INSURANCE (10.53)
INSURANCE    8.60      6       INCOME (10.53)
RENT         2.09      0       ------
```

Interpreting results: Income exhibits multicollinearity with Insurance. These variables should not be included concurrently in the same model. We keep Income, as it is significant in more cases (100%) and also has constant sign, compared to Insurance (71.54%). Income conceptually fits better the scopes of this analysis.

- **Section 5: Diagnostic Summary**

```
---------------------------------------------------------------

                              Summary of Residual Normality (JB)
      JB     AdjR2       AICc       K(BP)      VIF       SA     Model
0.996766 0.833190 911.084266  0.000000 2.139714 0.000005 +POPULATION*** +FOREIGNERS  +OWNERS  +INCOME***
0.964232 0.821283 917.289778 0.000000 1.730483 0.000000 +FOREIGNERS  +OWNERS*  +SECONDARYE  +INCOME***
0.879930 0.822030 915.667491 0.000000 1.562164 0.000000 +FOREIGNERS  +OWNERS*  +INCOME***
---------------------------------------------------------------------------------

                     Summary of Residual Spatial Autocorrelation (SA)
       JB     AdjR2       AICc      K(BP)       VIF       SA    Model
0.855801 0.876721 883.867623 0.000000 0.011252 3.695209  +POPULATION*  +UNIVERSITY**  +INCOME***  +RENT***
```

Exercise 6.1 (*cont.*)

```
0.604094 0.857425 895.710602 0.000000 0.012039 2.488589  +POPULATION***  +INCOME***  +RENT***
0.558420 0.878560 882.515318 0.000000 0.008369 2.643798  +OWNERS**  +UNIVERSITY***  +INCOME***  +RENT***
_____________________________________________________________________________________________________
      Table Abbreviations
      AdjR2 Adjusted R-Squared
      AICc  Akaike's Information Criterion
      JB    Jarque-Bera p-value
      K(BP) Koenker (BP) Statistic p-value
      VIF   Max Variance Inflation Factor
      SA    Global Moran's I p-value
      Model Variable sign (+/-)
      Model Variable significance (* = 0.10, ** = 0.05, *** = 0.01)
```

Interpreting results: The model with the highest *p*-value for errors normality (Jarque–Berra) includes the variables +POPULATION*** +FOREIGNERS +OWNERS +INCOME***.This model is unreliable, as most of the variables included are not statistically significant (Section 3).

The model with the highest *p*-value for spatial autocorrelation includes the variables +POPULATION* +UNIVERSITY** +INCOME*** +RENT***. From the geographical analysis perspective, spatial autocorrelation is not evident in this model. Still, it includes Population – which, as shown in Section 3, has unstable sign and is not advised to be included in the model.

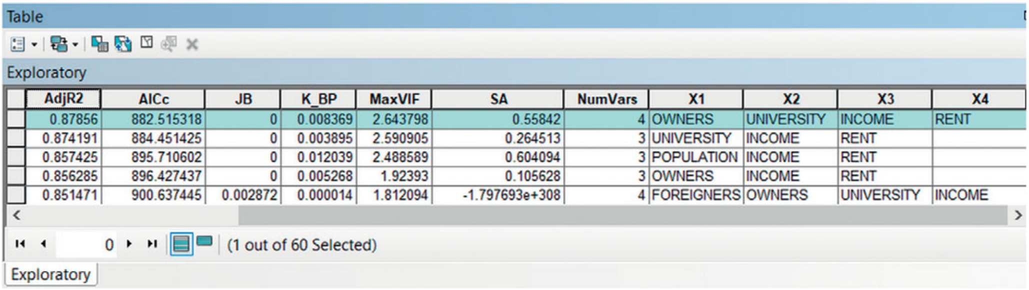
Finally, an additional output is a table summarizing the models that meet the maximum coefficient cutoff value and the maximum VIF cutoff value (see Figure 6.14). In conjunction with the previous findings, we use this table to select the appropriate model for OLS, even if some tests are rejected.

ACTION: Open Exploratory Table (List by Source)

RC Exploratory > Open > RC AdjR2 > Sort Descending

Sorting the table by adjusted *R*-squared, we get the best models relatively to this metric, no matter if they meet all passing criteria.

We only present the first five out 60 models included in the table (see Figure 6.14).



Table

Exploratory

AdjR2	AICc	JB	K_BP	MaxVIF	SA	NumVars	X1	X2	X3	X4
0.87856	882.515318	0	0.008369	2.643798	0.55842	4	OWNERS	UNIVERSITY	INCOME	RENT
0.874191	884.451425	0	0.003895	2.590905	0.264513	3	UNIVERSITY	INCOME	RENT	
0.857425	895.710602	0	0.012039	2.488589	0.604094	3	POPULATION	INCOME	RENT	
0.856285	896.427437	0	0.005268	1.92393	0.105628	3	OWNERS	INCOME	RENT	
0.851471	900.637445	0.002872	0.000014	1.812094	-1.797693e+308	4	FOREIGNERS	OWNERS	UNIVERSITY	INCOME

0 (1 out of 60 Selected)

Exploratory

Figure 6.14 Models' diagnostics.

Exercise 6.1 (*cont.*)

Interpreting results: The maximum adjusted *R*-squared value is 0.878, with four independent variables: Owners, University, Income and Rent. Still, as Owners is not a stable predictor (see table in Section 3), it should not be included in the model. The model with the second-highest adjusted *R*-squared includes University, Income and Rent. These three variables, embedded in many different models, seem to be the best available predictors of Expenses and will be further analyzed by OLS regression in the next exercise. This OLS model does not produce different results than those presented in the preceding table, but it provides us with more statistics, graphs and capabilities for further analysis.

Exercise 6.2 OLS Regression

Before we proceed to GWR analysis we run OLS to the three variables (University, Income, Rent) selected by the exploratory regression analysis in Exercise 6.1.

ArcGIS Tools to be used: Ordinary Least Squares, Generate Spatial Weights Matrix, Spatial Autocorrelation Moran's I

ACTION: OLS regression

Navigate to the location you have stored the book dataset and click on My_Lab6_Regression.mxd

ArcToolbox > Spatial Statistic Tools > Modeling Spatial Relationships > Ordinary Least Squares

Input Feature Class = City (see Figure 6.15)

Unique ID Field = PosctCode

Output Feature Class = I:\BookLabs\Lab6\Output\OLS.shp

Dependent Variable = Expenses

Explanatory Variable = University, Income, Rent

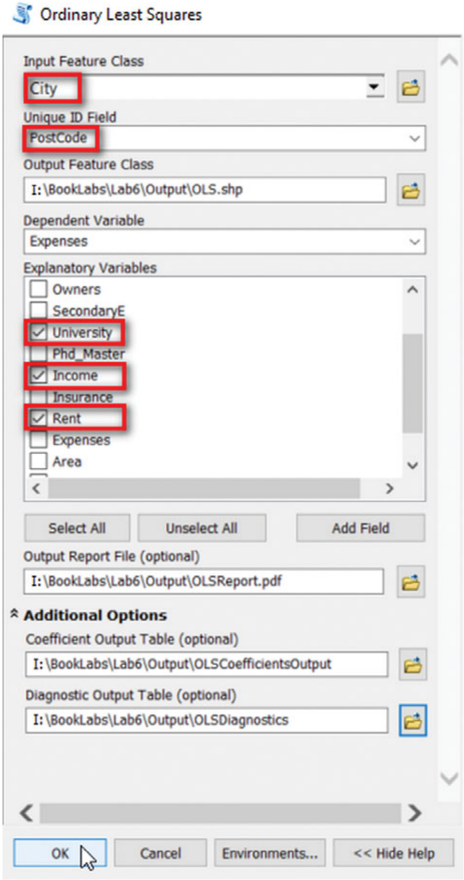
Output Report File = I:\BookLabs\Lab6\Output\OLSReport.pdf

Coefficient Output Table = I:\BookLabs\Lab6\Output\OLSCoefficientsOutput

Diagnostic Output Table = I:\BookLabs\Lab6\Output\OLSDiagnostics

OK

Exercise 6.2 (*cont.*)



Ordinary Least Squares

Input Feature Class
City

Unique ID Field
PostCode

Output Feature Class
I:\BookLabs\Lab6\Output\OLS.shp

Dependent Variable
Expenses

Explanatory Variables
- □ Owners
- □ SecondaryE
- ☑ University
- □ Phd_Master
- ☑ Income
- □ Insurance
- ☑ Rent
- □ Expenses
- □ Area

Output Report File (optional)
I:\BookLabs\Lab6\Output\OLSReport.pdf

Additional Options

Coefficient Output Table (optional)
I:\BookLabs\Lab6\Output\OLSCoefficientsOutput

Diagnostic Output Table (optional)
I:\BookLabs\Lab6\Output\OLSDiagnostics

Figure 6.15 OLS settings.

The standardized residuals are mapped automatically (Figure 6.16).

The software also provides a very informative report in PDF format, including statistics, coefficients and diagnostics tables. We briefly present the report next.

`Main Menu > Geoprocessing > Results > Current Session > Ordinary Least Squares > DC on Output Report File: OLSReport.pdf`

```
Variable   Coefficient StdError t-Statistic Probability Robust_SE Robust_t Robust_Pr  VIF
Intercept  -150.468653 15.988300 -9.411173   0.000000*  16.417585 -9.165090 0.000000*  --
UNIVERSITY    2.748526 0.617052  4.454288   0.000027*  0.990945   2.773640 0.006796* 1.664086
INCOME        0.010503 0.001089  9.646756   0.000000*  0.001486   7.067846 0.000000* 2.590905
RENT          0.165248 0.032568  5.073918   0.000003*  0.034200   4.831819 0.000007* 1.912884
```

Exercise 6.2 (*cont.*)

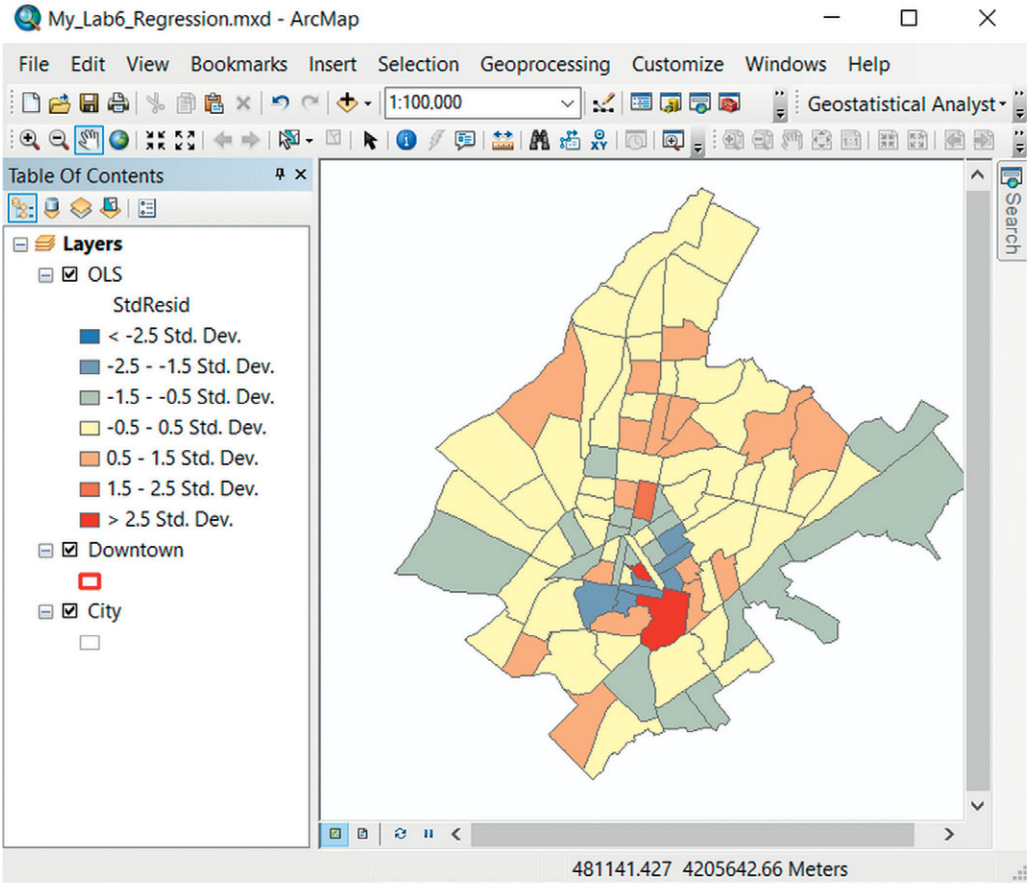


Figure 6.16 Standardized residuals.

This report presents the robust standard errors and VIF as well. Robust errors should be used if the Koenker (BP) test is statistically significant. Results are the same with the model from the exploratory regression (see Figure 6.14, second line).

```
Input Features: City
Dependent Variable: EXPENSES
Number of Observations: 90
Sum of Residual Squares: 86692.533
Akaike's Information Criterion (AICc) [d]: 884.451425
Multiple R-Squared [d]: 0.878431
Adjusted R-Squared [d]: 0.874191
```

Exercise 6.2 (*cont.*)

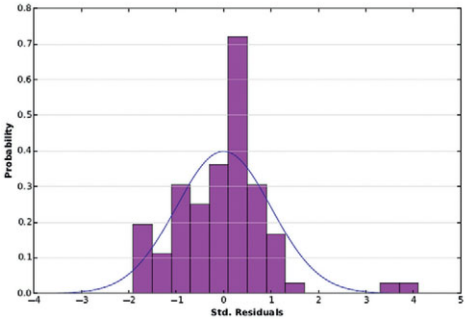


Figure 6.17 Histogram of standardized residuals.

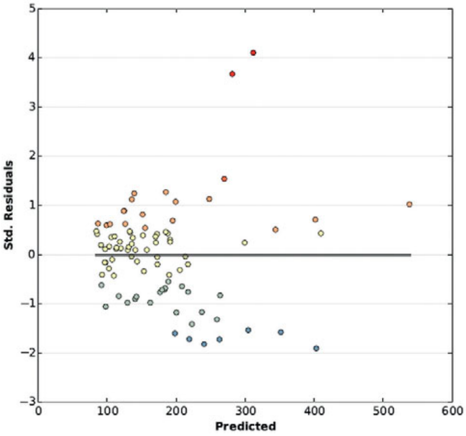


Figure 6.18 Residuals vs. predicted plot.

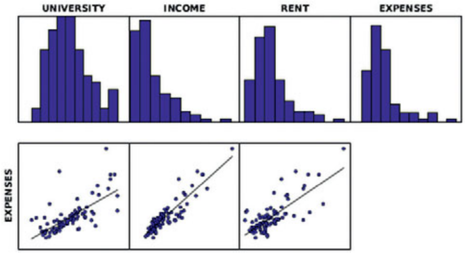


Figure 6.19 Variable distributions and relationships.

Exercise 6.2 (*cont.*)

```
Joint F-Statistic [e]: 207.139777 Prob(>F), (3,86) degrees of freedom: 0.000000*
Joint Wald Statistic [e]: 415.966846 Prob(>chi-squared), (3) degrees of freedom: 0.000000*
Koenker (BP) Statistic [f]: 13.373289 Prob(>chi-squared), (3) degrees of freedom: 0.003895*
Jarque-Bera Statistic [g]: 70.548247 Prob(>chi-squared), (2) degrees of freedom: 0.000000*
```

Interpreting results:

- First, we should check the Koenker (BP) statistic. When it is statistically significant, we should check the joint Wald statistic to determine the overall model significance and the robust probabilities to determine the coefficients significance. The Koenker (BP) statistic is statistically significant (0.003895*), indicating the existence of heterogeneity. As such, we should check the joint Wald statistic (instead of the F-significance) and the robust probabilities.
- The joint Wald statistic is statistically significant (0.000000*), and we accept the overall significance of the model with a high adjusted *R*-squared (0.874).
- In addition, all robust probabilities are statistically significant, and we accept the importance of each independent variable to the model.
- As such, the model can be written as:

```
Expenses = -150.46 + 2.75University + 0.01Income + 0.16 Rent
```

- These coefficients are not the beta coefficients, so we cannot infer which variable is more important for the model. Still, from the conceptual perspective, the model states that the more people with a university degree, of higher income and higher rent, the higher expenses for daily purchases. Following the typical analysis of the coefficients, if Income (annual income) increases by 2,000 euros and all other variables remain constant, then the mean change of Expenses (monthly expenses) will increase by 2,000X0.01 =+20 euros.
- Still, the Jarque–Berra statistic reveals violation of the normality assumption, as it is statistically significant (Jarque–Bera: 0.000000*), something that is also evident in the histogram of the standardized residuals (the blue line indicates a normal curve; see Figure 6.17).
- Moreover, the standardized residuals vs. predicted values scatter plot reveals heteroscedasticity, which increases especially for predicted values larger than 300 euros (see Figure 6.18). Ideally, residuals should exhibit no structure, shaping a random point pattern. Two outliers (more than three standard deviations) are also spotted (see the red dots in Figure 6.18). The postcodes linked to these outliers are rendered in red in Figure 6.16, indicating that the predicted value

Exercise 6.2 (*cont.*)

(fitted value) for Expenses deviated a lot from the actual one. Inspecting the robust standard errors reveals that all coefficients are robust to heteroscedasticity, as they are statistically significant.

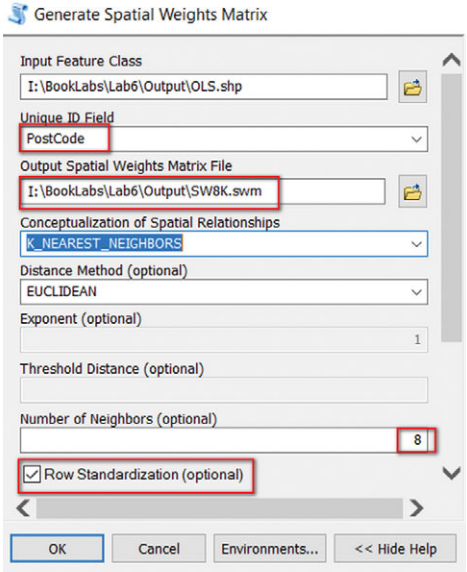
- Heteroscedasticity and nonlinearity infer bias to the model when used for predictions. To deal with heteroscedasticity and nonlinearity, we can inspect the distributions of the independent to the dependent variables (see Figure 6.19).
- The histograms and scatter plots among Expenses and the other independent variables show in general linear patterns. Still, in cases of strongly skewed variables, we may apply a transformation. Due to the significant extent of such analysis, we do not apply any transformation in this example, but you are encouraged to transform income and investigate the outcomes (keep in mind that transformations most of the time are not easily interpreted and, for this reason, many researchers avoid them).
- Finally, we should check if residual errors are spatially autocorrelated.

ACTION: Spatial Autocorrelation Moran's I

OLS does not automatically calculate spatial autocorrelation for residuals. Before applying Global Moran's I, we have to create the weights matrix. The spatial autocorrelation test used in the exploratory regression analysis in Exercise 6.1 chose by default the eight nearest neighbors to build the weights matrix. Still, this option is not included in the Moran's I tool, so we have to build the weights matrix first (see Box 6.3).

```
ArcToolBox > Spatial Statistics Tools > Modeling Spatial
Relationships > Generate Spatial Weights Matrix

Input Feature Class = I:\BookLabs\Lab6\Output\OLS.shp

Unique ID Field = PostCode (see Figure 6.20)

Output Spatial Weights Matrix File =

I:\BookLabs\Lab6\Output\SW8K.swm

Conceptualization of Spatial Relationships =

K_NEAREST_NEIGHBORS

Distance Method = EUCLIDEAN

Number of Neighbors = 8

Row Standardization = Check

OK
```

Exercise 6.2 (*cont.*)



Generate Spatial Weights Matrix

Input Feature Class: I:\BookLabs\Lab6\Output\OLS.shp

Unique ID Field: PostCode

Output Spatial Weights Matrix File: I:\BookLabs\Lab6\Output\SW8K.swm

Conceptualization of Spatial Relationships: K_NEAREST_NEIGHBORS

Distance Method (optional): EUCLIDEAN

Exponent (optional): 1

Threshold Distance (optional):

Number of Neighbors (optional): 8

☑ Row Standardization (optional)

Figure 6.20 Spatial weights matrix dialog box.

Continue by calculating Moran's I.

ArcToolbox > Spatial Statistic Tools > Analyzing Patterns > Spatial Autocorrelation (Morans I)

Input Feature Class = OLS (see Figure 6.21)

Input Field = Residual

Generate Report = Check

Conceptualization of Spatial Relationships =

GET_SPATIAL_WEIGHTS_FROM_FILE

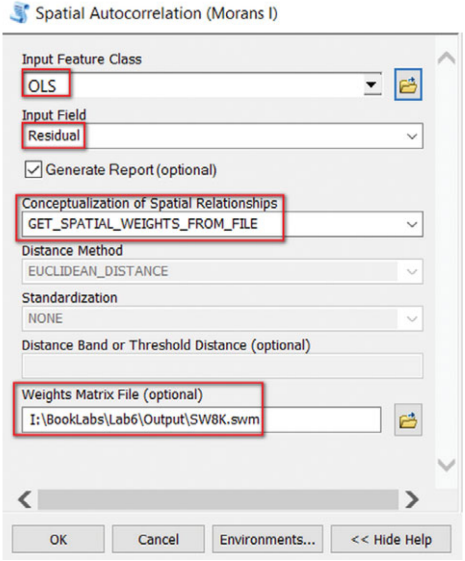
Weights Matrix File = I:\BookLabs\Lab6\Output\SW8K.swm

OK

Main Menu > Geoprocessing > Results > Current Session > Spatial Autocorrelation > DC Report File: MoransI_Results.html

Main Menu > File > Save

Exercise 6.2 (*cont.*)



Spatial Autocorrelation (Morans I)

Input Feature Class
OLS

Input Field
Residual

☑ Generate Report (optional)

Conceptualization of Spatial Relationships
GET_SPATIAL_WEIGHTS_FROM_FILE

Distance Method
EUCLIDEAN_DISTANCE

Standardization
NONE

Distance Band or Threshold Distance (optional)

Weights Matrix File (optional)
I:\BookLabs\Lab6\Output\SW8K.swm

Figure 6.21 Spatial autocorrelation (Moran's I).

The *p*-value is 0.26453, indicating no spatial autocorrelation (same as the results of exploratory regression).

Box 6.3 Keep in mind that for the same problem, a different conceptualization method (e.g., Queen's contiguity) may lead to contradicting results (e.g., spatial autocorrelation existence instead of nonexistence), showcasing how a single choice might change results significantly. Each parameter choice should be made after thorough investigation. Conceptualization of spatial relationships presented in Chapter 1 is very important in the output of many spatial statistics tools. We can run exploratory regression with various conceptualization methods to inspect how results are differentiated. Analysis is not a straight road, and we may have to take loops and start once again when new evidence comes to light. For practice, you are encouraged to run exploratory regression again with the Queen's contiguity method for calculating the weights matrix file.

Interpreting results: The overall interpretation for OLS model is presented in Table 6.4:

$$\texttt{Expenses} = -150.46 + 2.75\texttt{University} + 0.01\texttt{Income} + 0.16\ \texttt{Rent}$$

Exercise 6.2 (*cont.*)

Table 6.4 Regression summary results.

Value		Significant	Interpretation	Remarks	Evaluation
Model Overall significance/ performance					
R-Squared	0.878	Not applicable (N/A)	The model explains 87.8% of Expenses variation.	Adjusted *R*-squared should also be checked to avoid model overfitting. Adjusted *R*-squared is preferred, especially when their difference is large.	☑
Adjusted *R*-Squared	0.874	(N/A)	Adjusted *R*-squared is 0.874, indicating high goodness of fit.	In cases of low goodness of fit, we should add extra variables and/or test other regression methods (e.g., spatial lag model).	☑
F-Statistic (Also called Joint F-Statistic	207.139	Significant at the 0.01 significance level	We reject the null hypothesis and accept that the regression model is statistically significant. The trend is a real effect, not due to random sampling of the population.	F-statistic is of minor importance as most of the times is significant. It should be cross-checked with joint Wald statistic if Koenker (KB) test is statistically significant.	☑
Joint Wald Statistic	415.966	Significant at the 0.01 significance level	As the joint Wald test is statistically significant, it is an indication of overall model significance.	Joint Wald statistic is used to analyze the overall performance of the model especially when Koenker BP statistic is statistically significant ($p < 0.05$).	☑
Akaike Info Criterion corrected (AICc)	884.451	N/A	It is used for comparison with the spatial model. AICc indicates better fit when values decrease.	AICc is a measure of the quality of models for the same set of data. It is used to compare different models and select the most appropriate.	
Coefficients					
t-statistic	Uni/sity: 0.000027* Income: 0.000000* Rent: 0.000003*	Variables significant at the 0.01 significance level	All variables significant	As Koenker BP statistic is statistically significant, robust probabilities should be checked instead.	☑
White standard errors (Robust probabilites)	Uni/sity: 0.006796* Income: 0.000000* Rent: 0.000007*	Variables significant at the 0.01 significance level	All variables significant	Regression coefficients remain the same as in the OLS model. Standard errors and *p*-values change. Coefficients' values significant in OLS might be different after the white standard calculation. The new coefficients that are significant are robust to heteroscedasticity.	☑
Multicollinearity					
Variance Inflator Factor	Uni/sity: 1.664086 Income: 2.590905 Rent: 1.912884	VIF = 1–4 No collinearity VIF = 4–10 Further analysis needed VIF > 10 Severe collinearity	All variables have values smaller than 4. No multicollinearity issues.		☑

Exercise 6.2 (*cont.*)

Table 6.4 (*cont.*)

Value		Significant	Interpretation	Remarks	Evaluation
Normality of residual errors					
Jarque-Bera test	70.548	Significant at the 0.01 significance level	We reject the null hypothesis (errors normality) as the value of the test is statistically significant.	Biased model	☒ Violating errors normality assumption
Heteroscedasticity					
Koenker-Bassett test	13.373	Significant at the 0.01 significance level	We reject the null hypothesis (homoscedasticity) as the value of the test is statistically significant.	When the test is significant, relationships modeled through coefficients are not consistent. We should look at standard errors probabilities (White, HAC) to determine coefficients' significance. Large difference between the values of Breusch-Pagan test and Koenker (KB) test also indicates violation of the errors normality assumption.	☒ Violating the assumption of constant variance of errors. Still, standard errors are statistically significant.
Spatial Dependence					
Moran's I	0.26	Not significant at the 0.01 significance level	There is not sufficient evidence to reject the null hypothesis (no spatial autocorrelation).		☑
Overall conclusion	Model suffers from non-errors normality and heteroscedasticity				
Next step	Try a GWR model				

Although the OLS performs moderately well, it would be reasonable to attempt improving the reliability of the predictions from the models by using GWR. GWR would also allow us to map the coefficient estimates and to examine further whether the process is spatially heterogeneous.

Exercise 6.3 GWR

In this exercise, we model Expenses as a function of University, Income and Rent using GWR. With GWR, a local model is produced by fitting a regression equation to every single spatial feature in the dataset.

Before we run the tool, we show how ArcGIS presents the results of GWR. For more details on this section, you can visit ESRI's website (ESRI 2016b).

Exercise 6.3 (*cont.*)

GWR output: GWR as generated in ArcGIS yields three main outputs: (A) common diagnostics, (B) output feature class (shapefile) containing all local estimates and (C) coefficient raster surfaces mapping each coefficient separately. In more detail,

(A) **Common diagnostics:** This table describes the model variables and the diagnostic results. It is used for comparison reasons with the global OLS model (GWR is the local model). It contains the following:

- **Bandwidth or the number of neighbors:** This is the distance or the number of neighbors used in the spatial kernel adopted for each local estimation. It controls the degree of smoothing. The larger the values and the smoother the model, the more global the results are. The smaller the number, the more local the results are.
- **Residual squares:** It is the sum of the squared residuals of the model. The smaller the value, the better the fit of the GWR model.
- **Effective number of parameters:** It is a measure of the complexity of the model and is also used in the calculation of other diagnostics. It is equivalent to the number of parameters (intercept and coefficients) in OLS. The value of the effective number of parameters is a trade-off between the variance of the model and the bias in the coefficient estimates, and is heavily dependent on the bandwidth. Suppose that the global model (OLS) has *k* parameters for *n* data points (observations/locations). If the bandwidth value tends to infinity, then the local model tends to be global and the number of parameters tend to *k*. If the bandwidth value tends to zero, then the local models are based only on the regression points, as weights tend to zero as well. In this case, the parameters are *n* (as many as the data points). The effective number of parameters is a number ranging from *k* to *n* and might not be an integer. Additionally:
  - ➢ When the effective number of parameters tends to *k* (global model), then the weights tend to 1 and the local coefficient estimates have small variance but are biased.
  - ➢ When the effective number of parameters tends to *n* (data points), then the weights tend to zero and the local coefficient estimates have large variance but are not that biased.
- **Sigma** is the square root of the normalized residual sum of squares. It is the estimated standard deviation of the residuals. As in the case of residuals, the smaller the value, the better the model. As sigma is used to calculate AICc, we may only analyze AICc in our results.

Exercise 6.3 (*cont.*)

- **AIC corrected (AICc)** is a measure of the quality of models for the same set of data based on the AIC criterion. It is used to compare different models (i.e. GWR vs. OLS) in order to select the most appropriate one. The comparison is relative to those models available, which are not necessarily the most accurate. As a rule of thumb, if the AICc value of two models differs by more than 3, then the one with the smallest AICc is better. For small samples or for samples with a relatively large number of parameters, AIC corrected is advised to be used instead of uncorrected AIC.
- ***R*-squared** is a measure of goodness of fit indicating the explained variation of the dependent variable.
- **Adjusted *R*-square:** is a measure of goodness of fit that is more reliable than *R*-squared.

(B) **Output feature class** (shapefile): It is used primarily for mapping standardized residuals. It contains:

- **Condition number:** The condition number (field COND) reveals if local multicollinearity exists. As a rule of thumb, condition values larger than 30, equal to null or negative reveal multicollinearity, and results should be examined with caution. A high condition number reveals high sensitivity in the regression equation for small changes in the variables' coefficients. Problems of local multicollinearity for independent variables might exist when values (whether ratio, nominal or categorical) cluster geographically. For this reason, categorical variables with few categories, dummy or regime variables should not be used.
- **Local *R*-squared:** It is the *R*-squared value for the local model and is interpreted as a measure of goodness of fit like the global *R*-squared. By mapping *R*-squared, we observe where GWR model predicts well and where it predicts poorly. Local *R*-squared can be used to identify potential model misspecification.
- **Predicted values:** The estimated fitted values of the dependent variable.
- **Coefficients:** Coefficients and intercepts for each location.
- **Residual values:** The subtraction of the fitted values from the observed values (dependent variable).
- **Coefficient standard errors:** When standard errors are small in relation to absolute values, then coefficient values are more reliable. Large standard errors may reveal local multicollinearity issues.

Exercise 6.3 (*cont.*)

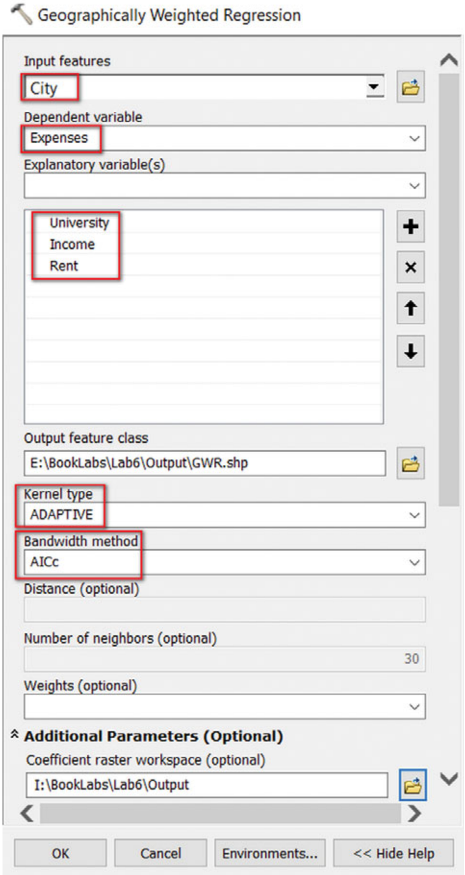
(C) **Coefficient raster surfaces:** A raster surface per coefficient is created. It shows the variation in the coefficient estimation for each independent variable.

ArcGIS Tools to be used: Geographically Weighted Regression, Spatial Autocorrelation Moran's I, Generate Spatial Weights Matrix, Create Graph

ACTION: GWR regression
Navigate to the location you have stored the book dataset and click on My_Lab6_Regression.mxd.

ArcToolbox > Spatial Statistic Tools > Modeling Spatial Relationships > Geographically Weighted Regression

Input Feature Class = City (see Figure 6.22)



Geographically Weighted Regression

Input features: City

Dependent variable: Expenses

Explanatory variable(s):
University
Income
Rent

Output feature class: E:\BookLabs\Lab6\Output\GWR.shp

Kernel type: ADAPTIVE

Bandwidth method: AICc

Distance (optional):

Number of neighbors (optional): 30

Weights (optional):

Additional Parameters (Optional)

Coefficient raster workspace (optional): I:\BookLabs\Lab6\Output

Figure 6.22 GWR dialog box.

Exercise 6.3 (*cont.*)

```
Dependent Variable = Expenses

Explanatory Variable = University, Income, Rent

Output Feature Class = I:\BookLabs\Lab6\Output\GWR.shp

Kernel Type = ADAPTIVE

Bandwidth method = AICc

Weights = Leave blank

Coefficient raster workspace = I:\BookLabs\Lab6\Output\ > Add

Leave all other fields blank or as filled by default

OK
```

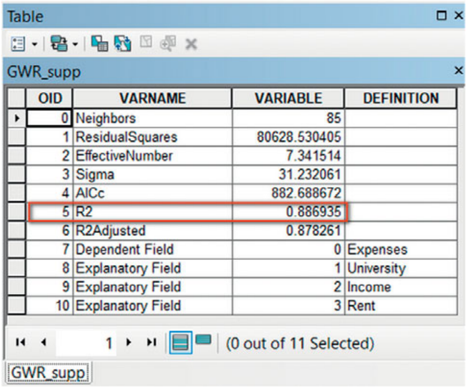
(A) Common Diagnostics (Supplementary Table)

ACTION: Add GWR_supp to the TOC

```
ArcCatalog > Navigate to I:\BookLabs\Lab6\Output\ > Drag and drop GWR_supp.dbf to the TOC

TOC > RC GWR_supp > Open
```

We begin with analyzing the supplementary table GWR_supp.dbf, which presents the primary results of the GWR model (see Figure 6.23).

Interpreting results: The base for analyzing GWR output (see Figure 6.23) is the OLS results, which are referred to as the global model as it provides a single model for the whole study area. GWR results along with a comparison with the global OLS model presented in Exercise 6.2 are listed and interpreted in Table 6.5.



OID	VARNAME	VARIABLE	DEFINITION
0	Neighbors	85	
1	ResidualSquares	80628.530405	
2	EffectiveNumber	7.341514	
3	Sigma	31.232061	
4	AICc	882.688672	
5	R2	0.886935	
6	R2Adjusted	0.878261	
7	Dependent Field	0	Expenses
8	Explanatory Field	1	University
9	Explanatory Field	2	Income
10	Explanatory Field	3	Rent

1 (0 out of 11 Selected)

GWR_supp

Figure 6.23 Common diagnostics of GWR.

Exercise 6.3 *(cont.)*

Table 6.5 Model overall significance/performance for the local GWR and OLS model

	GWR (local) Value	OLS (global) Value	Interpretation	Evaluation
Bandwidth or number of neighbors	85 nearest neighbors	No direct comparison. OLS uses all spatial objects (90).	85 out of 90 spatial objects are used for each local estimation. Under each kernel, about 94% of the data are used, which is quite large. The model tends rather to a global one.	
Sum of residual squares	80628.530	86692.533	The local model has a smaller sum of residual squares than the global model, indicating better performance.	☑
Effective Number	7.34	4	In this dataset, the effective value could vary between 4 and 90. The value of 7.34 is very close to 4, indicating that as the bandwidth value tends to 90, the local model tends to be global, and the number of parameters tends to 4.	
Akaike Info Criterion corrected (AICc)	882.688	884.451	Their value difference is smaller than 3, indicating that there is not any substantial improvement in the GWR model.	☒
R-Squared	0.886	0.878	Value is practically the same with the global model. The	☑

Exercise 6.3 (*cont.*)

Table 6.5 (*cont.*)

	GWR (local) Value	OLS (global) Value	Interpretation	Evaluation
Adjusted *R*-Squared	0.878	0.874	model explains 88.6% of expenses' variation. Value is practically the same with the global model. There is high goodness of fit.	☑
Overall conclusion	Local GWR model did not significantly improve the global OLS model. GWR is better applied when hundreds of spatial units exist. For the educational needs of this example, we use 90, which probably leads to a poor GWR model.			
Next Step	Continue by mapping residuals and coefficients.			

(B) Mapping Standardized Residuals (Output Feature Class) and Calculating Spatial Autocorrelation: Over- and Underpredictions

Interpreting results: The standardized residuals reveal how much of the spatial data variation is not explained by the independent variables. From the map inspection, we observe that overpredictions (blue) and underpredictions (red) are located around the downtown area, engulfing the historic and financial center (see Figure 6.24). To better analyze residuals, we plot them using a scatter plot. (see Figure 6.26).

ACTION: Create residual plot

Main Menu > View > Graphs > Create Graphs

Graph type = Scatter Plot (see Figure 6.25)

Layer/Table = GWR

Y field = StdResid

X field = Predicted

Uncheck Add to legend > Next > Title = Standardized Residual plot > Finish > RC on graph > Add to layout

Exercise 6.3 (*cont.*)

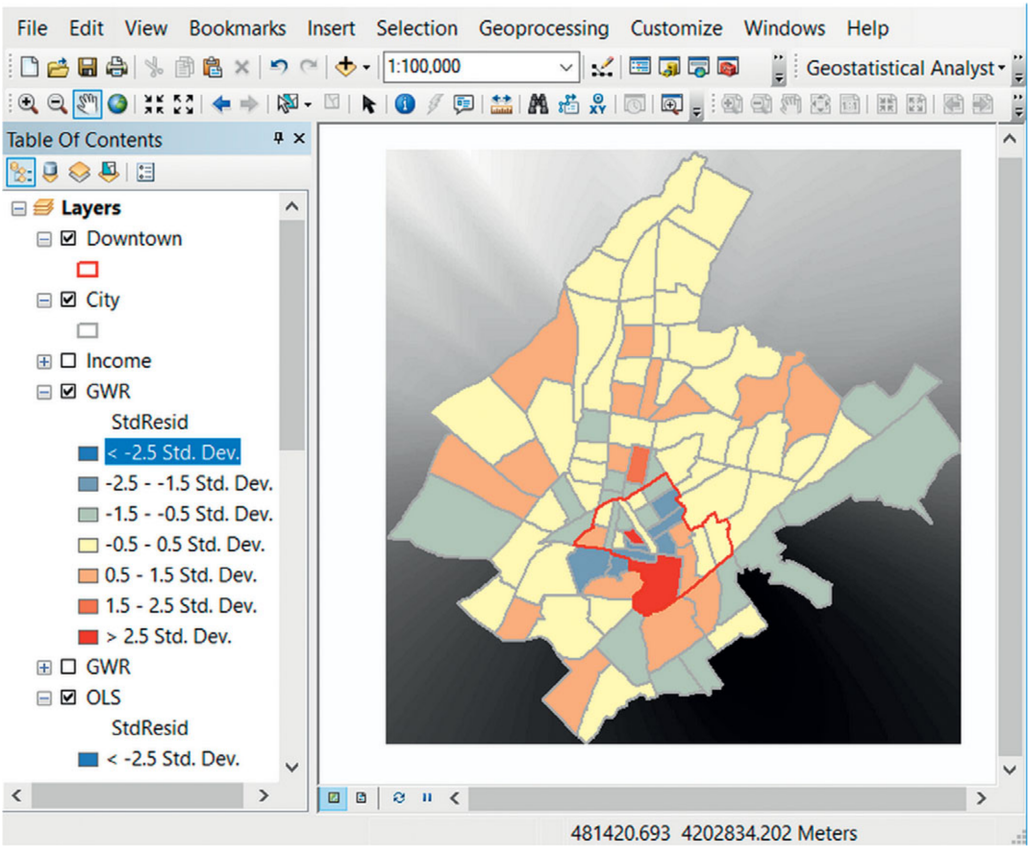


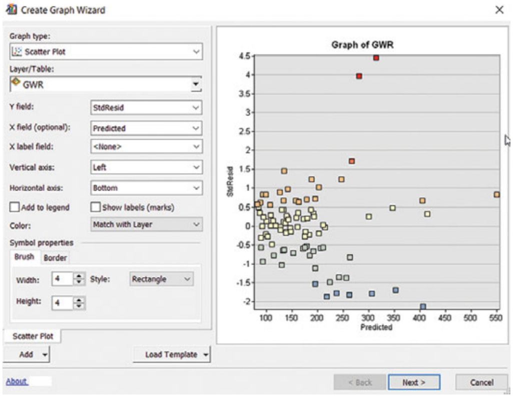
Figure 6.24 Standardized residuals mapping.

```
Get back to the Data View > Brush the dots in the upper-right
corner
```

```
Main Menu > Save
```

Interpreting results: In general, the scatter plot reveals heteroscedasticit,y although the majority of residuals seem to have a random pattern (see Figure 6.26). There are two residuals (center up – underprediction), and one residual (right down – overprediction) that deviate a lot from the general pattern (brush to see the related postcodes). These extreme residuals lie inside the historic center and the financial district (downtown) and should be closely inspected to discover the potential reasons for these significant differences from the observed values.

Exercise 6.3 (*cont.*)



Create Graph Wizard

Graph type: Scatter Plot

Layer/Table: GWR

Y field: StdResid

X field (optional): Predicted

X label field: <None>

Vertical axis: Left

Horizontal axis: Bottom

□ Add to legend

□ Show labels (marks)

Color: Match with Layer

Symbol properties

Brush | Border

Width: 4 Style: Rectangle

Height: 4

Figure 6.25 Creating scatter plot.

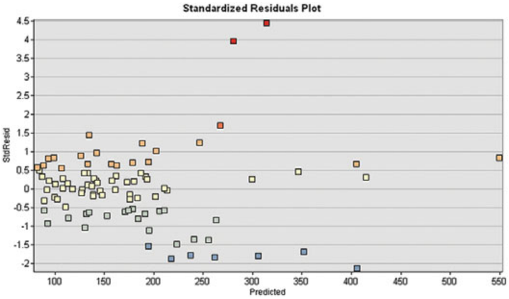


Figure 6.26 Standardized residual plot.

Spatial Autocorrelation Global Moran's I

Continue by calculating Moran's I (see Box 6.4).

ACTION: Spatial Autocorrelation (Morans I)

Box 6.4 Before running the tool, calculate again the weights matrix to ensure its smooth operation.

```
ArcToolBox > Spatial Statistics Tools > Modeling Spatial

Relationships > Generate Spatial Weights

Matrix Input Feature Class = GWR.shp

Unique ID Field = Source_ID
```

Exercise 6.3 (*cont.*)

Box 6.4 (*cont.*)

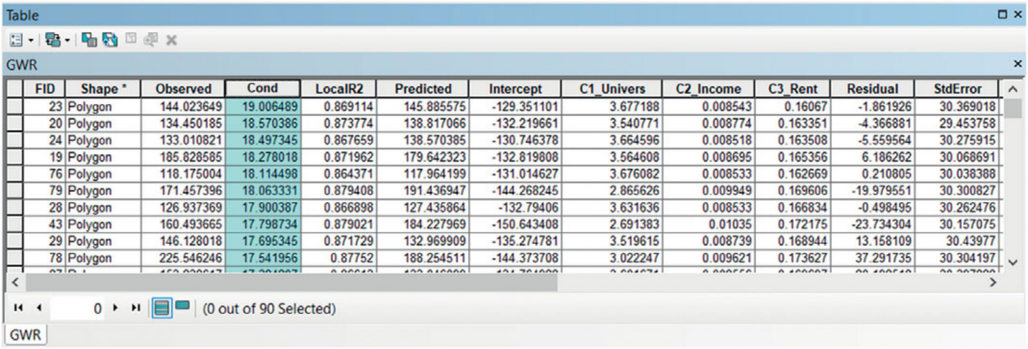
```
Output Spatial Weights Matrix File =
I:\BookLabs\Lab6\Output\SW8KGWR.swm
Conceptualization of Spatial Relationships = K_NEAREST_NEIGHBORS
Distance Method = EUCLIDEAN
Number of Neighbors = 8
Row Standardization = Check
OK
```

```
ArcToolbox > Spatial Statistic Tools > Analyzing Patterns > Spatial Autocorrelation Moran's I
Input Feature Class = GWR
Input Field = Residual
Generate Report = Check
Conceptualization of Spatial Relationships =
GET_SPATIAL_WEIGHTS_FROM_FILE
Weights Matrix File = I:\BookLabs\Lab6\Output\SW8KGWR.swm
OK
```

To view the results of the test:

```
Main Menu > Geoprocessing > Results > Current Session > Spatial Autocorrelation > DC Report File: MoransI_Results0.html
Global Moran's I Summary
Moran's I Index:-0.011492
Expected Index:-0.011236
z-score:-005704
p-value:0.995449
Save
TOC > RC GWR > Open Attribute Table > RC Cond > Sort Descending
Close table
```

Exercise 6.3 (*cont.*)



Table

GWR

FID	Shape *	Observed	Cond	LocalR2	Predicted	Intercept	C1_Univers	C2_Income	C3_Rent	Residual	StdError
23	Polygon	144.023649	19.006489	0.869114	145.885575	-129.351101	3.677188	0.008543	0.16067	-1.861926	30.369018
20	Polygon	134.450185	18.570386	0.873774	138.817066	-132.219661	3.540771	0.008774	0.163351	-4.366881	29.453758
24	Polygon	133.010821	18.497345	0.867659	138.570385	-130.746378	3.664596	0.008518	0.163508	-5.559564	30.275915
19	Polygon	185.828585	18.278018	0.871962	179.642323	-132.819808	3.564608	0.008695	0.165356	6.186262	30.068691
76	Polygon	118.175004	18.114498	0.864371	117.964199	-131.014627	3.676082	0.008533	0.162669	0.210805	30.038388
79	Polygon	171.457396	18.063331	0.879408	191.436947	-144.268245	2.865626	0.009949	0.169606	-19.979551	30.300827
28	Polygon	126.937369	17.900387	0.866898	127.435864	-132.79406	3.631636	0.008533	0.166834	-0.498495	30.262476
43	Polygon	160.493665	17.798734	0.879021	184.227969	-150.643408	2.691383	0.01035	0.172175	-23.734304	30.157075
29	Polygon	146.128018	17.695345	0.871729	132.969909	-135.274781	3.519615	0.008739	0.168944	13.158109	30.43977
78	Polygon	225.546246	17.541956	0.87752	188.254511	-144.373708	3.022247	0.009621	0.173627	37.291735	30.304197

0 (0 out of 90 Selected)

GWR

Figure 6.27 Inspecting Cond field.

Interpreting results: Given the z-score of −0.00570, the pattern does not appear to be significantly different than random. This is a sign of an effective GWR model (OLS lead to the same conclusion as well). In cases where Moran's I does not run due to an error in weights matrix, save the project, exit and run again.

By opening the attribute table of the output feature class, we observe no issues of local multicollinearity, as no value (Cond field in `GWR.shp` attribute table) is larger than 30, a threshold value to consider that multicollinearity exists (current values are less than 19; see Figure 6.27). Standard errors of coefficients are relatively small, something expected when effective number is small (close to OLS parameters). Local *R*-squared is high in every single model, with the lowest being 0.85.

(C) Mapping Local Coefficients (Coefficient Raster Surfaces)

We do not analyze every single coefficient, as this would require much space. We mostly focus on how to assess raster coefficients to trace potentially hidden underlying processes. In cases where GWR is not effective due to spatial autocorrelation, mapping coefficients remains informative. For example, deviations or trends of coefficients might reveal valuable information that will help us to build a better model in a later stage.

```
ACTION: Raster surface

TOC > RC Income (raster layer) > Properties > TAB = Symbology
Color Ramp = Blue to Red
OK
TOC > Move Income on the top before Downtown and City layers
(List By Drawing Order)
```

Exercise 6.3 (*cont.*)

We can also map coefficients per postcode by just using the symbology menu under the properties of the GWR.shp. We should first save the GWR residuals scheme as a new layer.

```
TOC > RC GWR > Save As Layer File >
I:\BookLabs\Lab6\Output\GWRStResid.lyr
Add layer to the TOC
TOC > RC GWR > Properties > TAB = Symbology > Quantities >
Graduated colors
Value = C2_Income
Color Ramp = Blue to Red > OK
Main Menu > File > Save
```

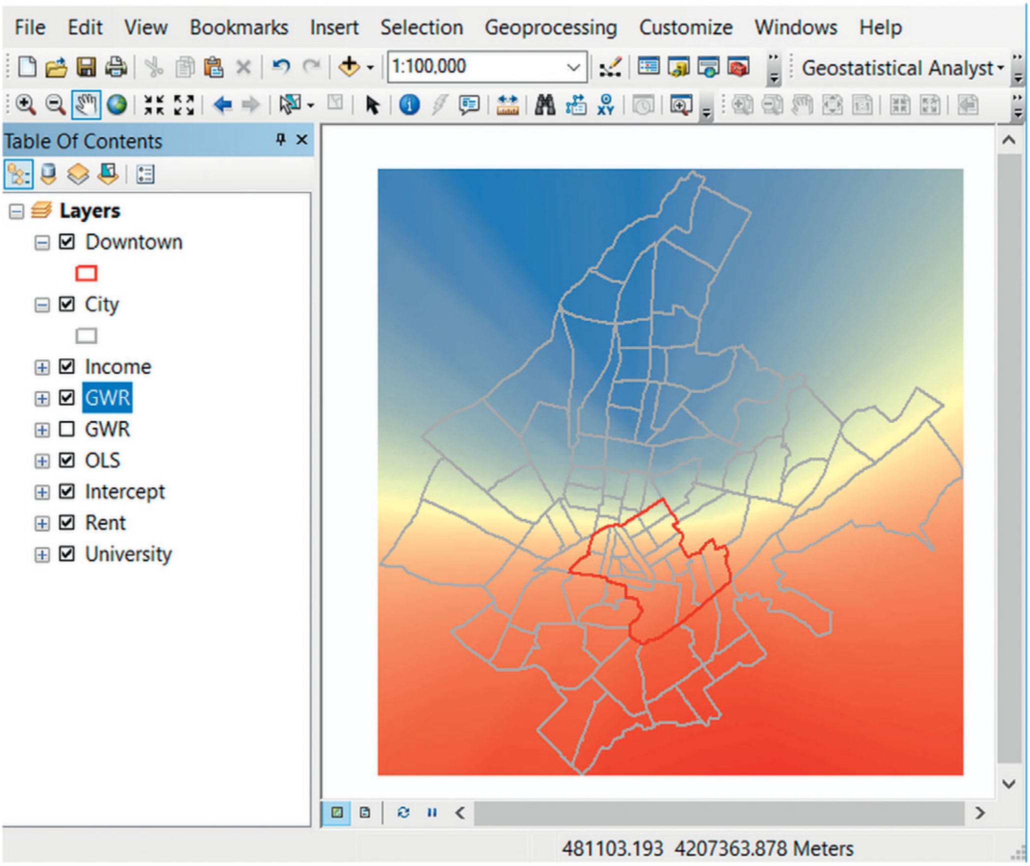


Figure 6.28 Mapping local coefficients of Income as a raster surface.

Exercise 6.3 (*cont.*)

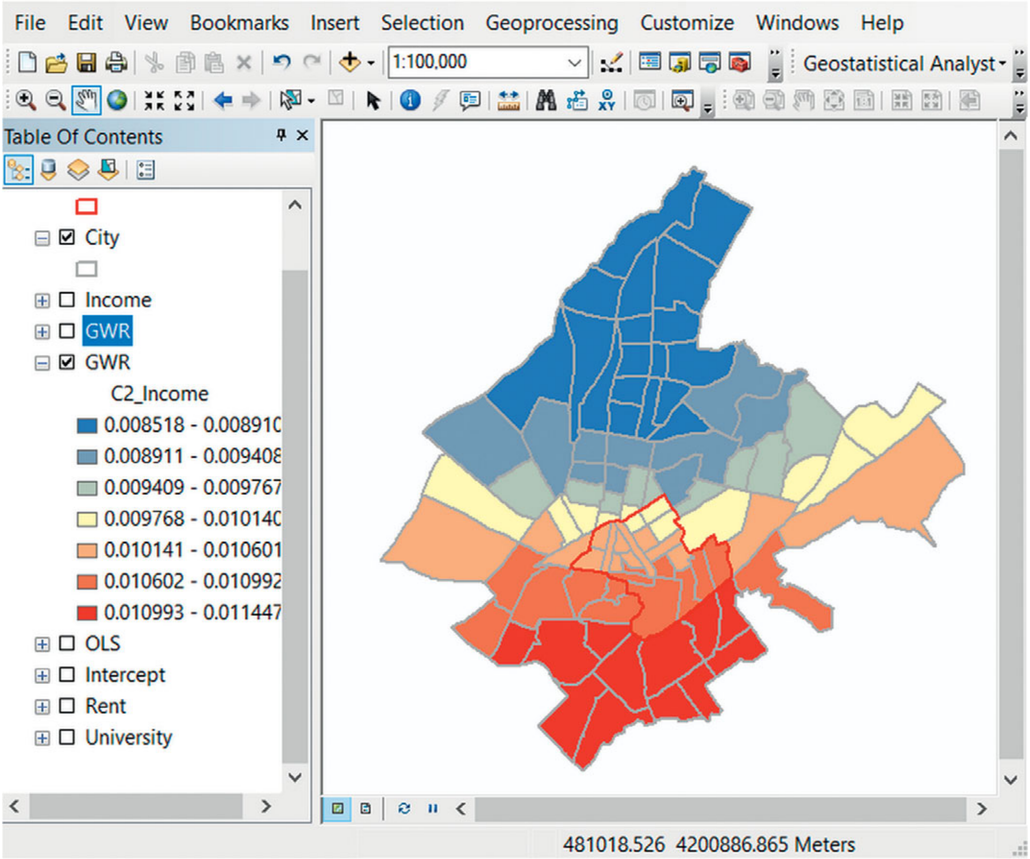


Figure 6.29 Mapping local coefficients of Income per spatial unit.

Interpreting results: Mapping `Income` coefficient estimates of the local model reveals a trend from north to south, with intensity of coefficients particularly starting to increase at the historical and financial center (see Figures 6.28 and 6.29). Moreover, coefficient values vary significantly from place to place (0.008 to 0.01). From the policy perspective, as the map for the local coefficients reveals that the influence of `Income` varies considerably across the study area (spatial heterogeneity presence), local policies seem more appropriate than regional policies. The global coefficient and the local coefficients are all positive for `Income`, so there is an agreement between the local and the global model on how this independent variable influences the dependent `Expenses`.

Exercise 6.3 (*cont.*)

As an overall comment, GWR performs relatively better than the global OLS model, providing us with rich mapping capabilities that better assist in identifying varying spatial relationships. Spatial autocorrelation is not present in residuals, and the model exhibits high goodness of fit and no multicollinearity issues. By mapping coefficients estimates, trends are traced, and valuable information may be extracted. GWR performs better with larger datasets (more than 160 spatial units) than those in the current layer.