**Chapter 11**

**Regression**

with Amr ElRafey

# [H1] Learning Objectives

1. Define different types of regression
2. Fit data to estimate equation of line
3. Transform nonlinear data
4. Choose a line that minimizes sum of square of residuals
5. Fit a nonlinear equation to data to examine interactions among variables
6. Interpret coefficient of determination
7. Check the assumption of correctness of the model form
8. Check assumptions of the independence of the error terms
9. Check the assumption of a stable variance over time
10. Check the assumption of normally distributed errors
11. Interpret effects of collinearity on model parameters
12. Cross-validate regression predictions
13. Prepare weighted regressions
14. Reduce the number of variables in analysis by ignoring irrelevant ones
15. Test hypotheses in subset of data

# [H1] Key Concepts

* Equation of a line
* Slope and intercept in an equation
* Multiple regression
* Linear model
* Independent variables
* Dependent variables
* Interaction among variables
* Main effect
* Collinearity
* Coefficient of determination
* Regression parameters
* Context-specific test of hypothesis
* Predicted values
* Q-Q plot
* Independence of errors
* Homoscedasticity
* Normal error
* Transformation of data
* Cross-validation

# [H1] Chapter at a Glance

This chapter covers regression, a widely used method of relating multiple variables to an outcome. Regression is used to fit the data to a mathematical equation that relates a dependent variable to one or more other variables. In many statistics books, regression is introduced by deriving its mathematical proofs or by using matrix algebra. In keeping with our focus on applications, we will not use this method. Instead, we introduce the basic ideas and assumptions and the interpretation of the findings.

There are some surprises not regularly seen in a chapter on regression. First, we discuss how the tests of parameters of regression are different from hypothesis tests discussed in chapter 6. The test of parameters in regression is a context-specific test that could lead to conclusions very different from those of methods presented in the chapter on means and rates. Second, we spend time discussing interactions among variables, something that we will repeatedly come back to in other chapters.

# [H1] Regression Is Everywhere

Regression is a general method that relates an outcome variable *Y* to one or more independent variables *X*1 … *X*n using an equation. Equations are important. They show, with precision, how variables affect each other. They are the language of science and allow scientists to express their theories. In management, equations can be used to quantify relationships among variables, forecast future events, set priorities for interventions, test hypotheses on the effect of a variable on outcomes, and for many other applications. The following are some examples of how regression (and the related equation) is used in healthcare management:

**[INSERT BL]**

* Regression can be used to predict future stock market values. For example, Maskawa (2016) used regression to model the buy and sell order for stocks using information about other participants in the market.
* Regression can be used to assess the impact of health promotion on recruitment of new patients. For example, Shafer and colleagues (2016) evaluated the impact of a media campaign aired on national cable television networks, radio, and other channels of communication, with supporting digital advertising to drive traffic to a website. Linear regression models were used to estimate the relationship between the components of the media campaign and traffic.
* Regression can be used to predict the likelihood of a sale or market penetration given the characteristics of the population in the zip code area. For example, Shimada and colleagues (2009) used regression to estimate enrollment in a Medicare managed-care program using patient and market characteristics.
* Regression can be used to predict the likelihood of project success based on its characteristics. For example, Gustafson and colleagues (2003) used project characteristics from 198 healthcare organizations to predict the project’s likelihood of success.
* Regression can be used to assess financial risk from historical patterns. For example, Ryan and colleagues (2015) used regression models to estimate physician compensation as a function of participation in accountable care organizations, risk for primary care costs, and other practice characteristics.
* Regression can be used to assess the likelihood of bad debt from patient characteristics. For example, Clyde, Hemenway, and Nagurney (1996) used regression to examine the association between a number of factors—including seat belt use, insurance status, admission to the hospital, driver or passenger status, age, and sex—and bad debt.
* Regression can be used to check for fraud by comparing the amount charged to the amount expected. For example, Chen, Goo, and Shen (2014) used regression to predict, on the basis of financial and nonfinancial variables, the probability of fraudulent financial statements.
* Regression can be used to assess the effectiveness of medical interventions. For example, Kheirbek Alemi, and Zargoush (2013) used regression to assess the comparative effectiveness of antidiabetic medications.
* Regression can be used to identify the relevant components of care that lead to cost overruns. For example, Carey and Stefos (2011) predicted excess cost of hospital inpatient care stemming from adverse safety events in US Department of Veterans Affairs hospitals.
* Regression can be used to summarize the impact of different surgeons on the overall cost of bundled care. For example, Yount and colleagues (2015) used regression to adjust for the cost of cardiac surgical bundles.
* Regression can be used to identify relevant risk factors that increase patients’ risk of illness, patients’ risk of compliance with treatment, or the likelihood of adverse outcomes. For example, Boucquemont and colleagues (2014) used regression to understand risk factors for kidney disease outcomes.

**[END BL]**

Because there are many ways to use equations in healthcare management, it is not surprising that regression is used often.

# [H1] Types of Regression

Depending on the type of the independent variables, there are four varieties of regressions. These include the following:

**[INSERT NL]**

1. *Ordinary* or *standard* regression, where the dependent variable is a continuous variable. Ordinary regression estimates the mean of the dependent variable as a function of the independent variables.
2. *Logistic* regression, where the dependent variable is a binary variable. In these regressions, the log odds of the dependent variable are estimated from independent variables. Logistic regression is discussed in chapter 12.
3. *Cox’s hazards* regression, where the dependent variable is time (e.g., number of days) until an event. This type of regression is not discussed in this book.
4. *Poisson* regression, where the dependent variable is the count of co-occurrences of the independent variables. We discuss this type of regression in chapter 19.  
   **[END NL]**

Chapter 11 focuses on ordinary regression. When there is more than one predictor, the regression may also be called *multiple* regression.

# [H1] Introduction to Equations

To understand regression, we can start with various equations used to relate the observed values of variables to a dependent variable, such as health outcome or cost. The equation for a straight line is referred to as linear, and it is written as

**[INSERT EQUATION]**

.

**[END EQUATION]**

Where and are constants (fixed numbers), *X* is referred to as an *independent variable* because it can change at will. *Y* is referred to as *dependent variable* because it changes as a function of *X*, although these naming procedures can be switched and we can write *X* as a function of *Y*, in which case the dependent and independent variable designation are switched. If we were to graph the line, is the intersection of the line with the *y*-axis (i.e., it is the value of *Y* when *X* is zero). The constant is the slope of the line; it shows how much *y* increases with one unit of increase in . Increase in is called *run*. The corresponding change in *Y* is called *rise*. Slope is the ratio of rise to run (as in exhibit 11.1).

**[INSERT EXHIBIT; render in gray scale. Make all lines solid black.]**

**Exhibit 11.1** Calculation of Slope in an Equation

**Rise**

**Run**

**[END EXHIBIT]**

The equation for a line can be specified by the coordinates of two points on the line. For example, a line can be used to describe the cost of practice in physician groups (Weil 2002). The fixed cost is the intercept of the line and the variable cost is the slope. Some physician groups may be above this linear prediction of cost, in which case they are less efficient than the average group. Other groups may be below this line, in which case they are using economies of scale to benefit their operations.

The relationship between two variables may be nonlinear, increasing in portion of time and decreasing other times. These equations are considered *nonlinear*. Data can be transformed so that the relationship between *Y* and transformed *X* is linear. Exhibit 11.2 shows examples of transformations that would make the relationship between the dependent variable and the transformed variable linear. For example, the equation for a parabola (also shown in exhibit 11.2) is

**[INSERT EQUATION]**

.

**[END EQUATION]**

In this graph, is the intersection of the parabola with the *y*-axis (i.e., it is the value of *Y* when *X* is zero). The constant is the slope of the line; it shows how much *Y* increases with one unit of increase in *X*. Note that if *X* is negative or positive, *Y* is still positive. Consider a transformation of the data so that . Now the relationship between *Y* and the transformed data is linear: . Exhibit 11.2 shows various equations relating two variables and how one of the two variables can be transformed to create a linear relationship. In exhibit 11.2, you can find the transformations for power, exponential, logarithmic, and reciprocal functions. The original form is nonlinear, but the subsequent form is linear. For example, the following function, at the bottom of exhibit 11.2, shows a nonlinear reciprocal relationship between *Y* and *X*:

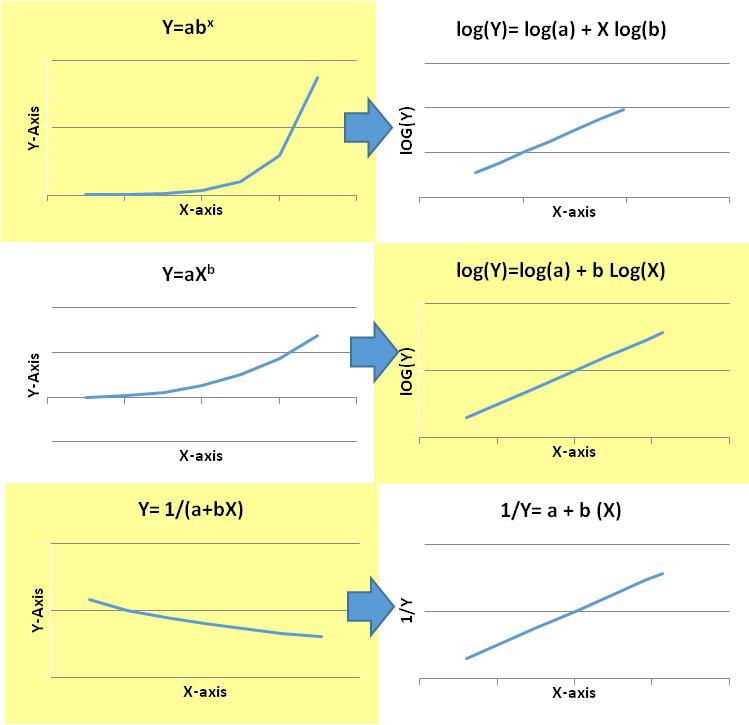
**[INSERT EQUATION]**

.**[END EQUATION]**

Yet, if we examine the relationship between the reciprocal of *Y*(i.e., 1 ÷ *Y*) and *X*, the relationship is linear. Similarly, if *Y* is a power function of *X*, then the logarithm of *Y* has a linear relationship with the logarithm of *X*. This logarithm transformation is shown in the middle row of exhibit 11.2.

**[INSERT EXHIBIT]**

**Exhibit 11.2** Transformation of Nonlinear Equations

****

**[END EXHIBIT]**

In practice, data relating *X* and *Y* may not follow any of the examples given in exhibit 11.2. In these situations, it may make sense to fit an equation to the data. Excel can be used to create a scatter plot. In the plot, a trend line equation can be fitted to the data. Polynomial, power, logarithmic, and exponential are examples of trend line equations that can be fitted to the data in Excel. Exhibit 11.3 shows the fitted trend line equation for a set of data.

**[INSERT EXHIBIT; make axis labels rom, not bold. Render in gray scale. Make line solid black.]**

**Exhibit 11.3** Example of a Trend Line Equation in Excel

**[END EXHIBIT]**

In exhibit 11.3, the trend line fits the data well. Because both the term and *X* term are present, it is not immediately clear what transformation would make the relationship linear. An approximation may be used to make the data more linear.

In regression analysis, the most common equation is a multiple regression equation in which the dependent variable *Y* is described as a function of multiple independent variables. The following shows the dependent variable cost, *Y*, as a function of two variables of the patient (age, and gender, of the patient. The equation includes a constant, a linear component of main effects, and a third component showing an interaction between the two variables, . So

**[INSERT EQUATION]**

**[END EQUATION]**

The interaction term is the product of the two variables. The main effect is the linear portion, and the interaction effect is the nonlinear portion of the regression.

# [H1] Fitting Data to an Equation: Residuals

As mentioned earlier, regression is a procedure for fitting data to an equation. The analyst specifies the equation, and the software fits the data to the equation by estimating the parameters of the equation from the data. By convention, parameters of an equation are shown using the Latin alphabet (e.g., ). Greek letters are used to indicate estimates obtained for these parameters through regression. For example,can indicate regression estimates of the parameters The Greek letter is read aloud as “BAY-tuh.” These equation parameters are also referred to as *regression coefficients*. We use the terms *coefficient* and *parameter* interchangeably to break the monotony of a commonly repeated word.

In regression, the equation for a straight line is used often. As in the vocabulary used for linear equations, regression parameters for a straight-line equation are the *intercept* and the *slope*.

**[INSERT EQUATION]**

**[END EQUATION]**

In this equation, is the intercept, is the slope, and *X* is an independent or predictor variable. We use the hat on top of *Y* to indicate that it represents an estimated variable calculated from independent variables. It is not the observed value of the outcome but an estimate from the formula. We show the slope and intercept using Greek lettering to emphasize that these are estimated from the sample data.

To accurately estimate the intercept and slope parameters, regression calculates a residual. The *residual* is the difference between the estimated dependent variable and the observed dependent variable. Regression minimizes the sum of the squared residuals. For a straight line, the intercept, , and the slope, , parameters are estimated to minimize

**[INSERT EQUATION]**

.

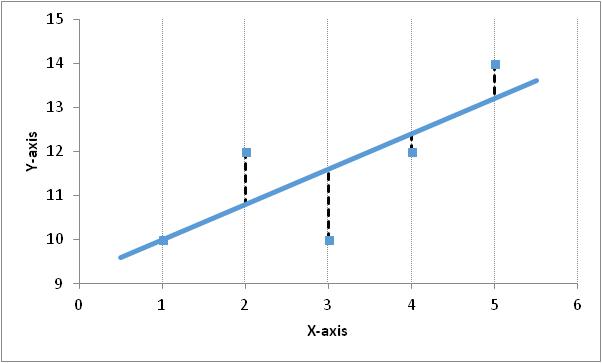
**[END EQUATION]**

In exhibit 11.4, the differences between the observed and predicted values are shown as a dashed line. At *X* = 1, there is no difference. At *X* = 4 there is a small negative difference. The largest difference is at *X* = 3. The positive differences are at *X* = 2 and *X* = 5. Each one of these differences is considered a residual. Regression is fitted to the data so that the sum of the squared residuals is the smallest possible number.

**[INSERT EXHIBIT; in graph, make blue line and blue squares black. Make axis labels and numbers rom, not bold. Remove “-axis” from the axis labels.]**

**Exhibit 11.4** Example Data and Plot of Residuals

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *y* | *x* | *Predicted y* | *Residual* | *Squared Residuals* |
| 10 | 1 | 10.00 | 0.00 | 0.00 |
| 12 | 2 | 10.80 | 1.20 | 1.44 |
| 10 | 3 | 11.60 | −1.60 | 2.56 |
| 12 | 4 | 12.40 | −0.40 | 0.16 |
| 14 | 5 | 13.20 | 0.80 | 0.64 |

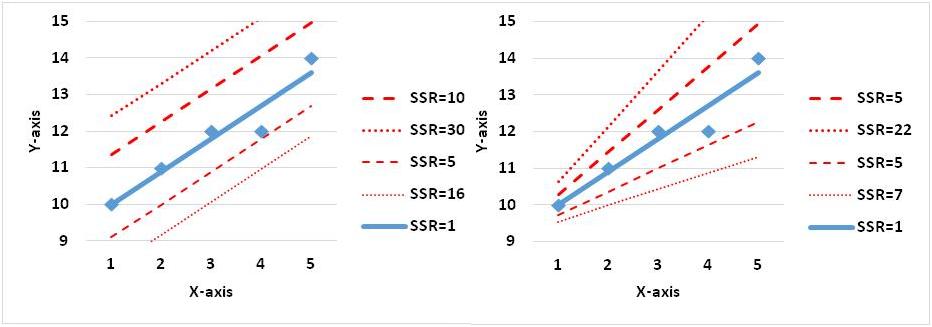


**[END EXHIBIT]**

The line in exhibit 11.4 has the sum of squared residuals calculated as (0)2 + (1.2)2 + (−1.6)2 + (−.4)2 + (.8)2 = 4.8. Different lines will have different residuals. The objective of regression analysis is to choose the regression parameters (intercept and slope, in cases of linear regression) so that the line has the minimum sum of squared residuals. Exhibit 11.5 shows the sum of squared residuals associated with lines that have different intercepts or slopes. The left side of exhibit 11.5 shows five lines with different intercepts but the same slope. Each line has a different sum of squared residuals. The one with the lowest sum is the solid blue line—this is the regression line. The same is true for the right side of the exhibit, where five lines are shown with different slopes. Each line has a different sum of squared residuals. Again, the line with the lowest sum is the regression line.

**[INSERT EXHIBIT; render in gray scale. Make axis labels rom, not bold. Remove “-axis” from axis labels. If possible, eliminate legends and label the lines instead. If that won’t fit well in the layout, I’d say that retaining the one legend between the graphs would be better than repeating the legend, so eliminate the right-hand legend]**

**Exhibit 11.5** Alternative Lines and the Sum of Squared Residuals



**[END EXHIBIT]**

The reader may reasonably ask—why sum the squared residuals? This step is arbitrary; there is no mathematical reason for it, except that such a sum allows one to easily calculate the parameters of regression. By “easy to calculate,” we mean that it is not necessary to repeatedly try different lines and see which one has the minimum sum of squared residuals. A formula for estimating the parameter that would produce the minimum sum of squared calculation has been worked out, and it guarantees that the estimated parameters would produce the minimum sum of squared residuals.

## [H1] Example: Analysis of Costs

To explain the ideas in this chapter, we use data from a study of the cost of nursing home care. One alternative to nursing home care is to allow the resident to rent a place in the community and bring the medical services to the resident. This alternative to nursing home is called a *medical foster home* (MFH). In the MFH, the resident lives in a small community home with a family but has access to the needed medical or nursing care. At this community home, the resident interacts with the entire family living at the same place, eats with the family, and is generally looked after by the entire family. Think of the MFH program as a shared home, akin to the services offered by Airbnb lodging, and think of the nursing home as a large hotel. The MFH is in the community and small, the nursing home is centralized and large. The foster home does not create the economy of scale introduced by having many elderly residents in one place; the patient benefits, however, from the constant attention of the family. Which of these arrangements is best for the resident? MFH residents receive more attention from the caregiver, they socialize across generations, and they do not face the risk of infection outbreaks in large nursing homes. On the other hand, the cost of organizing the home and taking care of these residents may be higher, as there are no economies of scale.

Obviously, the true cost is more than just the residential cost. Additional hospitalization and outpatient visits may increase the cost of residents in either program. The cost information should include residential, inpatient, outpatient, and pharmaceutical costs. While gathering the cost information is cumbersome, it is still possible. A bigger difficulty occurs when MFH residents and nursing home residents differ in their medical histories. Would the MFHs shun sicker residents and thus end up with residents who would be less expensive, regardless of location? After all, families are likely to select a resident who fits them well—they may avoid residents who drink too much or who are mentally ill.

To compare the cost of residents in the two programs, we must make sure to control for differences in the medical history of the residents. We examined diagnoses in the year prior to enrollment in either home type. Because patients arrive with many diagnoses, it is important that various diseases are organized into smaller sets of categories. We categorized the 14,400 diagnoses of the *International Classification of Diseases* into 250 related illnesses, using the Agency for Healthcare Research and Quality Clinical Classifications Software (Healthcare Cost and Utilization Project 2017). We will use these data to predict the cost of long-term care for both types of patients. If their residence is a significant predictor of cost even after adjustments are made for medical history, we might be able to claim a cost advantage for either of these places.

The purpose of the statistical evaluation of cost data is to describe relationships between cost and several variables, including program participation and components of medical history prior to program participation. In our example, we would want to predict total cost for each resident from participation in a MFH program, demographic differences (age, gender, race), and 250 predictors from the resident’s medical history. The variable to be explained (total cost) is called the dependent variable, or, alternatively, the response variable; the variables that predict it (participation in MFH, age, gender, race, 250 medical history indicators) are called independent variables or predictor variables. Participation in the MFH may also be called the “treatment variable.”

If the dependent and independent variables are continuous, correlation and scatter diagrams can describe the relationship between them. For example, correlation and scatter diagrams can describe the relationship between age and total cost. Multivariate regression can provide a mathematical formula that represents the relationship among all the independent variables and a single dependent variable. In this chapter, we use the data from our MFH example to introduce multiple regression.

## [H1] Example with a Single Predictor of Cost

Using data from the MFH project, we can regress cost on the age of the patient. The outcome, or dependent variable, is the total cost of patient care. The independent variable is the age of the resident. The regression examines the relationship between age of the patient and cost. Our initial assumption is that as age increases, the cost of care increases proportionally. If so, the regression equation is

**[INSERT EQUATION]**

.

**[END EQUATION]**

The intercept, , is the mean cost for patients who have the minimum age (when the *x* variable is zero). The slope, , is the average change in cost for each additional year of aging. This is the proportion of increase in cost when age increases by one year.

To use Excel to do regression you must first add in the analysis tool pack. To do so, click on the Excel logo, click on Add-Ins from the vertical menu list, select Manage Add-Ins, and click on Go. Once the Add-In menu comes up, select both of the analysis tool packs. See exhibit 11.6 for a visual presentation of these steps.

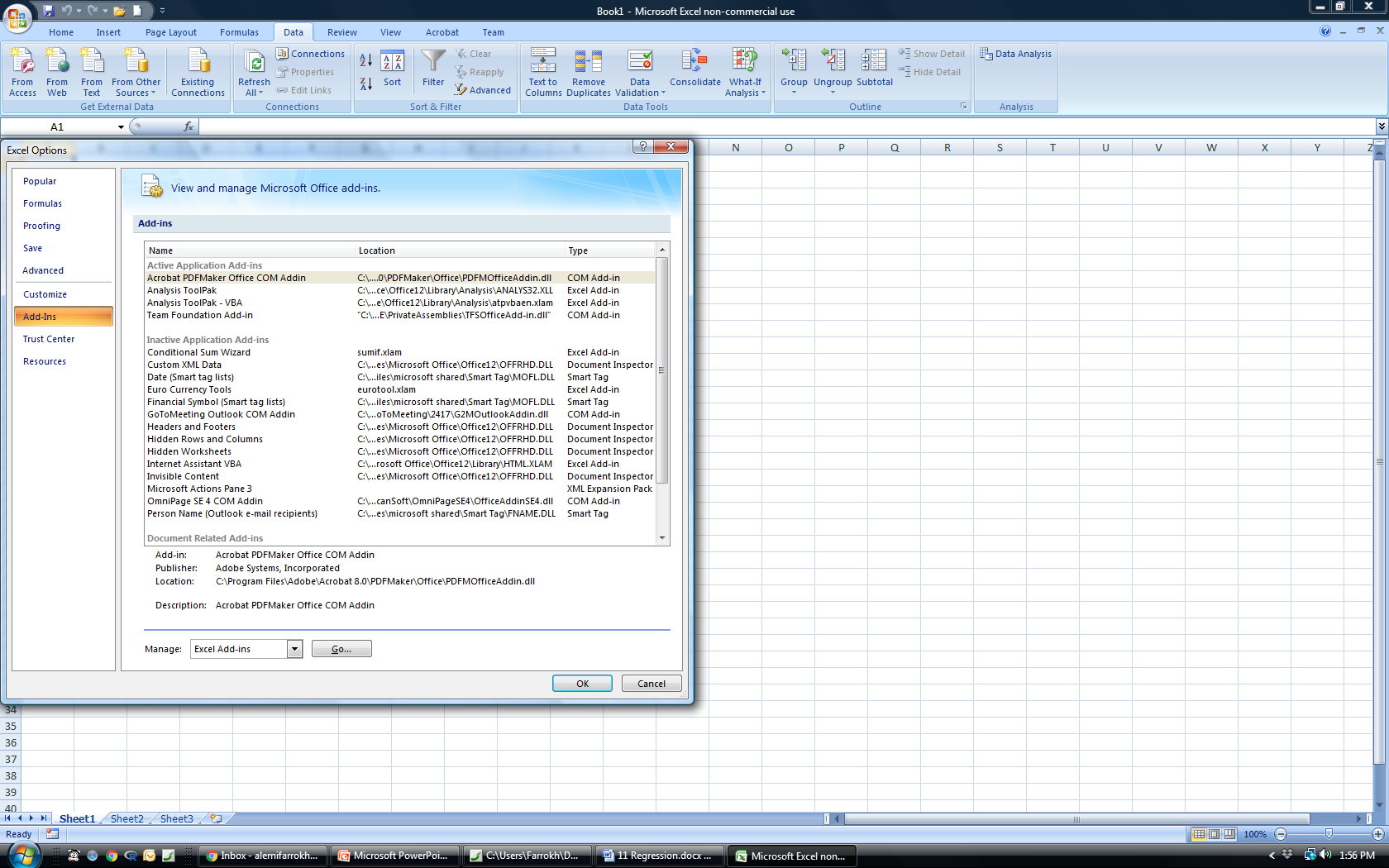
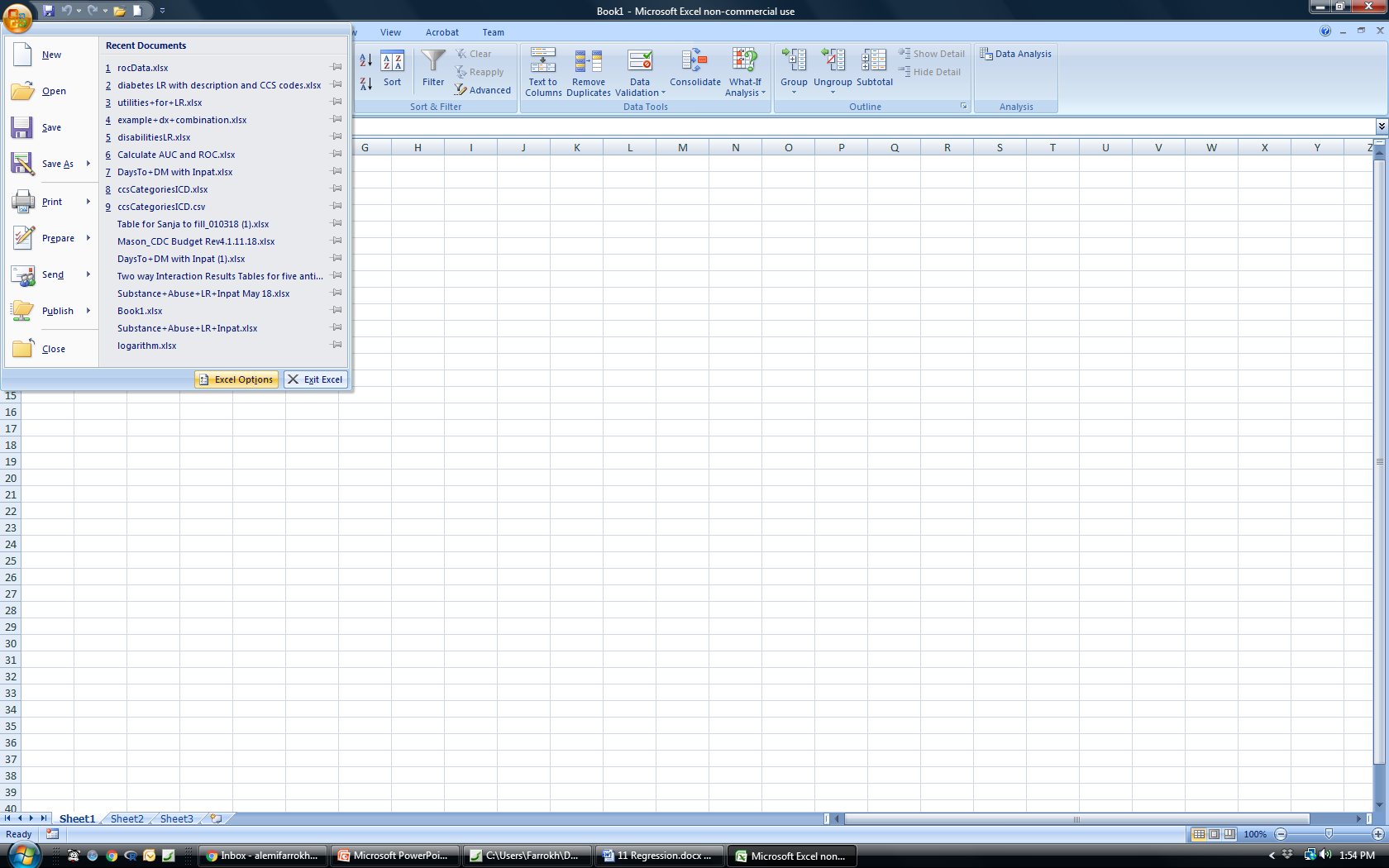
To actually do the regression, you need to select Data in the general Excel menu; select Data Analysis in the Data menu; select Regression under the Data Analysis menu. Then fill in the regression menu. These steps are also shown in exhibit 11.6.

**[INSERT EXHIBIT]**

**Exhibit 11.6** Specification of Regression in Excel 2013

3

4

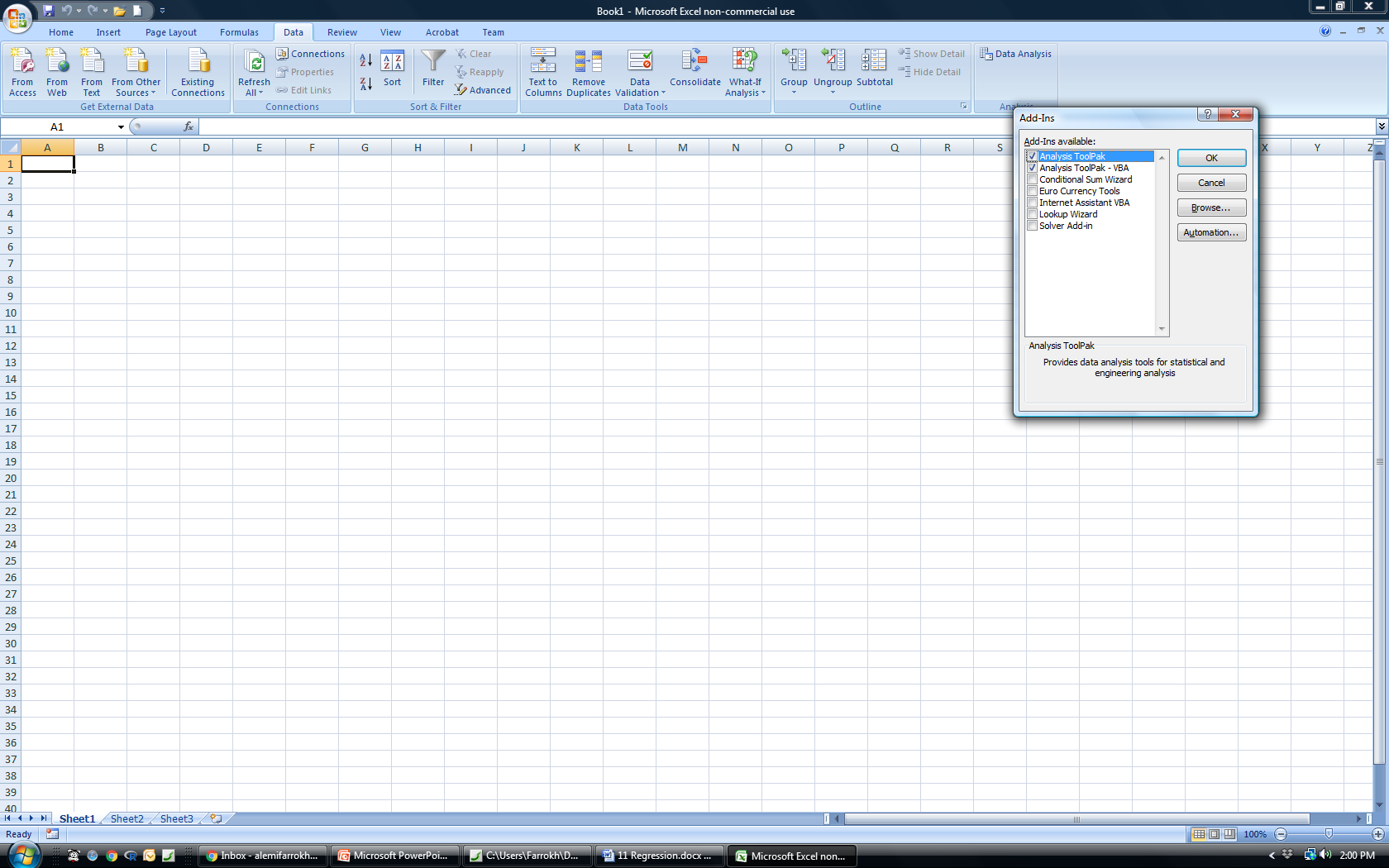


1

2

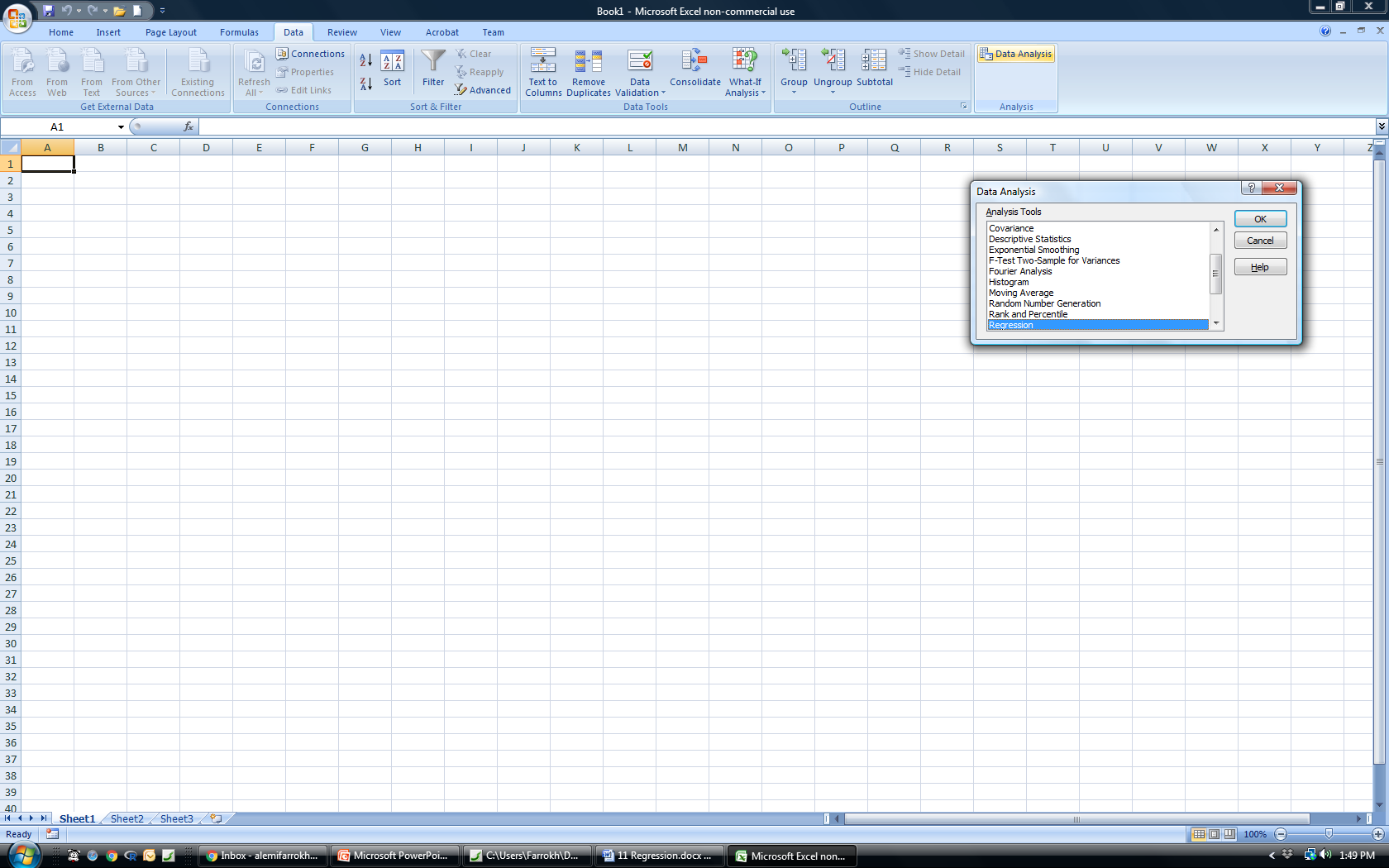
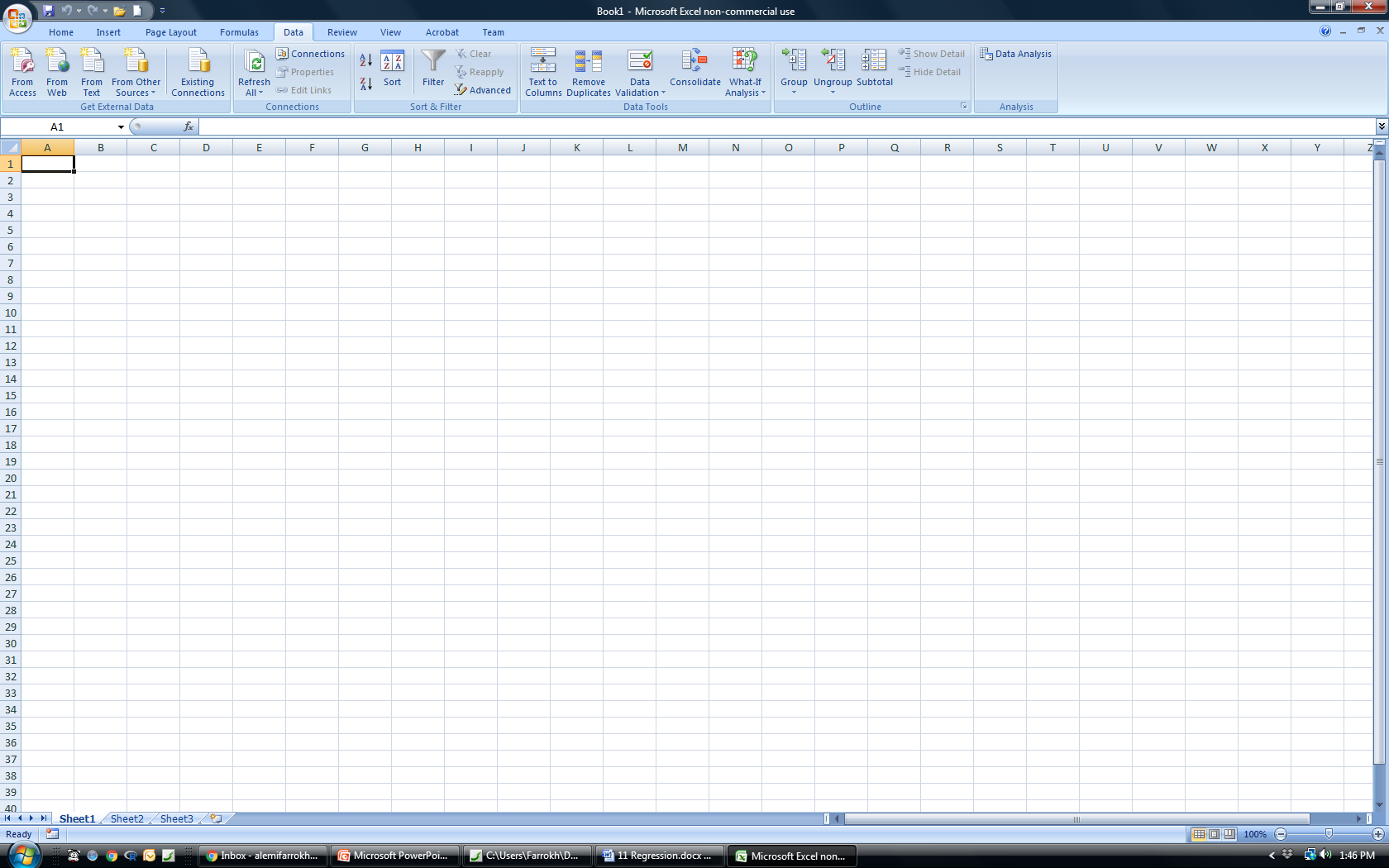
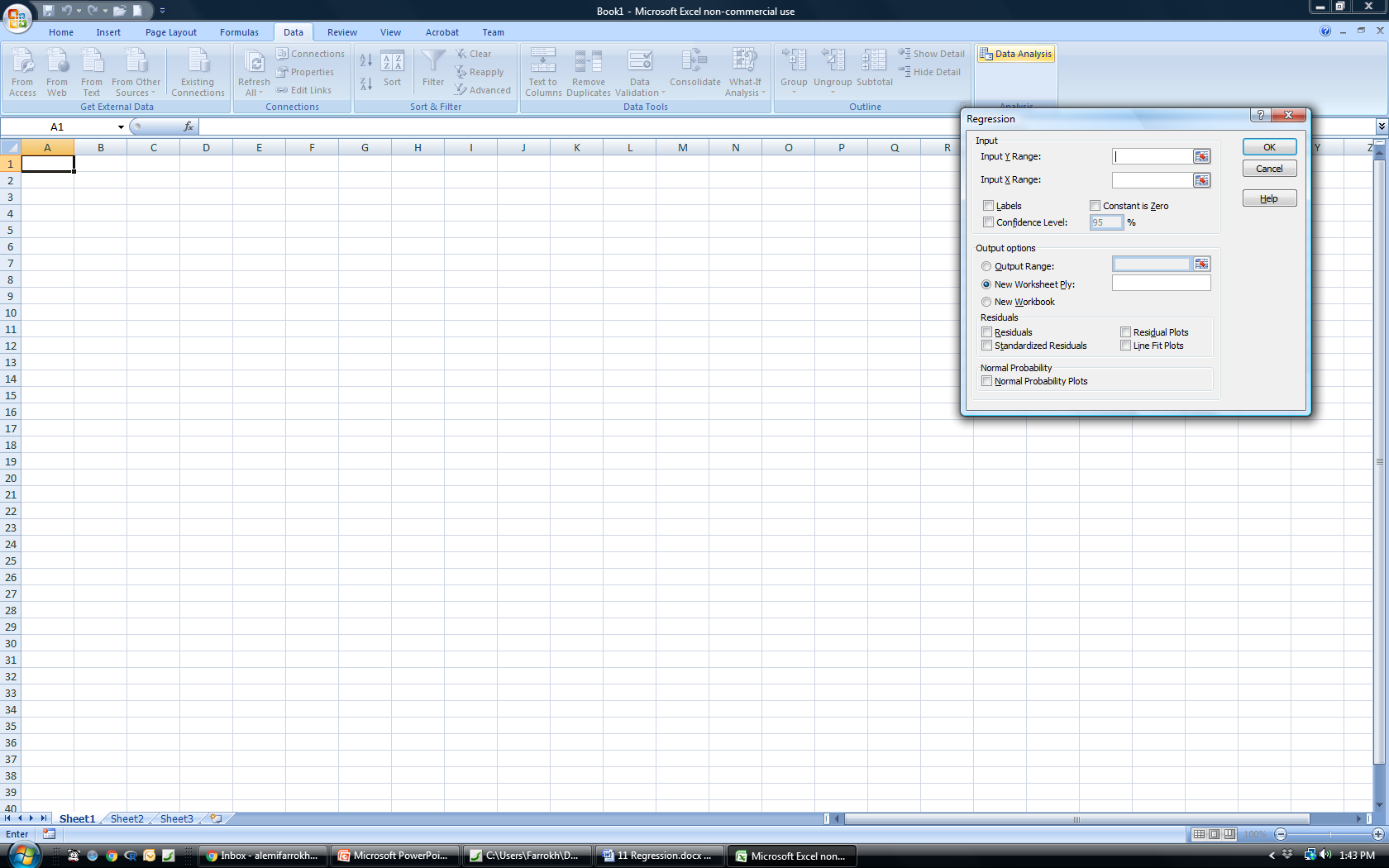
3

4



5

6



7

8

9

8

**[END EXHIBIT]**

The regression menu requires the specification of the input variables. The dependent, or *Y*, range must be specified. The dependent variable must be numerical, with no missing value or text. A common error in Excel is that one of the values in the dependent variable is entered as text (e.g., NA, or null), making it impossible to run the regression. The range for the independent variables should also be specified. This range must be as long as the dependent variable. A maximum of 250 variables can be included in this range. If the first row contains the name of the variable, check the box next to Labels. If there are no intercepts allowed in the regression equation—a rare situation—check the Constant Is Zero box. The output options must be specified; do not designate the place where you kept the data as the output. This will wipe out the data. Options on residual plots can also be specified. The normal probability plot tests the assumption of a normal distribution of variables.

The result of the regression of the cost data is provided in exhibit 11.7. These data show the *adjusted R-squared*, which is a measure of goodness of fit. They show analysis of variance (ANOVA). The heading SS refers to the sum of squared residuals. The column titled F shows the F-statistic. This statistic and the corresponding Significance F column show whether the regression explains a statistically significant portion of the variation in residuals. In these data, the F-statistic is 160.72. It is statistically significant at alpha levels less than 0.05, as the probability of observing this F-statistic is given under the column Significance F, and it is near 0, less than 0.05. The regression coefficients are also provided, starting with intercept and age. The *p*-value in this table tests whether the coefficient is significantly different from zero. It gives the probability of observing the coefficient. If the reported probability is smaller than 0.05, then it is less likely that it could have occurred by random chance.

**[INSERT EXHIBIT]**

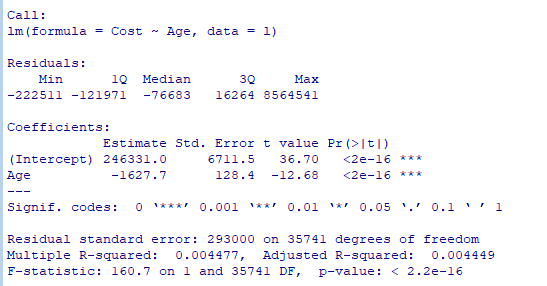
**Exhibit 11.7** Output of Excel Regression

****

**[END EXHIBIT]**

Using code from R, a free and widely used software package for statistical analysis, we get a similar output. R code requires us to call a function called lm(). In the following lines of code, the column listed before the tilde (pronounced “TIL-duh”) is the response or dependent variable and the column listed after it is the independent variable. The first set of output is the distribution of residuals shown in terms of five points: minimum, first quartile, median,third quartile, and maximum. One expects to see a normal symmetric distribution around the median. The Coefficients column lists the intercept and age parameters. The probability that the *t*-statistic calculated from the observed coefficient is different from zero is also provided. Note that the coefficients estimated by R and Excel (exhibit 11.7) are the same, within rounding error.

**[LIST FORMAT]**

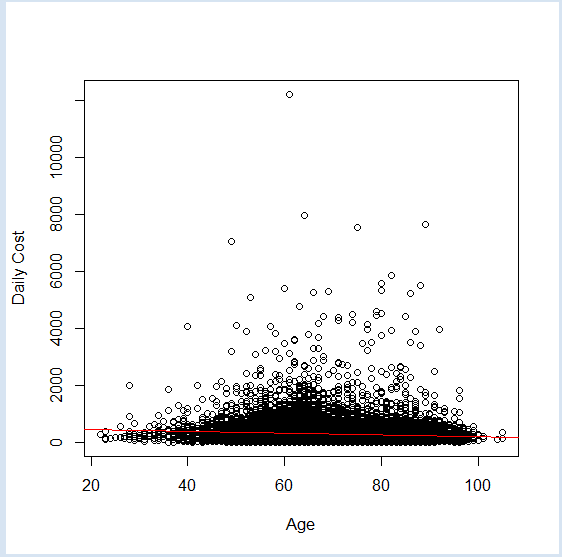


**[END LIST]**

Exhibit 11.8 plots cost against age, with the linear fitted line shown in red. The *x*-axis is age; the *y*-axis is daily cost. You may also wish to see this plot and resulting regression using R software. See the appendix for how to do this in R.

**[INSERT EXHIBIT]**

**Exhibit 11.8** Excel Plot of Cost and Age



**[END EXHIBIT]**

A regression produces coefficients that can be used to create an equation to relate the independent variables to the dependent variable. Using the reported regression results, we know that the equation should look like   
**[INSERT EQUATION]**

**[END EQUATION]**

The regression estimated the coefficients, so that the resulting equation is now

**[INSERT EQUATION]**

**[END EQUATION]**

That is, as age progresses, the cost declines by every year of age. This obviously does not fit our intuition. We expect that as age increases, patients will become sicker and therefore costlier. A quick examination of the regression output for the MFH raises another concern. Less than 1 percent of the variation in cost was explained by age. So not only does age reduce cost, but also age is not a good predictor of cost. Keep in mind that age is significant in predicting the cost, but the effect size is small. We will see shortly that this is in part the result of the violation of assumptions of regression.

# [H1] Independent Variables

In an ordinary regression, independent variables can be binary, categorical, continuous, or interaction variables. Binary or dummy variables are assigned two values: 0 or 1. They are used often to indicate the occurrence of an event, such as an intervention or an adverse event. Categorical variables have multiple discrete levels, and before regression is completed, categorical variables must be listed as multiple binary variables. Continuous or interval variables can enter regression as they are, without transformation. Interaction variables are the product of two or more variables.

This section describes the preparation of independent variables. We start with an example of a binary or dummy variable. If the analyst is interested in the impact of gender on cost of care, then she would regress cost on gender and obtain the equation

**[INSERT EQUATION]**

.

**[END EQUATION]**

In this regression, mean cost of care is a continuous outcome or dependent variable; gender is a binary independent or predictor variable. Gender is coded 0 for males and 1 for females. Based on this coding, which is arbitrary, there will be two estimates of mean cost. One is for females,

**[INSERT EQUATION]**

**[END EQUATION]**

and the other is for males,

**[INSERT EQUATION]**

**[END EQUATION]**

The slope estimates the per unit change in cost as a function of a one-unit increase in gender (i.e., a shift from 0, male, to female, 1). Remember that the coding is arbitrary and females could have been coded as zero and males as 1, which would not change the magnitude of the slope . If the slope is positive with one set of coding, then it would become negative with the reverse coding. In other words, the mean difference between females and males remains the same no matter the coding, but the sign of the value would change depending on which value is used as the reference value. The following shows the R code to estimate the coefficient for gender in predicting the cost of healthcare, when males are coded as 1 and females as 0. To recode the values of the gender variable to 0 and 1, we use the following simple loop.

**[LIST FORMAT]**

|  |
| --- |
| To perform this regression, again, we simply use the lm() function. Our dependent variable is Cost, which is column 1, and our independent variable is Gender, which is column 5.  And the results of this regression can be found using summary() as follows: |

**[END LIST]**

So what these results are saying is that there is no statistically significant difference between males and females.

Sometimes, the independent variable has several categories. These categorical variables must be turned into binary variables before conducting the regression. For each category, a new binary variable is created. One of these binary variables is dropped, as the combined absence of all other variables implies the dropped variable. The dropped category will act as a reference point for comparisons of all other categories. For example, consider patients’ race. It has multiple categories, including White, Black, Asian, and Other, four categories that we will use in this example for the sake of simplicity. We cannot use the 0 and 1 coding approach because race has more than two categories. Each category can be turned into a new binary variable. Thus, we can have a binary variable called “White”; its value will be 1 if the patient is White and 0 otherwise. Another binary variable called “Black” will be 1 if the patient is black and 0 otherwise. Still another variable, “Asian,” will be 1 if the patient is Asian or 0 otherwise. Finally, Other will be 1 if the patient is none of the listed races and 0 otherwise. The single categorical variable Race was broken into the following four binary variables, assuming values of 1 and 0:

**[INSERT NL]**

1. X1 White = 1, Otherwise = 0
2. X2 Black = 1, Otherwise = 0
3. X3 Asian = 1, Otherwise = 0
4. X4 Other = 1, Otherwise = 0  
   **[END NL]**

Then, instead of regressing cost on race—that is, , the statistician must regress cost on the three out of four binary variables that indicate the patient’s race. Suppose we select White as the reference value (the value when all three indicators are 0). The multiple regression will be

**[INSERT EQUATION]**

**[END EQUATION]**

The intercept in this equation estimates the cost for white patients (where black = 0, Asian = 0, and other = 0). The mean cost for black patients is, for Asian patients it is , and for patients of other races it is . The R code for estimating these regression parameters is provided in the following box. Again, we use the lm() function and we have

**[LIST FORMAT]**

|  |
| --- |
| And the results of the regression are as follows: |

**[END LIST]**

# [H1] Main Effect and Interactions

In multiple regression, the impact of the variable by itself is referred to as the *main effect*,and the *interaction effect* is the impact of a combined set of variables. Interaction between two or more variables is modeled by the product of the variables. For example, consider two binary variables; when the variable is present, it is coded as 1—otherwise, 0. The product of two binary variables is another binary variable, which is 1 when both variables are present and 0 otherwise. With two binary variables, four situations are possible. The only time the product is 1 is when both variables are present, as we can see in the following equations:

**[INSERT EQUATION]**

**[END EQUATION]**

If one variable is binary and another continuous, the product of the variable will be 0 when the binary variable or the continuous variable is 0; otherwise, it will be the same as the continuous variable. If both variables are continuous, then the product of both variables will be continuous, and it will be a larger number than the sum of the two variables. Including interaction terms in regression expands our ability to model events.

A typical hospital discharge includes five diagnoses. These diagnoses interact, meaning that the combination of the diagnoses has a different impact than the sum of the two diagnoses by themselves (Extermann 2007). In predicting patients’ prognosis, the interaction among these diagnoses must be modeled. The full model then includes the main, pair-wise, three-way, four-way, and five-way effects of combinations of the five diagnoses. The following equation shows the set of interacting variables on separate lines to highlight different combinations of variables:

**[INSERT EQUATION]**

.

**[END EQUATION]**

Because specifying interactions is time-consuming, R language has provided shorthand for specifying interactions. Typically, a set of independent variables in parentheses is shown to the power of a number. The software interprets the number as the highest interacting term that should be included in the regression. For example, the following regression specification indicates that all pair-wise terms should be included:

**[INSERT EQUATION]**

.

**[END EQUATION]**

This specification is the same as indicating a model with the terms

**[INSERT EQUATION]**

**[END EQUATION]**

Despite the ease of including interaction terms in regression equations, few investigators do so. Inclusion of a full set of interaction terms could exceed the number of cases available and thus over specify the model (i.e., more independent variables than cases in the sample). When a large set of variables is present, estimating interaction effects may not be possible—or tracking interaction effects may be beyond human cognitive abilities. At the same time, it is precisely when a large number of variables is present that interactions are most likely to be statistically significant and ignoring the interactions could lead to erroneous conclusions.

A note of warning is necessary about the effect of interactions on the interpretation of regression parameters. Interaction terms change the interpretation of the regression parameters. At the start of this chapter, we discussed the meaning of regression parameters. We did so with a single predictor or with linear models, where interaction terms are not present. When interaction terms are present, the regression coefficients no longer indicate the rise in outcome for the run in the variable. In fact, the notion of a run in a variable is no longer clear. Everything we mentioned about interpretation of the coefficient as slope of the line is not valid if interaction terms are present. In a sense, our effort to explain regression in ways that could help create intuitions about it has put us in a pickle. Our explanations do not work when interaction terms are present.

So, what is the meaning of the regression coefficients when interaction terms are present? There is no simple answer to this question. The coefficients are the parameters of the equation, and they may not have a meaning other than being an estimate. An example can demonstrate the problem with interpretation of the regression equation when interaction terms are present. Suppose we want to predict cost from the gender (male = 1, female = 0) and age of the patient. The regression equation would be

**[INSERT EQUATION]**

**[END EQUATION]**

A one-year increase in age now involves a unit increase in cost. The cost for males increases more quickly than for females. Similarly, a change from female to male increases cost by Also, the impact of change in gender depends on the age of the patient. In low ages, cost increases a small amount, but it increases by a larger magnitude for older people. Interactions change the meaning of the parameters of regression. With two variables, it is still possible to explain the meaning of the regression coefficients in words, but as the number of interacting variables increases, words may fail to describe what each coefficient means.

The code for regression of cost on gender and age is displayed in the following box. The interpretation of these findings requires careful statements of how the interaction affects cost. To perform a regression in R with two independent variables and a dependent variable while including an interaction term, we use the lm() function as follows:

**[LIST FORMAT]**

|  |
| --- |
| The results of this regression are as follows: |

**[END LIST]**

# [H1] Coefficient of Determination

The R-squared, , is used to measure the goodness of fit between the equation and the data. It reports the strength of association between the outcome variable denoted by *y* and the independent variables. The value of can be anywhere from 0 to 1, where 0 represents no association and 1 represents the strongest possible association. Anof 1 indicates that the independent variables perfectly predict the dependent variable. Because this only occurs when the prediction task is trivial, over specified, or tautological (predicting something from itself), one never sees an of 1. Only God has perfect knowledge (*R*2 = 1); humans are less certain (*R*2<1).

The statistic estimates the percentage of variability in the outcome variable explained by independent variables. The total sum of squares is a measure of the variability of observations around the mean of the observations. The unexplained sum of squares is the sum of the square of the residuals, where residuals are the difference between observed outcome and predicted outcome. The difference between total sum of squares and sum of squared residuals is the explained sum of squares, represented by the equation

**[INSERT EQUATION]**

Total sum of squares = Explained sum of squares + Unexplained sum of squares.

**[END EQUATION]**

The coefficient of determination is then calculated as

**[INSERT EQUATION]**

**[END EQUATION]**

One peculiar problem with is that it increases whenever a new variable is included in the model. In other words, one expects the fit to the data to become increasingly accurate as the number of variables increase. The adjusted R-squaredcorrects the value by the degrees of freedom available. Thus, it allows the analyst to examine whether the addition of a new variable increases the adjusted value and thus makes real improvement.

Keep in mind that the statistic is sensitive to the size of the data. As the size of the data increases, there is more variability in the data; a smaller percentage of variability could be explained by a fixed set of variables. Thus, for a very large data set (e.g., millions of patients), 100 independent variables may explain 25 percent of the total variability. Despite the low number, this may be a reasonable model. For a small data set (e.g., 2,000 cases), 100 variables may explain 80 percent of the variability. Anof 25 percent may be low in a small data set but reasonable in a massive database.

For each regression, the total variation in data—known as sum of squares total (SST)—is calculated as

**[INSERT EQUATION]**

**[END EQUATION]**

In this calculation, is the average of the observed outcomes. The sum of squares regression (SSR), or the explained sum of square, is calculated by the formula

**[INSERT EQUATION]**

**[END EQUATION]**

In this formula, is the predicted value for observation “i,” meaning that we use the regression equation to calculate the predicted value. The ratio of SSR to SST is the percentage of variation, explained by the regression equation

**[INSERT EQUATION]**

**[END EQUATION]**

The unexplained variation is calculated as sum of square of errors (SSE) (residuals):

**[INSERT EQUATION]**

.

**[END EQUATION]**

Mean SSE is a point estimate for the variance of the residuals, and the standard error of the residuals is given by the square root of the mean sum of the square of error. These data are usually reported as the output of regression equations. The hypothesis that at least one of the variables is predictive of the outcome (e.g., one of the beta coefficients in the regression is not zero) can be tested using the following test statistic, known as the F test statistic:

**[INSERT EQUATION]**

.

**[END EQUATION]**

In this equation, *k* is the number of independent variables in the model, and *n* is the number of observations.

We can reject the null hypothesis that there is no relationship between the regression equation and the outcome variable if the probability of observing the F-statistic is less than a predetermined acceptable error rate. We can find the sum of squares of a regression model, along with the analysis of variance, using the anova() function in R, as follows:

**[LIST FORMAT]**

|  |
| --- |
|  |

**[END LIST]**

The box reports the mean square and F value test for each of the variables in the regression. To estimate the sum of square associated with the regression, we need to add up the sum of squares associated with each variable in the regression. The ratio of the sum of squares associated with regression and the total sum of squares is the coefficient of determination, which was reported in an earlier R code to be 1 percent.

# [H1] Model Building

As you develop regression models, it is important to keep two competing considerations in mind. The first is that an analyst wants a model that accounts for most of the variation of the dependent variable. In other words, develop a model that maximizes the coefficient of determination—the explained proportion of variation in the dependent variable. A larger portion of the variation in outcome can be explained by including more independent variables in the model. The second consideration is that the regression model should be as parsimonious as possible, meaning that the simplest model with the highest coefficient of determination should be chosen. This criterion is maximized by selecting as few independent variables as possible. The analyst must examine several models, then choose the model that has the best predictions with the fewest independent variables.

There are several methods for doing so. These methods try out different regression models, then chose the one that is most attractive. The first method is called *hierarchical regression*. In hierarchical regression, predictors are selected based on the knowledge of experts or literature review. Variables known to have a strong relationship to outcome are put into the first regression. The percentage of variation in outcome explained by the regression is noted. Then, an additional independent variable or a set of additional variables are added, and a second regression is done. If the percentage of variation in outcome explained by the second regression is significantly higher than the percentage explained by the original regression, the new variable or variables are retained in the regression. Otherwise, the variables are discarded. The process is continued until all relevant independent variables are examined.

In this approach, the analyst selects the order in which various sets of variables are entered into the regression. When the order is decided by software, using the variable with the highest percentage of variation explained first, then adding another variable, the process is called the *forward stepwise selection of variables*. In forward selection, once a variable is entered into the regression equation, the variable is not taken out if other variables undermine its influence on the outcome. When we start with all independent variables and progressively drop variables that do not explain variation in the outcome, then the process is called *backward stepwise selection of independent variables*.

To avoid overfitting the data, it is important that the performance of various models—whether hierarchical, forward, or backward—should be checked in new data sets, a concept known as cross-validation. *Overfitting* the data happens when the number of variables in the regression exceeds the number of data points. In these situations, regression would perfectly predict all cases. Yet the equation is misleading, as chance variation in the data—what is often referred to as noise—is mistakenly attributed to the variables in the model.

The choice of variables becomes more important in predictive medicine, where the goal is to predict patients’ diagnoses ahead of time. Large data sets and many independent variables (sometimes thousands) are analyzed to predict the probability of future disease. The healthcare field is currently expressing a great deal of enthusiasm about predictive medicine, but this excitement is unfounded if the coefficient is not large. Without a large coefficient of determination, clinicians may have too many false predictions, and thus the regression may not be useful in a clinical setting. For example, this is precisely what happened in the initial efforts to use electronic health records (EHRs) to anticipate influenza outbreaks. The idea seemed simple, and massive data were pulled together. A model was learned and used to predict influenza outbreaks. However, the predictions were not accurate (thanks to the low coefficient of determinations), and clinicians who were trying to use the predictions were frustrated. There were too many false positive predictions, and the effort was abandoned. For more details on using mathematical modeling to anticipate outbreaks, see the work of Alemi and colleagues (Alemi et al. 2013; Alemi et al. 2012).

# [H1] Regression Assumptions: Correctness of Model Form

There are four major assumptions for linear regression models. One of the most important assumptions is the correctness of the model form. In regression, the model form is specified by the analyst, and it is important to check and see if the particular equation used fits well with the data. Typically, analysts use a linear relationship between dependent and independent variables. They focus on main effects with no interaction terms. If the true relationships among the variables are nonlinear, then the misspecified model form could lead to erroneous conclusions. Nonlinearity can be examined in plots of residuals versus predicted values. If there are patterns in these plots, the model form is misspecified. To show the shape of residuals under violations of assumption of residuals, we generated several different types of data and plotted the residuals.

Exhibit 11.9 shows plots of residuals versus predicted values for situations in which polynomial, exponential, logarithmic, inverse, or power equations are incorrectly modeled as a straight line. We begin with a simple linear model, whereby the dependent variable is twice the independent variable and errors are randomly distributed using a normal distribution with a mean of 0 and a standard deviation of 1:

**[INSERT EQUATION]**

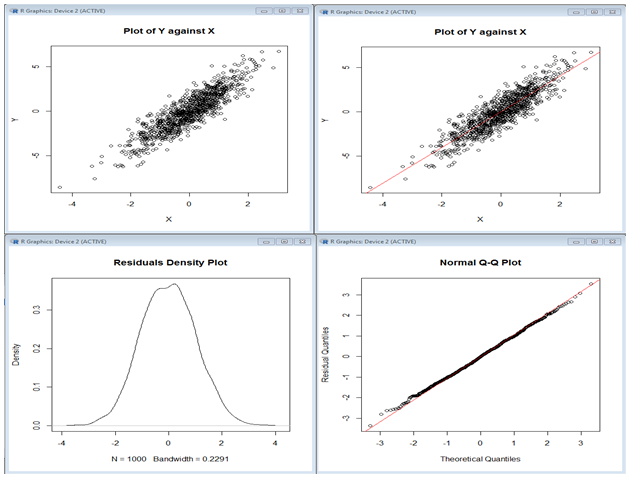
*y* = 2 × *X* + .

**[END EQUATION]**

We generated 1,000 points from this simple model. We made three plots. First, we plotted the *Y* against the *X*-variable. If the normal distribution is reasonable, we should see a linear relationship. Then, we plotted the distribution of residuals. If data are normal, we should see a symmetric distribution with a unimodal peak and concentration of the data in the middle. Last we made Q–Q plots. Q–Q plots are plots of quartiles in the data against quartiles of standard normal distributions. If the data distribution is normal, Q–Q plots should fall along a 45-degree line. These displays show how the plots would look when the assumption of linearity is appropriate. Note how the residual density plot seems symmetric and nearly unimodal, which reassures us that the data are normally distributed. The Q–Q plot is also linear, especially at the two ends of the plot; the plot of *Y* against *X* seems to be linear, supporting the assumption of normal distribution of the data.

**[INSERT EXHIBIT]**

**Exhibit 11.9** Diagnostics for Linear Regression of Linear Relationships

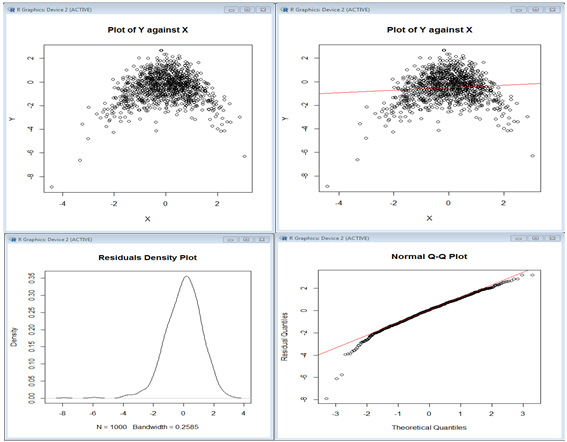


**[END EXHIBIT]**

Next we generated data from a model of the form *y* = + . This is a nonlinear model—*Y* is a power function of *X*. We modeled these data using linear regression—obviously the wrong thing to do. The same three diagnostic plots are reproduced in exhibit 11.10. Note in the plot of Y against *X*, the two tails of data do not follow a linear pattern. This diagnostic should have warned us against using a linear main-effect model. The density plot is no longer symmetric, which tells us that the assumption of a normal distribution of residuals is not met. The Q–Q plot also diverges from the line at the two ends. We presented these diagnostics so the reader would have some experience in detecting situations in which regression assumptions are violated.

**[INSERT EXHIBIT]**

**Exhibit 11.10** Diagnostics for Linear Regression of Power Relationships

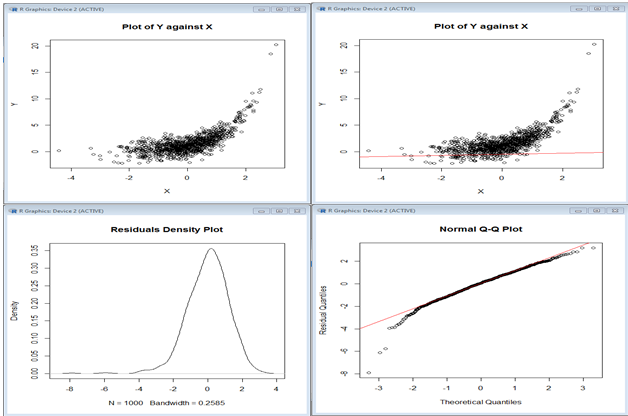


**[END EXHIBIT]**

Next, we generated data from a model of the form *y* = + ; *y* is an exponential function of *X*, but it is modeled using linear regression. Again, the diagnostic plots show that a line is a poor fit (see exhibit 11.11). The residual density plot is not symmetric. Residuals are near zero for some data and increase for others. The plot of *y* versus *X* shows curvature. There is divergence from a linear Q–Q plot, especially at the two ends.

**[INSERT EXHIBIT]**

**Exhibit 11.11** Diagnostics for Linear Regression of Exponential Relationships

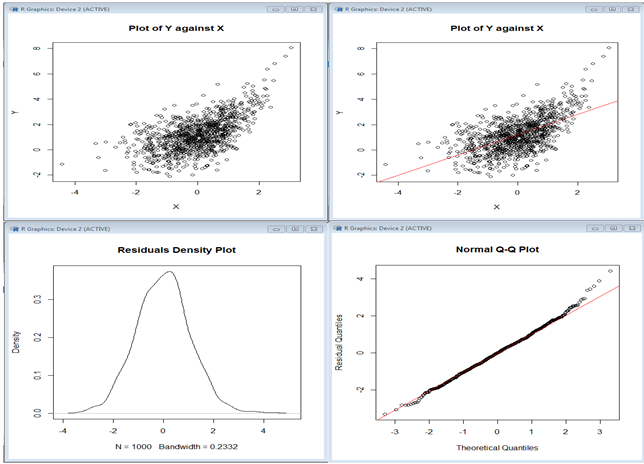


**[END EXHIBIT]**

Finally, we generated data from a model of the form *y* = + . Obviously, the model does not fit a linear regression well (see exhibit 11.12). As in exhibit 11.11, we see a curvature in the *x–y* plot. Again, the residual density plot is not symmetric. The Q–Q plot is not a good fit at higher quartiles.

**[INSERT EXHIBIT]**

**Exhibit 11.12** Diagnostic Plots of the Linear Regression of Nonlinear Data



**[END EXHIBIT]**

The shape of diagnostic plots can tell us whether the linearity assumption is met. The easiest way to see this is in *x–y* plots. The density of residuals and the Q–Q plot tell us if the assumption of normal distribution of errors is met. These plots are informative, and whenever a regression is done, these diagnostic plots should also be examined.

# [H1] Regression Assumptions: Independence of Error Terms

In regression, each observation is assumed to be independent from previous ones. Thus, if patients are the unit of observation, the values of independent and dependent variables are assumed to be independent from each other. This assumption breaks down when one patient affects the health of another, such as contagious diseases or related family members. If, over time, the nature of the patients examined changes or the outcomes improve, this assumption falls apart. It can be tested by predicting the value of a residual at time *t* with residuals at prior periods. The existence of autocorrelation among the residuals is also a sign of a violation of this assumption. A test called the *Durbin-Watson statistic* could be used to examine this assumption.

To perform the Durbin-Watson test on a regression model in R, we first need to download the package named “car” from the web and load it into our local computer’s R library. This can be achieved using the command install.packages(), as shown.

**[LIST FORMAT]**



**[END LIST]**

Once the package is installed, it must be loaded into your R workspace using the command library():

**[LIST FORMAT]**



**[END LIST]**

We are now ready to perform a standard Durbin-Watson test on a regression model. First, let us create a simple regression, which reads

**[LIST FORMAT]**



**[END LIST]**

Next, we will tell R to run a Durbin-Watson on the regression model titled “model.fit” using the command durbinWatsonTest() as follows:

**[LIST FORMAT]**



**[END LIST]**

In this case, R is informing us that the autocorrelation between the residuals is 0.038 with a *p*‑value of 0.218. Hence, residuals are uncorrelated. We can also verify that the residuals are noncorrelated using the acf() command in R.

**[LIST FORMAT]**

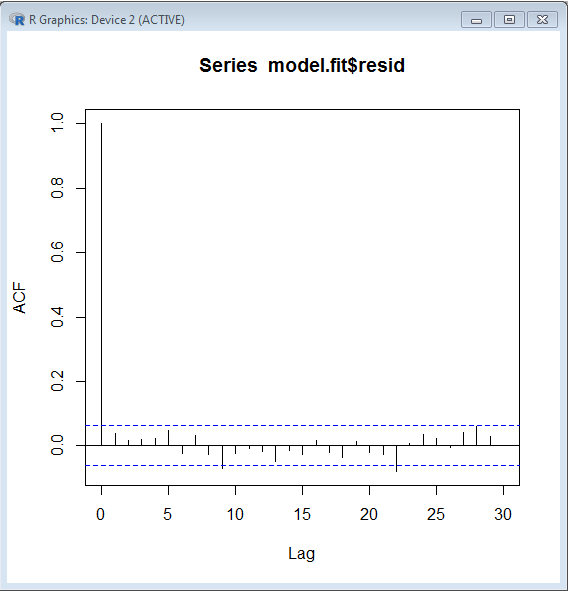


**[END LIST]**

R will plot the simple autocorrelation plot seen in exhibit 11.13.This plot indicates that residuals are noncorrelated, at any time lag. The correlations at lag of 0 is 1, and the remaining correlations at different lags are relatively small.

**[INSERT EXHIBIT]**

**Exhibit 11.13** Autocorrelation Plot



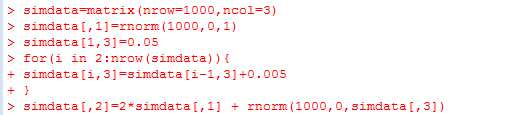
**[END EXHIBIT]**

# [H1] Regression Assumptions: Homoscedasticity

Regression assumes that the standard deviation of an observation does not change over the independent variables. *Homoscedasticity* is a situation in which error terms are distributed randomly. *Heteroscedasticity* is a situation in which error terms have some kind of pattern—for example, when the standard deviation of the sample changes over time. This situation can be detected by plotting residuals over time or over any of the independent variables. If the dispersion of residuals increases or decreases, the assumption may be violated. Exhibit 11.14 shows three plots in which residuals are increasingly becoming larger as the value of the independent variable increases.

To simulate data where there is heteroscedasticity, we ran the following R code:

**[LIST FORMAT]**

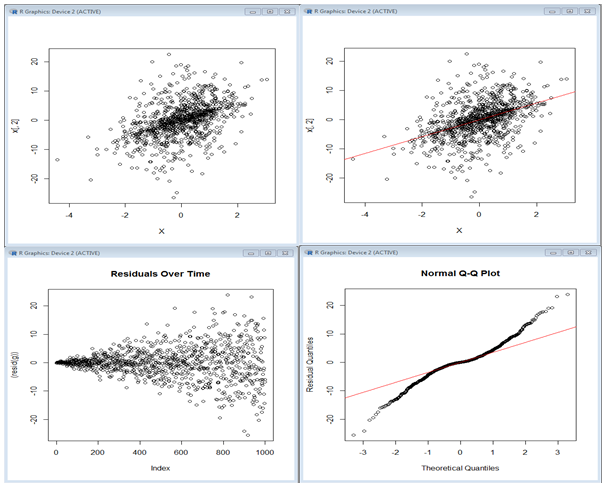


**[END LIST]**

In this code, we have made the variance of the true model increase over time by creating a third column containing increasing values (in increments of 0.05) and creating an error term in our model with mean 0 and variance equal to the third column. Running a regression on these data, we get the diagnostic plots in exhibit 11.14. The plot of residuals over time suggests that the variation in residuals is changing over time; the results become less accurate (bigger residuals) over time. The Q–Q plot also shows a violation of normal assumptions. The variance of the error terms has increased over time.

**[INSERT EXHIBIT]**

**Exhibit 11.14** Diagnostics for Data with Heteroscedasticity



**[END EXHIBIT]**

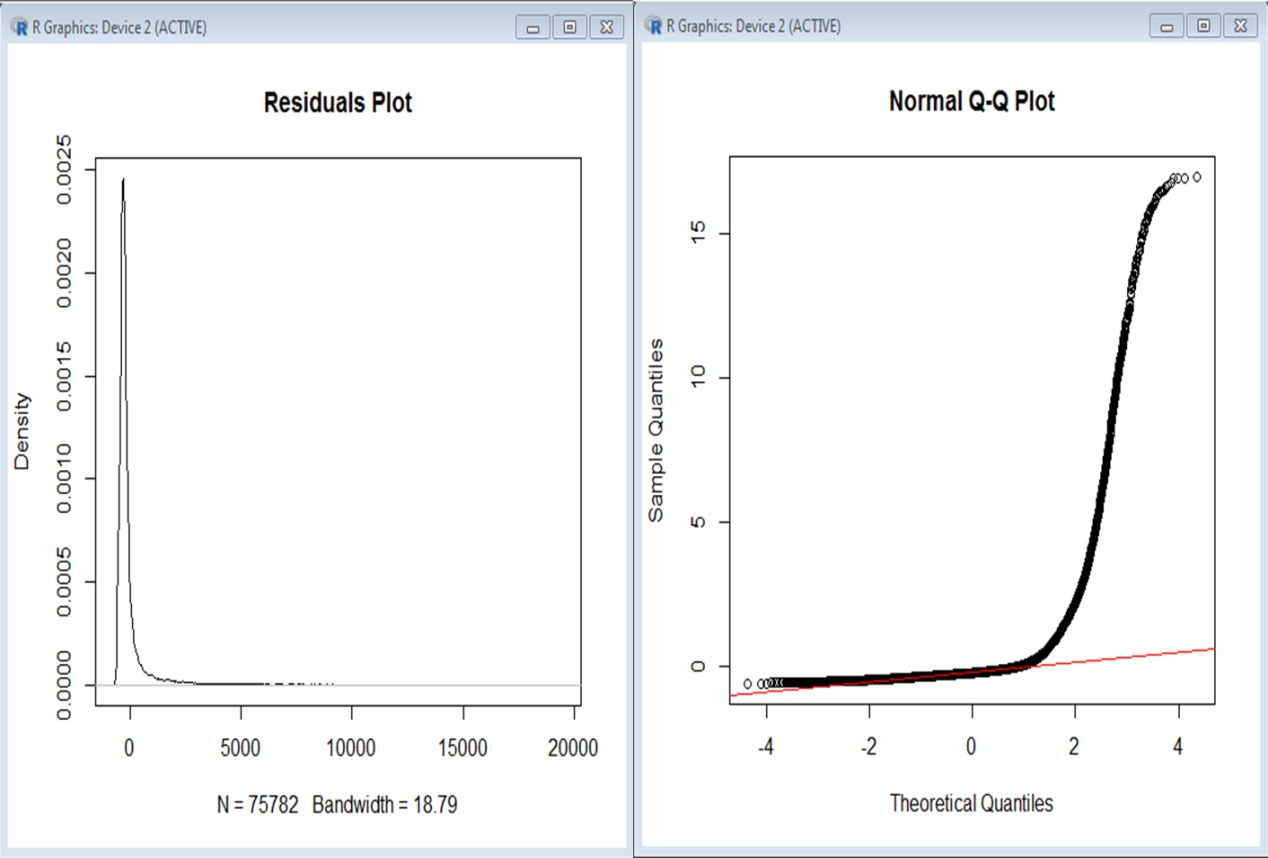
If statistical tests are desired, the continuous changes in variance can be tested by something called a *White test* or a *Breush test*, whereas the discrete changes in variance can be tested with the *Goldfeld-Quandt test*. The White test is generally used for large samples and does not require much input from the analyst. The Breush test requires the analyst’s input on which independent variables are included—otherwise, it is similar to the White test. The Goldfeld-Quant test is used to compare the differences in error terms across discrete subgroups in the data. The analyst has to assign or create the subgroups. Often there is no need to calculate specific statistics—a quick look at the diagnostic plots will disclose all we need to know.

# [H1] Regression Assumptions: Normally Distributed Errors

Ordinary regression assumes that the error term is normally distributed. This can be visually depicted in a normal quintile plot of the residuals. In these plots, quintiles of the observed data are plotted against quintiles in a standard normal distribution. Exhibit 11.15 shows an example of normal probability plots for the cost variable in our MFH example. It shows a long, asymmetric tail for the density of the residuals. If it were normally distributed, we would see a symmetric density function. The Q–Q plot also shows radical departures from normality. A quick look shows that the quartiles do not fall where normal distribution quartiles would. Clearly our effort to regress cost on age or other variables is not reasonable, as many assumptions, including the normal distribution of errors, are violated.

**[INSERT EXHIBIT]**

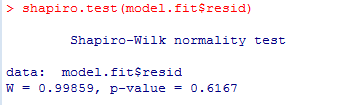
**Exhibit 11.15** Normal Probability Plot for Cost Data in MFH Example

****

**[END EXHIBIT]**

A number of statistics are also available to test normality: the Kolmogorov-Smirnov, the Shapiro-Wilk, the Jarque-Bera, or the Anderson-Darling tests can be used (Horber 2018). For example, the Shapiro-Wilk test of normality can be carried out using the following test.

**[LIST FORMAT]**



**[END LIST]**

# [H1] Transformation of Data to Remedy Model Violations

If the assumption of normal distribution of error is violated, then data transformations can be used to remedy the situation (van Emden 2008). The log transformation is carried out by adding 1 to the observed value and then converting it to the logarithm with base ten. The addition of 1 to the observed values is a precautionary measure that allows us to take the log of zero values in the data. The following R code shows the transformation of cost data in the MFH example using log transformation. We begin by transforming our cost-per-day variable by applying the log() function to the data. In Excel, we would transform the data using the Ln function.

**[LIST FORMAT]**

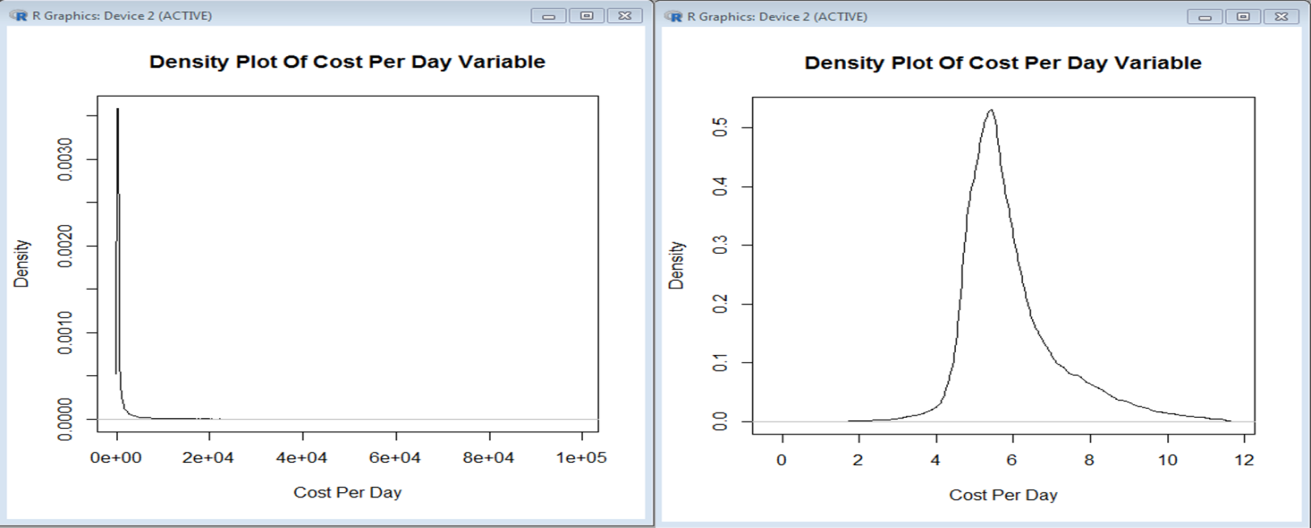


**[END LIST]**

To see what effect this transformation has had on the data, exhibit 11.16 shows a density plot of the data before (on the left) and after transformation (on the right). The transformation has reduced the long tail and made the shape look more symmetric, like a normal distribution.

**[INSERT EXHIBIT]**

**Exhibit 11.16** Density Function of Cost Data Before and After Logarithm Transformation

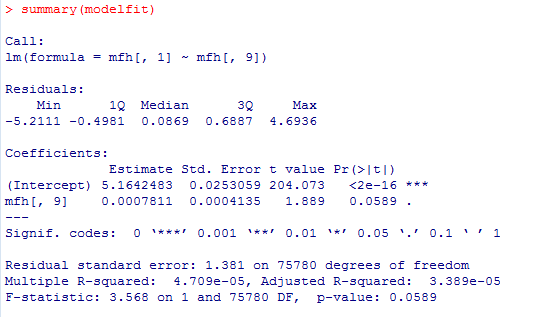


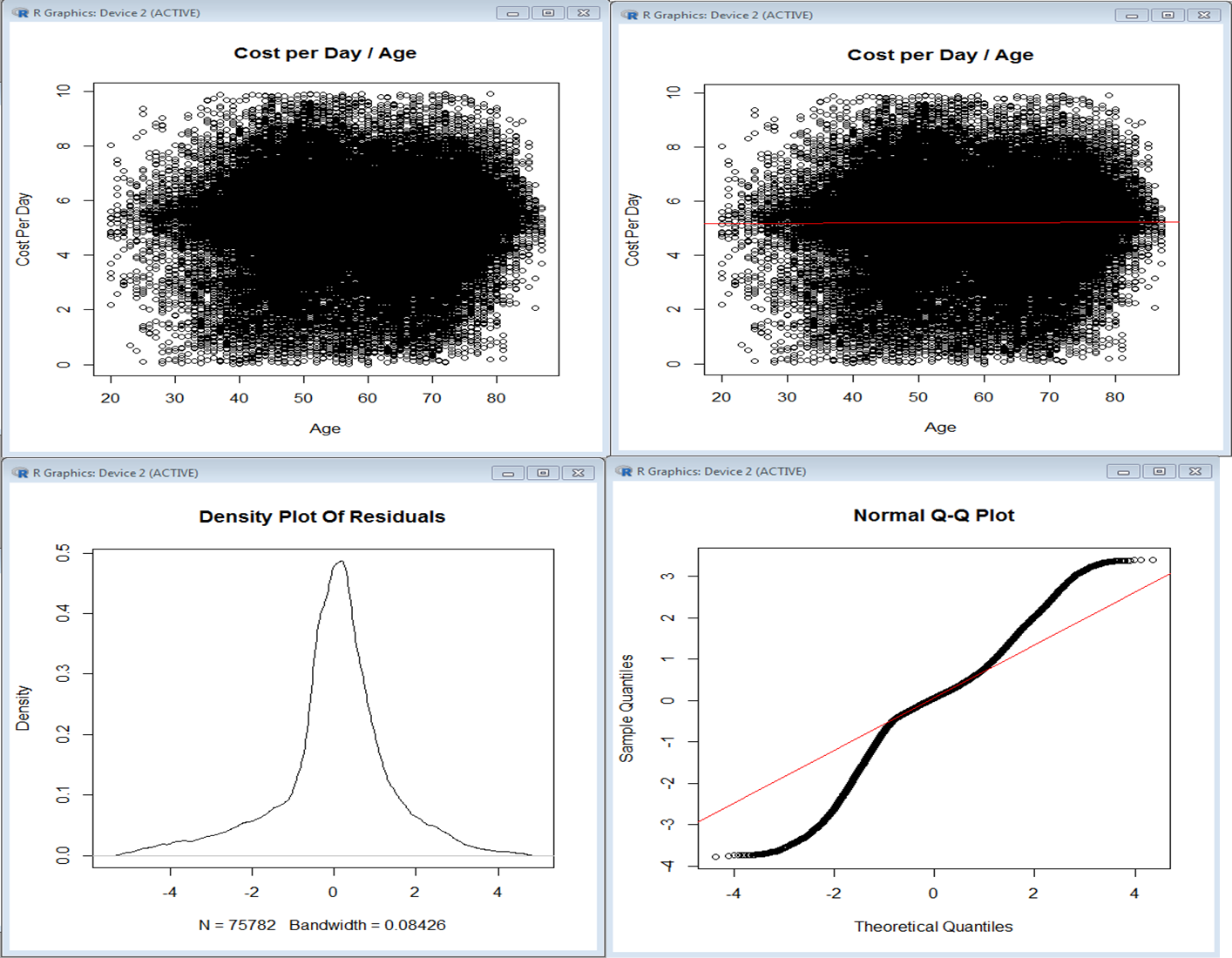
**[END EXHIBIT]**

Next we perform a simple regression of cost per day on age, just as we had done previously, and we obtain the diagnostic plots in exhibit 11.17. This image also shows the regression summary. Although the cost data were made to look more normal, the residuals appear to be still nonnormal. The Q-Q plot shows large violations of normal distribution. The plot of cost versus age suggests that there may be two different sets of data; at low ages, we have one spread of residuals, and at higher ages, we have another, as if there are two overlapping circles. We might need to create two models for the relationship between cost and age, one for younger people and one for older people.

**[INSERT EXHIBIT]**

**Exhibit 11.17** Effect of Log Transformation on Regression of Cost on Age





**[END EXHIBIT]**

Another way to transform data is to square the independent variable. X-squared moderately affects the shape of the error distribution and is generally used to address the left skewness. This is a practical transformation to use when the aim is to fit a response by a quadratic function and when the variable or variables concerned are either zero or positive (Cox 2007).

Square root transformation is appropriate for situations in which the variance of the data is greater than the mean. This usually occurs when observations are clumped. Again, a constant needs to be added to each observed value to make it positive, or the square root transformation cannot be applied. Angular transformation becomes valuable when the percentage data consist of extreme ends of the spectrum—that is, when the percentage data are either very high or very low. The angular transformation normalizes skewed percentage data.

# [H1] Effects of Collinearity

*Collinearity* is the term used to describe strong correlation among some of the independent variables in a regression model. Collinearity affects the variance of the regression coefficient; therefore, in the presence of collinearity, the regression coefficients change. A simple view of relationships between two variables assumes that one variable affects another. In regression, this simple view is often proven wrong. The effect of two variables on each other could depend on yet another variable inside the regression equation. For example, in predicting the cost of care, the effect of aging on cost could depend on gender, a different variable in the equation. If two independent variables (*x*1 and *x*2) are correlated, then the impact of each on the outcome *y* depends on the other. In other words, if the two variables have strong correlations, then one of these variables can change a significant relationship between the other variable and the outcome to an insignificant relationship. In short, the impact of variables depends on other variables that are also in the regression equation. In this sense, relationships found in regression are context dependent. With one set of variables, the relationship may exist; with another, it may vanish.

The existence of collinearity can be anticipated by calculating the correlation matrix (the correlation between any two pairs of independent variables). It is not always clear what to do with collinearity. One could drop one of the two correlated variables, but that may reduce predictive accuracy, and it may also make models that are sensitive to missing values. The analyst could also continue with both variables in the model, assuming that the observed coefficients may be affected.

# [H1] Importance of Cross-Validation

Cross-validation is the process by which a portion of the data is set aside and the remaining data can be used for model training. The data that were set aside will be used for testing the model. This usually gives the analyst a sense of the model’s effectiveness when dealing with unseen data—in other words, it lets the analyst know how the model will perform when using real-world, unseen data. It also protects against modeling the noise in the training data; as such, a model will perform poorly in cross-validation. There are various methods of cross-validation; the following list contains a few.

**[INSERT BL]**

* *Validation set*. In this approach, a portion of the data is used for model training. The parameters of the regression model are estimated from the training data set. The remaining portion is reserved for validation. When data are separated into training and validation sets, the impact of rare events becomes harder to capture.
* *K-fold cross-validation*. This approach randomly splits the data set into k-folds; for each fold, the model would train on k – 1 folds of the data and the kth fold is reserved for testing the model effectiveness. The errors seen in each prediction are recorded. This process is repeated until all k-folds have served as tests.
* *Leave-one-out cross-validation*. In this approach, only one data point of the entire data set is reserved for testing. The remaining data set is used for training the model. The advantages of this type of cross-validation include low bias. However, because the analyst uses the majority of the data for model training, it reports higher cross-validated accuracy.

**[END BL]**

# [H1] Weighted Regression

Instead of treating each observation equally, the weighted regression gives each observed value an appropriate value through which it affects the parameter estimation by the model. Excel does not provide an easy method for weighting the data. The following R code shows how to run a weighted regression. We show the performance of weighted regression in simulated data that we have called *simdata*. We can perform a weighted regression in R by simply telling the lm() function where the weights are:

**[LIST FORMAT]**



**[END LIST]**

The syntax is reserved word weights equals a column of data (here, the third column in the simdata data frame).

Weighted regression is useful in the presence of heteroscedasticity, which occurs if the variance of the dependent variable differs across different values of independent variables. In other words, the standard deviation of error terms is not constant across different values of independent variables. In such scenarios, weighted regression assigns each observed value a relative weight based on its variance. The bigger the variance, the smaller the weight. Weighted regression is not associated with any specific function and can therefore be used with any linear or nonlinear function; it adds nonnegative constants to observed values of independent variables for each data point. The value of the constant, or *weights*, is inversely proportional to the variance at each observed data point (MedCalc 2019; Croarkin and Tobias 2019).

In later chapters, weighted regression is used to remove confounding that occurs when patients select treatment (i.e., when there is no randomization). Weights are chosen to remove the effects of certain covariates or alternative explanations for change in outcome.

# [H1] Shrinkage Methods and Ridge or Lasso Regression

When there are many variables in the regression model, multicollinearity is a major concern, and removal of some variables from the analysis could be helpful in reducing misspecified models. Ridge or lasso regression are methods of restricting the number of variables examined in the regression model. In these regressions, small effect sizes are ignored. In massive EHR data, all variables—even those with a small effect—will have statistical significance. A large number of statistically significant variables is what leads to paralysis. By focusing on large effects, ridge or lasso regressions help make the model more practical.

# [H1] Context-Specific Hypothesis Testing

Regression parameters are estimated from a sample of data. Even if multiple samples are taken from the same population, the estimates for the parameters will be different for each sample. This sampling variability gives rise to the standard error associated with each regression parameter—namely the intercept and slope coefficients. The common way to address the uncertainty with estimates is to develop 95 percent confidence intervals and *p*-values. The intercept (which represents the estimated mean) and slope (the mean differences) are normally distributed for large samples with more than 60 observations. The 95 percent confidence interval for the intercept is given by the equation

**[INSERT EQUATION]**

.

**[END EQUATION]**

Similarly, the 95 percent confidence interval for the regression coefficient of an independent variable is given by the formula

**[INSERT EQUATION]**

**[END EQUATION]**

Regression coefficients and their associated standard error can be used to test hypotheses. The null hypothesis states that the real population-level mean difference,is equal to 0, whereas the alternate hypothesis would be that this difference is not 0:

**[INSERT EQUATION]**

**[END EQUATION]**

For the coefficient, assuming the null hypothesis to be true, the standardized *t*-statistic is calculated with the equation

**[INSERT EQUATION]**

**[END EQUATION]**

The *t*-statistic calculates the number of standard errors the estimated coefficient is away from 0. Testing hypotheses using regression coefficients may lead to findings that are different from the direct tests of hypothesis covered in chapters 4 and 5. This chapter has already mentioned how the coefficients of the regression equation are affected by the set of variables in the equation. When certain variables are present, a hypothesis may be rejected, but when these variables are absent, the same hypothesis may be accepted. In short, regression provides a context-specific test of hypotheses.

# [H1] Changing Units of Measurement

A regression equation is an equation like any other, analysts should pay attention to units of measurement. Changing the unit of measurement for a variable will lead to a corresponding change in coefficients and standard error but will not change the statistical significance of the coefficients or the interpretation of the model. Changing the units of an independent variable is the same as multiplying the independent variable by a constant. The corresponding regression coefficients will be divided by the constant. If the outcome *y* is multiplied by a constant, all regression coefficients will be multiplied by the same constant. This technique does not, however, affect the statistical significance of the relationship between the independent and dependent variables.

# [H1] Interpretation of Regression as Cause and Effect

Many investigators use a structural model, in which multiple simultaneous regressions are assumed, to model causes and effects. Strictly speaking, regression is a test of association, so its use in causal modeling requires additional assumptions. The cause-and-effect interpretation requires four assumptions:

**[INSERT NL]**

1. the independent variables must be associated with the dependent variable, showing a relatively large coefficient of determination;
2. the independent variables must be measured before the dependent variable;
3. there should be a hypothesized mechanism between the independent variables and the dependent variable; and
4. the effect of independent variables should be measured on the dependent variable after removing confounding.

**[END NL]**

The first assumption can be verified through regression. The second assumption could be arranged by the judicious use of periods to observe dependent and independent variables. The other two are more difficult to verify or arrange ahead of time. Chapter 20 shows how regression analysis can be used to identify causal networks.

# [H1] Chapter Summary

In this chapter, we covered multiple regression, with a special emphasis on diagnostic tools. You should be able to use software to make a regression equation. To do this well, there are six requirements, all of which must be mastered.

**[INSERT NL]**

1. You should be able to check the assumptions of regression and transform data to meet these assumptions.
2. You should be able to run a regression using the software of your choice. Excel was used in this chapter, but R code is introduced in the appendix. You can also do regressions in SAS, STATA, and other software.
3. You should be able to interpret the findings from regression, including how well the regression equation fits the data and whether specific variables have a statistically significant relationship with the dependent variable.
4. You should be able to understand interactions among variables.
5. You should be able to understand the effect of collinearity on regression parameters and why it would make the test of regression parameters a context-specific test.
6. You should be able to use diagnostic plots to figure out how data should be transformed.

**[END EXHIBIT]**

Regression is widely used and fundamental to the rest of this book, so it is important that you practice it again and again until you understand it thoroughly.

## [H1] Supplemental Resources

Problem set, solutions to problems, multimedia presentations, SQL code, and other related material are in the course website

**[H1] REFERENCES**

Alemi, F., M. J. Atherton, D. C. Pattie, and M. Torii. 2013. “Continuously Rethinking the Definition of Influenza for Surveillance Systems: A Dependent Bayesian Expert System.” *Medical Decision Making* 33 (6): 860–68.

Alemi, F., M. Torii, M. J. Atherton, D. C. Pattie, and K. L. Cox. 2012. “Bayesian Processing of Context-Dependent Text: Reasons for Appointments Can Improve Detection of Influenza.” *Medical Decision Making* 32 (2): E1–9.

Boucquemont, J., G. Heinze, K. J. Jager, R. Oberbauer, and K. Leffondre. 2014. “Regression Methods for Investigating Risk Factors of Chronic Kidney Disease Outcomes: The State of the Art.” *BMC Nephrology* 15: 45.

Carey, K., and T. Stefos. 2011. “Measuring the Cost of Hospital Adverse Patient Safety Events.” *Journal of Health Economics*. 20 (12): 1417–30.

Chen, S., Y. J. Goo, and Z. D. Shen Z. 2014. “A Hybrid Approach of Stepwise Regression, Logistic Regression, Support Vector Machine, and Decision Tree for Forecasting Fraudulent Financial Statements.” *Scientific World Journal* article ID 968712. www.hindawi.com/journals/tswj/2014/968712/.

Clyde, A. T., D. Hemenway, and J. T. Nagurney. 1996. “Seat Belt Use, Insurance Status, and Hospital Bad Debt.” *Journal of Trauma* 41 (1): 100–104.

Cox, N. J. 2007. “Transformation: An Introduction.” Durham University. Last revised July 25. <http://fmwww.bc.edu/repec/bocode/t/transint.html>.

Croarkin, C., and P. Tobias, (eds.). 2019. “Weighted Least Squares Regression.” *Engineering Statistics Handbook*. NIST/SEMATECH. Accessed April 23. www.itl.nist.gov/div898/handbook/pmd/section1/pmd143.htm.

Extermann, M. 2007. “Interaction Between Comorbidity and Cancer.” *Cancer Control* 14 (1): 13–22.

Gustafson, D. H., F. Sainfort, M. Eichler, L. Adams, M. Bisognano, and H. Steudel. 2003. “Developing and Testing a Model to Predict Outcomes of Organizational Change.” *Health Services Research* 38 (2): 751–76.

Healthcare Cost and Utilization Project. 2017. “Clinical Classifications Software (CCS) for ICD-9-CM.” Agency for Healthcare Research and Quality. Published March. www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

Horber, E. 2018. “Analysis of Residuals.” University of Geneva. Accessed January 7, 2019. www.unige.ch/ses/sococ/cl//stat/action/analyse\_residuals0.html.

Kheirbek, R. E., F. Alemi, and M. Zargoush. 2013. “Comparative Effectiveness of Hypoglycemic Medications Among Veterans.” *Journal of Managed Care and Specialty Pharmacy*. 19 (9): 740–44.

Maskawa, J. 2016. “Collective Behavior of Market Participants During Abrupt Stock Price Changes.” *PLoS One* 11 (8): e0160152.

MedCalc. 2019. “Weighted Regression: Worked Example.” Retrieved January 7. www.medcalc.org/manual/weighted-regression-worked-example.php.

Ryan, A. M., S. M. Shortell, P. P. Ramsay, and L. P. Casalino. 2015. “Salary and Quality Compensation for Physician Practices Participating in Accountable Care Organizations.” *Annals of Family Medicine* 13 (4): 321–4.

Shafer, P.R., K. C. Davis, D. Patel, R. Rodes, and D. Beistle. 2016. “Association Between Media Dose, Ad Tagging, and Changes in Web Traffic for a National Tobacco Education Campaign: A Market-Level Longitudinal Study.” *Journal of Medical Internet Research* 18 (2): e39.

Shimada, S. L., A. M. Zaslavsky, L. B. Zaborski, A. J. O'Malley, A. Heller, and P. D. Cleary. 2009. “Market and Beneficiary Characteristics Associated with Enrollment in Medicare Managed Care Plans and Fee-for-Service.” *Medical Care* (5): 517–23.

van Emden, H. 2008. *Statistics for Terrified Biologists*. New York: Wiley Blackwell.

Weil, T. P. 2002. “Multispecialty Physician Practices: Fixed and Variable Costs, and Economies of Scale.” *Journal of Ambulatory Care Management* 25 (3): 70–77.

Yount, K. W., J. M. Isbell, C. Lichtendahl, Z. Dietch, G. Ailawadi, I. L. Kron, J. A. Kern, and C. L. Lau. 2015. “Bundled Payments in Cardiac Surgery: Is Risk Adjustment Sufficient to Make It Feasible?” *Annals of Thoracic Surgery* 100 (5): 1646–52.

# Appendix: Regression Using R

## [H1] Introduction

Previous chapters have used Excel to analyze and plot data. When it comes to regression analysis, Excel has widely available tools for conducting linear regression but fewer tools for logistic or Poisson regression. Furthermore, Excel cannot handle massive data, the type of data often found in EHRs. For this reason, chapter 11 initiates a switch from Excel to R. R is a widely available and free software that has extensive regression capabilities.

Using new software is always problematic as users learn new commands. The R environment is quite flexible, however, and the lessons learned will be of direct use in future projects. To orient the reader to R, this section shows (a) how to download and install R, (b) how to read the data used throughout this chapter into R, (c) how to recode null values, and (d) how to examine a summary of the data. Once these basic data manipulations are covered, we show how to plot data and do the regressions that were reported in the chapter. With both R and Excel skills, you have a broader set of options.

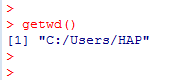
## [H1] Download

R is open-source software that can be installed after clicking the following link: <https://cran.r-project.org/bin/windows/base/old/3.3.1/>.

## [H1] Working Directory

Before reading data into R, we need to check that we have set our working directory properly. In R, this can be achieved by using the “getwd()” command. It provides the working directory for R. In our computer, it resulted in the following:

**[LIST FORMAT]**



**[END LIST]**

This means that R is currently set up to look for files in the “Users/HAP” directory on our machine. Let us assume that our data is stored in the “Users/HAP/Documents” directory. To change the R working directory, we need to use the “setwd()” command as follows:

**[LIST FORMAT]**



**[END LIST]**

## [H1] Reading Comma-Separated Variable (CSV) Files

Our data file is titled “MFHcleanedup.csv,” and to read this into R we need to use the read.csv() command.

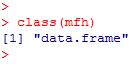
**[LIST FORMAT]**



**[END LIST]**

This code tells R to read the data into an object called “mfh” in our R workspace. We can see what kind of object “mfh” is by using the class() function, like this:

**[LIST FORMAT]**



**[END LIST]**

Now we know that we have set up a data frame. A data frame has all the data needed for analysis in a matrix format, with headings and related information.

## [H1] Recoding Data

To get rid of missing and not available (“NA”) values, use the na.omit() function.

**[LIST FORMAT]**



**[END LIST]**

To recode NULL values, for example to turn them into 0s and 1s, we can use a simple loop. As an example, to recode NULL values in the fifth column of our data frame, we can use the following:

**[LIST FORMAT]**



**[END LIST]**

Here we have told R to loop through the data and to change any NULL elements to 0.

## [H1] Summary Data

To view a summary of the data, we need to use the summary() function.

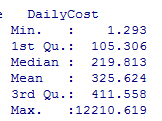
**[LIST FORMAT]**



**[END LIST]**

This function will print out a summary of each variable (column) in the data. As an example, in our data, the second variable is DailyCost. R will return the following:

**[LIST FORMAT]**



**[END LIST]**

This information tells us that the DailyCost variable takes a minimum value of 1.293, a maximum value of 12,210, and a mean value of 325.624.

## [H1] Plot and Density Functions

We can also plot the probability density of variables in R. For example, we can use the plot() and density() functions to plot the density of the Days Survived variable in R (see exhibit A.1).

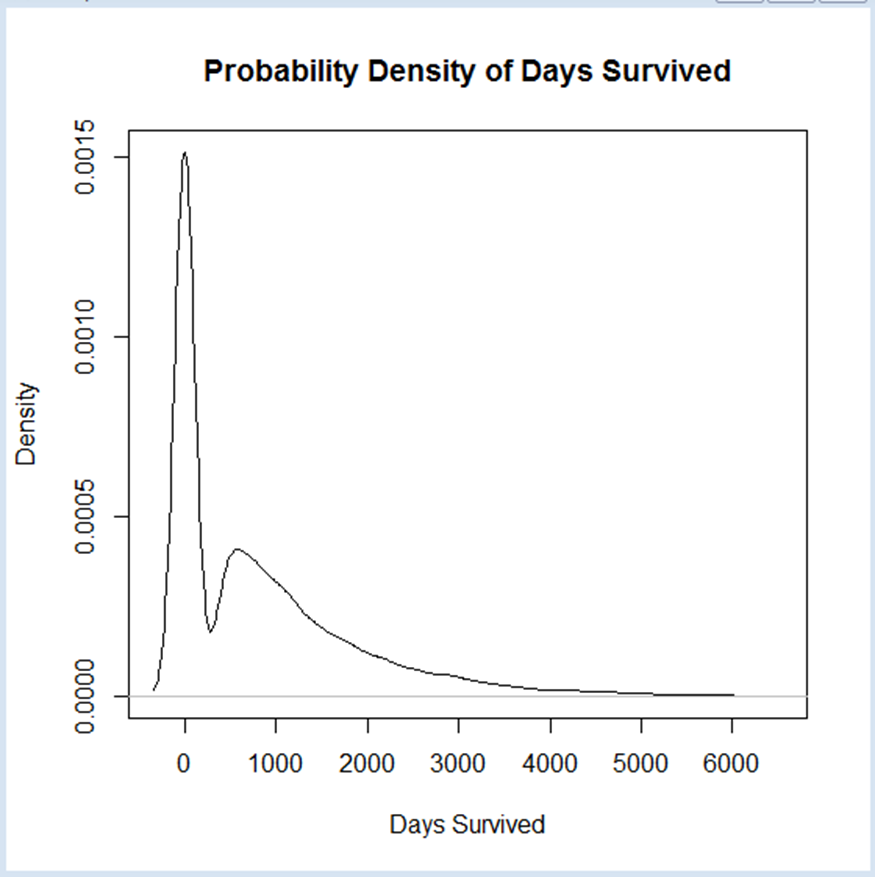
**[LIST FORMAT]**



[END LIST]

**[INSERT EXHIBIT; if you could RENDER THIS SO THAT THE LETTERS AND NUMBERS LOOK LESS DISTORTED, THAT WOULD BE GREAT]**

**Exhibit A.1** Probability Density of Days Survived



## [END EXHIBIT]

## [H1] Errors in R Code

The formats for commands used in R are provided online. When you are not sure how to do something in R, the best remedy is to search the web. Do the same when you get an error message—chances are, someone else has already received the same message and found a solution for it.

## [H1] Referring to Variables by Numbers

Data frames can be manipulated using columns and row references. To know what column number refers to a specific variable, one needs to use the function colnames(), which looks like

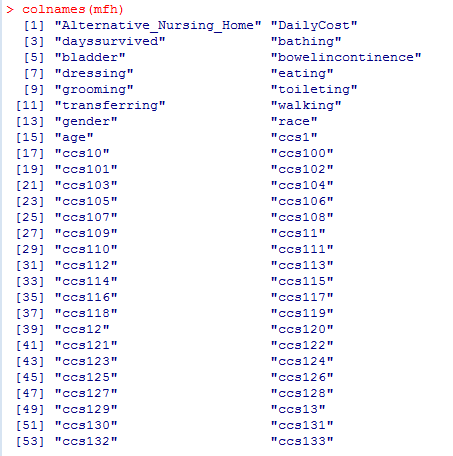
**[LIST FORMAT]**



**[END LIST]**

The colnames command gives the column names in a data frame. The only parameter needed for this command is the data frame name. R produces the following output:

**[LIST FORMAT]**



**[END LIST]**

As we can see, the variable DailyCost is column 2, and the variable “age” is column 15. To produce our scatterplot, we need to use the function plot() as described in the following section.

## [H1] Plotting in R

R has a number of standard plotting routines, including basic scatter plots and classic linear regression diagnostics. To create a scatter plot, we first need to determine the column names of the variables Cost and the variable Age. Then we refer to these columns in the function plot().

**[LIST FORMAT]**

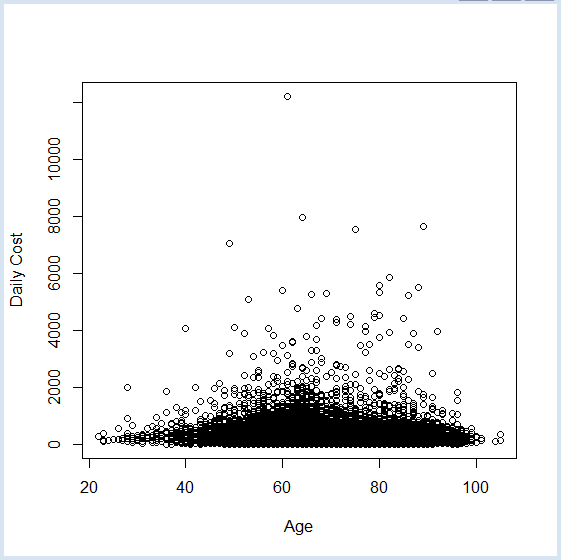


**[END LIST]**

In the plot command, we first give the *x*-variable to be plotted, then the *y*-variable. These variables are defined relative to the position of the variable in the data frame. So, mfh[,15] tells the computer to use the fifteenth column in the data frame mfh. Similarly, mfh[,2] tells the computer to use the second column from data frame mfh. The parameters xlab and ylab provide the labels for the *x*- and *y*-axes. In the command, we tell R to plot column 15 along the *x*-axis and variable 2 along the *y*-axis. The outcome of this code can be seen in exhibit A.2. These data suggest that daily cost and age may not have a strong relationship. This is the same conclusion we drew using Excel.

**[INSERT EXHIBIT]**

**Exhibit A.2** Outcome



## [END EXHIBIT]

## [H1] Ordinary Regression

To perform ordinary linear regression we need to use the linear model function, lm(), as follows:

**[LIST FORMAT]**



**[END LIST]**

This line of code tells R to regress variable 2 on variable 15 and to store the results of the regression in an object titled model.fit. The variable to the left of ~is the dependent or response variable. The variable to the right of ~ is the independent variable. We can view the results of the regression using the summary() function:

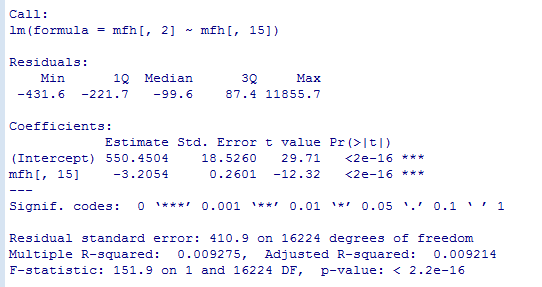
**[LIST FORMAT]**



**[END LIST]**

The syntax of summary command is the reserve word “summary” followed by the model name inside parentheses. This command returns the following summary of the regression results:

**[LIST FORMAT]**



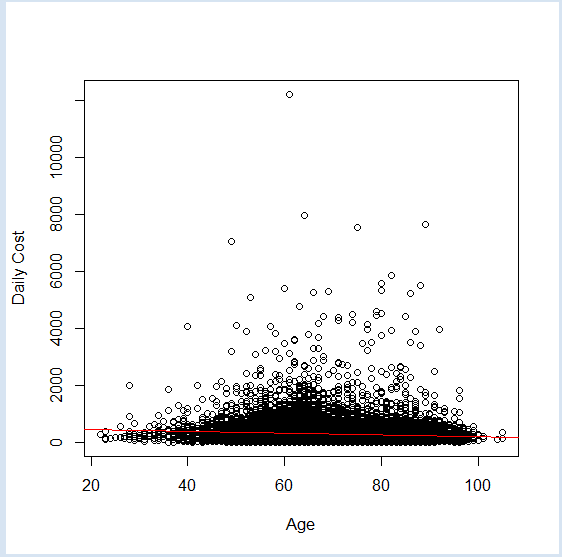
**[END LIST]**

The summary report starts by giving the formula of the variables examined. It then gives the distribution of the residuals, which is the difference of predicted and actual values. This distribution is given by providing the minimum, first quartile, median, third quartile, and maximum residual values. Next, the summary provides the estimated parameters for the equation. Here two parameters are given, the intercept and the coefficient for age (variable in column 15 of the data frame). The statistical significance of the coefficient is indicated with a number of stars; the higher the number of stars, the greater the significance. In the last three lines of the summary, residual standard error, degrees of freedom, adjusted R-squared, and F-statistics are provided. The adjusted R-squared is a measure of goodness of fit. The *p*-value for the F‑statistic shows the probability of obtaining the fit randomly. The lower the *p*-value, the better the fit, and the more significant the findings. Note that the intercept constant, the coefficient for age, and the overall fit to the model are all statistically significant. Yet, the fit is not high. Most of the variation in cost remains unexplained.

R also allows us to fit a regression line to the plot we produced earlier (see exhibit A.3). This can be done using the abline() function.

**[INSERT EXHIBIT]**

**Exhibit A.3** Regression Line



**[END EXHIBIT]**