# Applied Survey Data Analysis

## Second Edition

Steven G. Heeringa, Brady T. West, and Patricia A. Berglund


CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742




CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed on acid-free paper

International Standard Book Number-13: 978-1-4987-6160-4 (Hardback)



**Library of Congress Cataloging-in-Publication Data**

Names: Heeringa, Steven, 1953- author. | West, Brady T., author. | Berglund, Patricia A., author.
Title: Applied survey data analysis / Steven G. Heeringa, Brady T. West, Patricia A. Berglund.
Description: Second edition. | Boca Raton, FL : CRC Press, [2017] | Includes bibliographical references and index.
Identifiers: LCCN 2016050459| ISBN 9781498761604 (hardback cover) | ISBN 9781498761611 (e-book)
Subjects: LCSH: Social sciences--Statistics. | Social surveys--Statistical methods.
Classification: LCC HA29 .H428 2017 | DDC 001.4/22--dc23
LC record available at https://lccn.loc.gov/2016050459

**Visit the Taylor & Francis Web site at**
**http://www.taylorandfrancis.com**

**and the CRC Press Web site at**
**http://www.crcpress.com**

Printed and bound in the United States of America by Sheridan

# 7

# *Linear Regression Models*

## 7.1 Introduction

> Study regression. All of statistics is regression.

This quote came as a recommendation from a favorite professor to one of the authors while he was in the process of choosing a concentration topic for his comprehensive exam. The broader interpretation of the quote requires placing the descriptor in quotes, "regression"; but ask individuals with backgrounds as varied as social science graduate students or quality control officers in a paper mill to decipher the statement and they will think first of the linear regression model. Given the importance of the linear regression model in the history of statistical analysis, the emphasis that it receives in applied statistical training and its importance in real world statistical applications, the narrower interpretation is quite understandable.

This chapter introduces linear regression modeling for complex sample survey data—its similarities and how it differs (theoretically and procedurally) from standard ordinary least squares (OLS) regression analysis. We assume that the reader is familiar with the basic theory and methods for simple (single predictor) and multiple (multiple predictor) linear regression analysis for continuous dependent variables. Readers interested in a comprehensive reference on the topic of linear regression are referred to Draper and Smith (1981), Kleinbaum et al. (1988), Neter et al. (1996), DeMaris (2004), Faraway (2014), Fox (2008), or many other excellent texts on the subject.

Focusing on practical approaches for complex sample survey data, we emphasize "aggregated" *design-based approaches* to the linear regression analysis of survey data (sometimes referred to as *population-averaged modeling*), where design-based variance estimates for weighted estimates of regression parameters in *finite* populations are computed using nonparametric methods such as the Taylor series linearization (TSL) method, balanced repeated replication (BRR), jackknife repeated replication (JRR), or bootstrapping. *Model-based approaches* to the linear regression analysis of

complex sample survey data, which may explicitly include fixed effects of sampling strata and/or random effects of sampling clusters in the regression models and may or may not utilize the sampling weights (e.g., Skinner et al., 1989; Pfeffermann et al., 1998; Little, 2003; Pfeffermann, 2011), are introduced in Chapter 13. Over the years, there have been many contributions to the survey methodology literature comparing and contrasting these two approaches to the regression analysis of survey data, including papers by DuMouchel and Duncan (1983), Hansen et al. (1983), and Kott (1991).

We present a brief history of important statistical developments in linear regression analysis of complex sample survey data to begin this chapter. Kish and Frankel (1974) were two of the first to empirically study and discuss the impacts of complex sample designs on inferences related to regression coefficients. Fuller (1975) derived a linearization-based variance estimator for multiple regression models with unequal weighting of observations and introduced variance estimators for estimated regression parameters under stratified and two-stage sampling designs. Shah et al. (1977) further discussed the violations of standard linear model assumptions when fitting linear regression models to complex sample survey data, discussed appropriate methods for making inferences about linear regression parameters estimated using survey data, and presented an empirical evaluation of the performance of variance estimators based on TSL.

Binder (1983) focused on the sampling distributions of estimators for regression parameters in finite populations and defined related variance estimators. Skinner et al. (1989, Sections 3.3.4 and 3.4.2) summarized estimators of the variances for regression coefficients that allowed for complex sample designs (including linearization estimators) and recommended the use of linearization methods or other robust methods (such as JRR) for variance estimation. Kott (1991) further discussed the advantages of using variance estimators based on TSL for estimates of linear regression parameters: protection against correlated random errors within primary sampling units (PSUs), protection against possible nonconstant variance of the random errors, and the fact that a within-PSU correlation structure does not need to be identified to have a nearly unbiased estimator. Fuller (2002) provided a modern summary of regression estimation methods for complex sample survey data, and Pfeffermann (2011) discussed the relative merits of a variety of possible design-based and model-based approaches to fitting linear regression models to complex sample survey data, presenting empirical support for the use of a "q-weighted" method. In this method, the modified sampling weights used to compute estimates of the regression coefficients are the original probability weights divided by the *expectations* of the weights as a function of the covariates in the model of interest (i.e., predicted values of the weights based on a model regressing the weights on the predictor variables in the model of interest). We consider this approach in the illustration presented later in this chapter.

## 7.2 Linear Regression Model

Regression analysis is a study of the relationships among variables: a *dependent variable* and one or more *independent variables.* Figure 7.1 illustrates a simple linear regression model of the relationship of a dependent variable, $y$, and a single independent variable $x$. The regression relationship among the observed values of $y$ and $x$ is expressed as a regression model, for example, $y = \beta_0 + \beta_1 x + \varepsilon$, where $y$ is the dependent variable, $x$ is the independent variable, $\beta_0$ and $\beta_1$ are model parameters, and $\varepsilon$ is an error term that reflects the difference between the observed value of $y$ and its conditional expectation under the model, $\varepsilon = y - \hat{y} = y - \beta_0 - \beta_1 x$.

In statistical practice, a fitted regression model may be used to simply predict the expected outcome for the dependent variable based on a vector of independent variable measurements $x$, $E(y \mid x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$, or to explore the functional relationship of $y$ and $x$. Across the many scientific

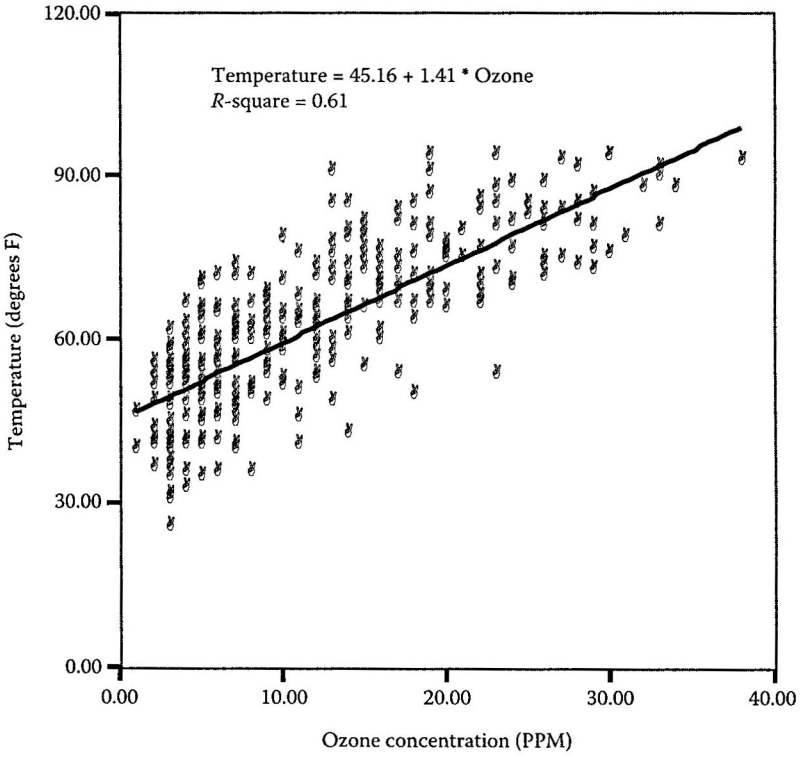


**FIGURE 7.1**
Linear regression of air temperature on ozone level. (From Breiman and Friedman, 1985).

disciplines that use regression analysis methods, dependent variables may also be referred to as response variables, regressands, outcomes, or even "left-hand side variables." Independent variables may be labeled as predictors, regressors, covariates, factors, cofactors, explanatory variables, or "right-hand side variables." We primarily refer to response variables and predictor variables in this chapter, but other terms can be used interchangeably.

This chapter will focus on the broad class of regression models known as *linear models*, or models for which the conditional expectation of $y$ given $x$, $E(y|x)$, is a linear function of the unknown parameters. Consider the following three specifications of linear models:

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{7.1}$$

Note in this model that the dependent variable, $y$, is a linear function of the unknown parameters and the independent variable $x$.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon. \tag{7.2}$$

In this model (7.2), the response variable $y$ is still a linear function of the $\beta$ parameters for $x$ and $x^2$; however, the linear model defines a *nonlinear* relationship between $y$ and $x$.

$$\begin{aligned} y &= \boldsymbol{x}\boldsymbol{\beta} + \varepsilon \\ &= \sum_{j=0}^{p} \beta_j x_j + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon. \end{aligned} \tag{7.3}$$

Here, the linear model is first expressed in *vector notation*. Vector notation may be used as an abbreviation to represent a complex model with many parameters and to facilitate computations using the methods of matrix algebra.

When specifying linear regression models, it is useful to be able to reference specific observations on the subjects in a survey data set:

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i, \tag{7.4}$$

where:

$\boldsymbol{x}_i = [1 \quad x_{1i} \quad \ldots \quad x_{pi}]$ and $\boldsymbol{\beta}^T = [\beta_0 \quad \beta_1 \quad \ldots \quad \beta_p]$.

In this notation, $i$ refers to sampled element (or respondent) $i$ in a given survey data set.

### 7.2.1 Standard Linear Regression Model

Standard procedures for unbiased estimation and inference for the linear regression model involve the following assumptions:

1. The model for $E(y|x)$ is linear in the parameters (Equation 7.2).
2. Correct model specification—in short, the model includes the predictor variables and regression coefficients that accurately reflect the true model under which the data were generated.
3. $E(\varepsilon_i|\boldsymbol{x}_i) = 0$, or the expected value of the residuals given a set of values on the predictor variables is equal to 0.
4. Homogeneity of Variance—$var(\varepsilon_i|\boldsymbol{x}_i) = \sigma^2_{y\cdot x}$, or the variance of the residuals given values on the predictor variables is a constant parameter equal to $\sigma^2_{y\cdot x}$.
5. Normality of residuals (and also $y$): for continuous outcomes, we assume that $\varepsilon_i \,|\, \boldsymbol{x}_i \sim N\left(0, \sigma^2_{y\cdot x}\right)$, or that given values on the predictor variables, the residuals are independently and identically distributed (i.i.d.) as normal random variables with mean 0 and constant variance $\sigma^2_{y\cdot x}$.
6. Independence of residuals: As a consequence of (5), $cov(\varepsilon_i, \varepsilon_j|\boldsymbol{x}_i, \boldsymbol{x}_j) = 0$, $i \neq j$, or residuals on different subjects are *uncorrelated* given values on their predictor variables.

There are several implications of these standard model assumptions. First, we can write:

$$\hat{y} = E(y \,|\, \boldsymbol{x}) = E(\boldsymbol{x}\boldsymbol{\beta}) + E(\varepsilon) = \boldsymbol{x}\boldsymbol{\beta} + 0 = \boldsymbol{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{7.5}$$

This equation for the predicted value of $y$ is the *regression function*, or the expected value of the dependent variable $y$ conditional on a set of values on the predictor variables (of which there are $p$). Further, we can write

$$var(y_i \,|\, \boldsymbol{x}_i) = \sigma^2_{y\cdot x}. \tag{7.6}$$

$$cov(y_i, y_j \,|\, \boldsymbol{x}_i, \boldsymbol{x}_j) = 0. \tag{7.7}$$

These assumptions, therefore, imply that the dependent variable has constant variance given values on the predictors and that no two values on the dependent variable are correlated given values on the predictors. Putting all of the implications together, we have

$$y_i \sim N\left(\boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2_{y\cdot x}\right). \tag{7.8}$$

Values on the dependent variable, $y$, are therefore assumed to be i.i.d. normally distributed random variables with a mean defined by the linear combination of the regression parameters and the predictor variables and a constant variance.

### 7.2.2 Survey Treatment of the Regression Model

Since the late 1940s and early 1950s, when economists and sociologists (Kendall and Lazarsfeld, 1950; Klein and Morgan, 1951) first applied regression analysis to complex sample survey data, survey statisticians have sought to relate design-based estimation of regression relationships to the standard linear model. The result was the linked concepts of a finite population and the superpopulation model, which are described in more detail in Chapter 3 and also Theory Box 7.1.

**THEORY BOX 7.1 FINITE POPULATIONS AND SUPERPOPULATION MODELS**

In theory, weighted estimation of a linear regression model, $y = \beta_0 + \beta_1 x + \varepsilon$, from complex sample survey data results in unbiased estimates of the regression function, $E(y|x) = B_0 + B_1 x$, where $B = [B_0, B_1]$ are *finite population regression parameters*. If instead of observing a sample of size $n$ of the $N$ population elements a complete census had been conducted, the finite population regression parameter $B_1$ for this simple "line" could be computed algebraically as follows:

$$B_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}.$$

The theory suggests that we distinguish between the true regression model parameters, $\boldsymbol{\beta}$, and the particular values, $\boldsymbol{B}$, that characterize the current finite survey population that was sampled for the survey. For example, consider a simple linear regression model of the effect of years of education on adults' earned income. The model can be interpreted as:

$$E(\text{Income} \mid \text{Education}) = \beta_0 + \beta_1 \cdot \text{Education(years)}$$

or

$$\text{Income} \mid \text{Education} \sim N\left(\beta_0 + \beta_1 \cdot \text{Education(years)}, \sigma^2_{I \cdot E}\right).$$

In contrast, a strict finite population regression interpretation is that for the particular population that has been sampled, the best prediction of adult income, given their education level, is found on the line $E(\text{Income}|\text{Education}) = B_0 + B_1 \cdot \text{Education(years)}$. The concept of a *superpopulation* links these two interpretations by introducing the assumption that although the surveyed population is finite (size $N$), the individual relationship between income and education in that finite population conforms to an underlying superpopulation model. Therefore, the fixed set of pairs of income and education values in the finite population of $N$ elements is itself a sample from an infinite number of possible data pairs that could be generated by a stochastic superpopulation model, denoted by $\zeta$: $\text{Income} = \beta_0 + \beta_1 \cdot \text{Education(years)} + \varepsilon$.

Under the superpopulation model, the finite population parameters $\boldsymbol{B}$ will vary about the true model parameters $\boldsymbol{\beta}$. However, the bias of making inferences for $\boldsymbol{\beta}$ based on estimates, $\hat{\boldsymbol{B}}$ (which are unbiased for $\boldsymbol{B}$), is small, and of the order $O(N^{-1/2})$ (meaning that the bias is a function of $N^{-1/2}$, and will thus become small as the population size becomes larger). Therefore, for large populations, an unbiased estimate of $\boldsymbol{B}$ (generally computed using weighted least squares [WLS]) can serve as an unbiased estimate of $\boldsymbol{\beta}$. Model diagnostics can then be used to question the hypothesized superpopulation model or the suitability of $\boldsymbol{B}$ as a summary measure of the relationships.

When survey analysts conduct a regression analysis using complex sample survey data, they choose between two targets—the finite population or a more universal superpopulation model—for their estimation and associated inference. Sometimes this is a conscious choice, and sometimes the choice is less conscious and only implicit in the analyst's presentation of the results and the inferences that are drawn.

In theory, weighted, design-based estimation of a regression equation permits the survey analyst to make unbiased inferences concerning values of the regression relationship as it exists within the finite population that corresponds to the geographic, demographic, and temporal definition of the survey population. To extend this inference beyond the survey population requires the analyst to make generalizing assumptions about the relationship of the finite population that has been studied to an overarching superpopulation model that may govern the relationships of the variables of interest. Theory Box 7.1 discusses the relationship of the superpopulation model and the finite population regression function in more detail. In practice, if the survey population is large and the regression model is correctly specified, then the survey data analyst may treat these two conceptual approaches as virtually equivalent (Skinner et al., 1989; Korn and Graubard, 1999).

## 7.3 Four Steps in Linear Regression Analysis

There are four basic steps that analysts should follow when fitting regression models to complex sample survey data: specification, estimation, evaluation (diagnostics), and inference. These steps apply to all types of regression models and not just those discussed in this chapter for continuous response variables. The following sections describe each step, considering the standard approach first, followed by the adaptation of the step for complex sample survey data.

### 7.3.1 Step 1: Specifying and Refining the Model

Survey data are typically observational data. The process of initial specification and subsequent refinement of a regression model for survey data involves multiple iterations of the four-step process. At the beginning of each cycle in this iterative process, it is important for the survey analyst to step back from the "number crunching" and critically evaluate the scientific interpretation and the plausibility of the emerging model.

A model is initially postulated based on subject matter knowledge and empirical investigation of the data. The specific aims of the analysis will often determine the choice of the dependent variable, $y$, and one or more independent variables of particular interest, $x$. Scientific subject matter knowledge and information gleaned from prior studies and publications can be used to identify additional independent variables (i.e., covariates that are known predictors of the dependent variable) or variables that may *mediate* or *moderate* (DeMaris, 2004) the relationships of the independent variables of primary interest with the dependent variable $y$. For example, an epidemiologist aiming to model the effect of obesity on systolic blood pressure (BP, in mmHg) decides to include age, gender, and race of the respondent as additional covariates. Based on prior research conducted by a colleague, she has evidence that advanced age moderates the relationship between systolic BP level and obesity. She will test for an interaction between age and the obesity measure (as well as other potential interactions—Section 7.4).

Empirical results from simple descriptive and graphical statistical analysis of the survey data itself can also be used to identify independent variable candidates for the model. The epidemiologist in our example might conduct an exploratory analysis, plotting systolic BP against respondent age. At older ages, the resulting scatter plot shows a curvilinear relationship to systolic BP, suggesting the addition of a quadratic term to the initial model. See Cleveland (1993) for a good resource on data exploration.

Contemporary survey data sets may contain hundreds of variables, and there is a temptation to bypass the scientific review and empirical investigations and "see what works." Regression programs in statistical software

packages often include variable selection algorithms such as *stepwise regression, forward selection,* and *backward selection* that are capable of culling a set of significant predictors from a large input of independent variable choices. These algorithms may prove useful in the model exploration and fitting process, but they are numerical tools and should not substitute for the survey analyst's own scientific and empirical assessment of the model and its final form. Careless use of such techniques leaves the analyst exposed to problems of *confounding* or *spurious relationships* that can distort the model and its interpretation.

A variety of variable selection and model-building approaches have been proposed for linear regression models. One such model-building "recipe" commonly used in our teaching and consulting practice is described in Section 7.4.5.

### 7.3.2 Step 2: Estimation of Model Parameters

After the survey data analyst has specified a linear regression model (Step 1), the next step in the modeling process involves computation of estimates of the regression parameters in the specified model. This section describes mathematical methods that can be used for estimation of those parameters.

#### *7.3.2.1 Estimation for the Standard Linear Regression Model*

By far, the most popular method of estimating unknown parameters in linear regression models is *ordinary least squares (OLS) estimation.* This method focuses on estimating the unknown set of regression parameters $\boldsymbol{\beta}$ in a specified model by minimizing the residual sum of squares (or sum of squared errors, SSE) based on the model:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2. \tag{7.9}$$

Once the estimate of $\boldsymbol{\beta}$ has been obtained analytically (7.11), an estimate of the variance of the random errors in the model, $\sigma^2_{y \cdot x}$, is obtained as follows:

$$\hat{\sigma}^2_{y \cdot x} = \frac{\sum_{i=1}^{n} (y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}})^2}{n - p}. \tag{7.10}$$

Here, $p$ is the number of regression parameters in the specified model.

The least squares estimate has several important properties. First, parameter estimates and their variances and covariances are analytically simple

to compute, requiring only a single noniterative algebraic or matrix algebra computation:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$
$$var(\hat{\boldsymbol{\beta}}) = \hat{\Sigma}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\hat{\sigma}^2_{y \cdot x}. \tag{7.11}$$

Second, the estimator is unbiased:

$$E(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \boldsymbol{\beta}. \tag{7.12}$$

Third, the estimator has the lowest variance among all other unbiased estimators that are also linear functions of the response values, making it the *best linear unbiased estimator* (BLUE). Finally, assuming normally distributed errors, the least squares estimates are equal to estimates derived based on maximum likelihood estimation (MLE).

As described in Section 7.2, one key assumption of the standard linear regression model is homogeneity of the error variance:

$$var(\varepsilon) = var(y_i \mid \boldsymbol{x}_i) = \sigma^2_{y \cdot x} = \text{constant}. \tag{7.13}$$

In practice, it is common to find that the variance of the residuals is heterogeneous—varying over the $i = 1, \ldots, n$ cases with differing values of $y$ and $\boldsymbol{x}$. For OLS estimation, the consequence of heterogeneity of variance is loss of efficiency (larger standard errors) in the estimation of the regression coefficients. *Weighted least squares (WLS)* estimation of the regression coefficients addresses this inefficiency by weighting each sample observation's contribution to the sums of squares by the reciprocal of its residual variance, $W_i = 1/\sigma^2_{y \cdot x,i}$. In matrix notation, the WLS estimator of the linear regression coefficients is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}. \tag{7.14}$$

Here, $\mathbf{W}$ is an $n \times n$ diagonal matrix (with zeroes off the diagonal) and the $n$ values of the inverse variance weights on the diagonal.

The standard linear regression model and the OLS estimator are statistically elegant, but the underlying assumptions are easily violated when analyzing real world data. To minimize the mean square error of estimation, techniques such as transformation of the dependent variable, WLS estimation, and other approaches such as ridge regression (Hoerl and Kennard, 1970) and robust variance estimation (Judge et al., 1985; Fuller et al., 1989) have been developed to address problems of non-normality, heterogeneity of variances, collinearity of predictors, and correlated errors.

### *7.3.2.2 Linear Regression Estimation for Complex Sample Survey Data*

Estimation of regression relationships for complex sample survey data alters the standard approach to estimation of coefficients and their standard errors. We first discuss what changes with estimation of the parameters, and then address what changes in terms of variance estimation.

#### *7.3.2.2.1 Estimation of Parameters*

The observed data from a complex sample survey are typically not "identically distributed." Due to variation in sample selection and sample inclusion probabilities, survey weights must generally be employed to develop unbiased estimates of the population regression parameters. Recall that in standard methods of regression analysis, WLS estimation incorporates a weight for each sample element inversely proportional to the residual variance. In the context of fitting regression models to complex sample survey data sets, where final survey weights have been calculated to compensate for unequal probability of selection, unit nonresponse, and possibly poststratification (Section 2.7), the survey weights can be incorporated into the estimation of the regression parameters via the use of WLS estimation. The contribution of each case to the residual sum of squares is made proportional to its population weight. This results in the following analytic formula for the WLS estimate of the finite population regression parameters:

$$\hat{B} = (X^T W X)^{-1} X^T W y. \tag{7.15}$$

Here, $W$ is an $n \times n$ diagonal matrix (with zeroes off the diagonal) and the $n$ values of the survey weights on the diagonal. Theory Box 7.2 provides mathematical motivation for the weighted survey estimator of $B$, and Theory Box 7.3 further considers the decision of whether to use analysis weights when estimating regression parameters from complex sample survey data.

Although the motives for conventional WLS estimation (heterogeneity of residual variances) and weighted survey estimation (unbiased population representation) are very different, the WLS computational algorithms that have been built into regression software for several decades serve perfectly well for weighted survey estimation of the finite population regression model. Consequently, analysts who naively specified the survey weight as the weight variable in a standard linear regression program (e.g., SAS PROC REG, `regress` in Stata, and SPSS Linear Regression) have obtained the correct design-based estimates of the population regression parameters. Unfortunately, the naïve use of weighted survey estimation in standard linear regression programs generally results in biased estimates of standard errors for the parameter estimates (see below). Stata explicitly recognizes the differences in weighting concepts and requires the user to declare probability weights, or "pweights." Stata also uses a robust estimator of variance when a "pweight" is specified.

### THEORY BOX 7.2 A "WLS" ESTIMATOR FOR FINITE POPULATION REGRESSION MODELS

When fitting regression models to complex sample survey data collected from a finite population, we compute estimates of the finite population parameters $B$ that minimize the following objective function:

$$f(B) = \sum_{i=1}^{N} (y_i - x_i B)^2.$$

We can think of this objective function $f(B)$ as a finite population "residual" sum of squares, $SSE_{pop}$. An unbiased sample estimate of this total incorporating the survey weights can be written as follows:

$$\hat{WSSE}_{pop} = \sum_{h}^{H} \sum_{\alpha}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} (y_{h\alpha i} - x_{h\alpha i} B)^2.$$

In this expression, $h$ is a stratum index, $\alpha$ is a cluster (or primary sampling unit) index, and $i$ is an index for elements within the $\alpha$-th cluster. WLS estimation is used to derive estimates of $B$ that minimize this sample estimate of the actual objective function for the finite population.

Correct interpretation of estimated regression parameters is essential for effectively communicating the results of investigations involving regression analyses in scientific publications. In a simple linear regression model for a continuous dependent variable, after obtaining estimates of the regression parameters, we can calculate the following expected values on the response variable, associated with a one-unit change in a given predictor variable $x$:

$$E(y \mid x = x_0) = \hat{B}_0 + \hat{B}_1 x_0. \tag{7.16}$$

$$E(y \mid x = x_0 + 1) = \hat{B}_0 + \hat{B}_1 x_0 + \hat{B}_1. \tag{7.17}$$

We can therefore write the following about the estimate of the regression parameter $\beta_1$:

$$\hat{B}_1 = E(y \mid x = x_0 + 1) - E(y \mid x = x_0). \tag{7.18}$$

### THEORY BOX 7.3 SHOULD SURVEY WEIGHTS BE USED TO ESTIMATE REGRESSION MODELS?

This is an important question that statisticians and data analysts working with complex sample survey data need to answer when fitting regression models. When employing the design-based approaches that we illustrate in this chapter and attempting to make inferences about finite populations, the use of survey weights ensures that the estimates of regression parameters will be unbiased with respect to the sample design. Weighted estimation does not, however, protect analysts from model misspecification; if an analyst is fitting a poorly specified model using survey weights, they will simply be computing unbiased estimates of the regression parameters in a model that does a poor job of describing relationships in the larger target population.

Statisticians and data analysts who advocate *model-based* approaches emphasize the importance of sound model specification. Those following this approach will generally argue that the use of survey weights in estimation should not be necessary if a model has been correctly specified; in these cases, the use of weights in estimation introduces the risk of computing inefficient estimates, with standard errors larger than they need to be (Korn and Graubard, 1999). If the computed survey weights are informative about a dependent variable of interest, then the variables used to compute the weights (or the weights themselves!) should be included in the model as covariates. Pfeffermann (2011) discusses alternative model-based and design-based approaches to fitting regression models to survey data, clarifying the differences between these approaches and finding general support (via simulation) for a variant of probability-weighted estimation.

Given that this is an area of healthy debate in the survey statistics literature (Gelman, 2007; Heeringa et al., 2015), we encourage data analysts to not choose a side, per se, but rather to consider the sensitivity of their inferences to alternative choices of estimation approaches. Today's statistical software makes it very easy to fit a carefully specified regression model with and without survey weights (holding the variance estimation approach constant), and examine the sensitivity of the results to the use of weights in estimation. In short, if the use of weights leads to substantially different parameter estimates and inferences, a model may be misspecified (e.g., Korn and Graubard, 1999), and the weighted estimates should be reported (as they will be unbiased). However, if the use of weights in estimation does not change parameter estimates substantially and only results in large increases in the standard errors of the estimates, a model may indeed be well-specified, and the use of weights in estimation may be unnecessary. One can also

formally compare weighted and unweighted regression parameters to assess the significance of the differences using a method described by Fuller (1984) and outlined on the web page for this book. Bollen et al. (2016) provide a comprehensive review of related methods for comparing weighted and unweighted regression parameters to assess the need for using weights in regression analysis.

Overall, we feel that the design-based approaches illustrated in this and later chapters, where the use of weights in estimation results in unbiased population estimates of regression parameters, generally work well in practice and are easier to defend as having favorable properties. We are usually never certain that regression models have been correctly specified, despite our best efforts, and we are willing to accept slight losses in the efficiency of estimates as long as they are unbiased. See Heeringa et al. (2015), Binder (2011), Pfeffermann (2011), or Korn and Graubard (1999) for more perspective on this issue. We consider the "q-weighted" approach for increasing the efficiency of weighted survey estimates of regression parameters that was proposed and evaluated by Pfeffermann (2011) in our illustration later in this chapter.

The parameter estimate therefore describes, on average, the expected change in the continuous response variable $y$ for a one-unit change in the predictor variable $x$.

When an additional predictor variable $z$ is added to the model, representing a theoretical control variable or possibly a confounding variable, a portion of the relationship between $x$ and $y$ is attributable to the variable $z$, and the interpretation of the regression parameter for the predictor variable $x$ requires that the predictor variable $z$ be fixed at a constant value, $z_0$:

$$E(y \mid x = x_0, z = z_0) = \hat{B}_0 + \hat{B}_1 x_0 + \hat{B}_2 z_0. \tag{7.19}$$

$$E(y \mid x = x_0 + 1, z = z_0) = \hat{B}_0 + \hat{B}_1 x_0 + \hat{B}_1 + \hat{B}_2 z_0. \tag{7.20}$$

$$\hat{B}_1 = E(y \mid x = x_0 + 1, z = z_0) - E(y \mid x = x_0, z = z_0). \tag{7.21}$$

We therefore interpret the estimate of the parameter $B_1$ in this case as the expected difference in $y$ associated with a one-unit increase in $x$, holding the value of the predictor variable $z$ constant. When multiple control variables are added to a linear regression model, we hold all of them at fixed values when interpreting the regression parameter for a primary predictor of interest. The interpretation of estimated regression parameters does not change

at all when analyzing sample survey data sets collected from finite populations, aside from the fact that the regression parameters are describing relationships in a finite population of interest. We will consider interpretations of regression parameters in detail in all examples presented in this chapter.

*7.3.2.2.2 Estimation of Variances of Parameter Estimates*

As described in Chapter 3, the complex designs of most large probability samples (stratification, cluster sampling, unequal selection probabilities) preclude the use of conventional variance estimators that can be derived for MLE for data that are presumed to be i.i.d. draws from a probability distribution (i.e., normal, binomial, Poisson). Instead, robust, nonparametric methods based on the TSL of the estimator or replication variance estimation methods (BRR, JRR, bootstrapping) are employed (Wolter, 2007).

The general approach to variance estimation using TSL for linear regression coefficients can be illustrated for the simple linear regression model (which involves a single predictor variable). Extension of the method to multiple linear regression is straightforward but algebraically more complex. In the case of a simple linear regression model with a single predictor $x$, an analytic formula for the calculation of the associated finite population regression parameter $B$ (given all data for the finite population with size $N$) can be written as follows:

$$B_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{T_{xy}}{T_{x^2}}. \tag{7.22}$$

This formula can be written as a ratio of two totals: $T_{xy}$ and $T_{x^2}$. We can calculate an estimate of this ratio by applying the survey weights to the observed sample data:

$$\hat{B}_1 = \frac{\sum_{h}^{H}\sum_{\alpha}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}(y_{h\alpha i} - \bar{y}_w)(x_{h\alpha i} - \bar{x}_w)}{\sum_{h}\sum_{\alpha}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}(x_{h\alpha i} - \bar{x}_w)^2} = \frac{t_{xy}}{t_{x^2}}. \tag{7.23}$$

Note that the sample estimate $\hat{B}$ is also a ratio of sample totals. Under the TSL approximation method, an estimate of the sampling variance of this ratio of two sample totals can be written as follows:

$$var(\hat{B}) \cong \frac{var(t_{xy}) + \hat{B}^2 var(t_{x^2}) - 2\hat{B}cov(t_{xy}, t_{x^2})}{(t_{x^2})^2}. \tag{7.24}$$

In multiple linear regression, TSL approximation methods require weighted sample totals for the squares and cross-products of all of the $y$ and $x = \{1\ x_1 \cdots x_p\}$ combinations. The computations are more complex but the approach is a direct extension of the technique shown here for a simple linear regression model. Replication methods like JRR or BRR (Section 3.6.3) can also be used to estimate sampling variances of estimated regression parameters (Kish and Frankel, 1974).

Statistical software designed for regression analysis of complex sample survey data applies the TSL, BRR, JRR, or bootstrapping methods to estimate the sampling variance of each parameter estimate, $var\left(\hat{B}_j\right)$, $j = 1, \ldots, p$, as well the $p(p+1)/2$ unique covariances, $cov(\hat{B}_j, \hat{B}_k)$, between the parameter estimates. These estimates of sampling variances and covariances are then assembled into the estimated *variance–covariance matrix* of the parameter estimates:

$$\boldsymbol{var}(\hat{\boldsymbol{B}}) = \hat{\Sigma}(\hat{\boldsymbol{B}}) = \begin{bmatrix} var(\hat{B}_0) & cov(\hat{B}_0, \hat{B}_1) & \cdots & cov(\hat{B}_0, \hat{B}_p) \\ cov(\hat{B}_0, \hat{B}_1) & var(\hat{B}_1) & \cdots & cov(\hat{B}_1, \hat{B}_p) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{B}_0, \hat{B}_p) & cov(\hat{B}_1, \hat{B}_p) & \cdots & var(\hat{B}_p) \end{bmatrix}. \quad (7.25)$$

The estimated variances and covariances can then be used to develop Student $t$-statistics and Wald chi-square or Wald $F$-statistics required to test hypotheses concerning the population values of the regression parameters (Section 7.3.4).

A variety of software procedures exist in statistical software packages for fitting linear regression models to survey data using the methods discussed in this section; we consider procedures in some of the more common general-purpose statistical software packages here. In SAS, PROC SURVEYREG can be used; IBM SPSS Statistics offers the General Linear Model procedure in the Complex Samples Module (CSGLM); Stata offers the `svy: regress` command, which will be considered in this book; SUDAAN offers PROC REGRESS; the `survey` package in R offers the `svyglm()` function. See Appendix A for more details on software options for linear regression analysis of complex sample survey data.

### 7.3.3 Step 3: Model Evaluation

The standard linear regression model is an "elegant" statistical tool, but its simplicity and "best" properties hinge on a number of model assumptions (Section 7.2). Standard texts on linear regression modeling (e.g., Neter et al., 1996) provide detailed coverage of procedures designed to evaluate the model *goodness of fit* (GOF), examine how closely the data match the basic model assumptions, and determine if certain observations are unduly influencing

the fit of the model. The process does not change substantially when fitting finite population models to complex sample survey data sets. In this section, we briefly consider some of the model diagnostics that can be used to evaluate the model and discuss some current work on model diagnostics adapted for regression models fitted to complex sample survey data (see also Theory Box 7.4):

1. *Explained variance and GOF*: A standard measure of the "fit" of the regression model to the data is the *coefficient of multiple determination* or the $R^2$ *statistic*, which is interpreted as the proportion of variance in the dependent variable explained by regression on the independent variables:

$$R^2 = 1 - \frac{SSE}{SST}. \tag{7.26}$$

   In Equation 7.26, SST refers to the total sum of squares, or the sum of squared differences between the response values and the mean of the response variable, and SSE is given in Equation 7.9 above. The use of $R^2$ as a measure of explained variance carries forward to regression modeling of complex sample survey data, although the statistic that is output by the analysis software is generally a *weighted* version, where each squared difference contributing to the sums is weighted by the corresponding survey weight:

$$R^2_{weighted} = 1 - \frac{WSSE}{WSST}. \tag{7.27}$$

   Although in theory it could be argued that this weighted $R^2$ statistic estimates the proportion of population variance explained by the population regression of $y$ on $x$, in practice it is safe to simply view it as the fraction of explained variance in $y$ attributable to the regression on $x$. Analysts who are new to regression modeling of social science, education, or epidemiological data should not fret if the achieved $R^2$ values are lower than those seen in their textbook training. Physicists may be disappointed with $R^2 < 0.98$–$0.99$ and chemists with $R^2 < 0.90$, but social scientists and others who work with human populations will find that their best regression model will often explain only 20%–40% of the variation in the dependent variable.
2. *Residual diagnostics*: In the standard regression context, analysis of the distributional properties of the residual terms, $\varepsilon_i = (y_i - \hat{y}_i)$, is used to evaluate how well the assumptions of the normal linear model are met (Section 7.2). Despite the theoretical distinction between the

## THEORY BOX 7.4 MODEL EVALUATION FOR COMPLEX SAMPLE SURVEY DATA

When fitting regression models to complex sample survey data collected from a finite population, outlier statistics used for model evaluation are simply adapted by replacing unweighted values with weighted values. This theory box presents mathematical expressions for some of these adaptations, which again are not yet widely implemented in general purpose statistical software packages. To keep the expressions as general as possible, we use notation from generalized linear model theory. Generalized linear models will be discussed in more detail in Chapters 8 and 9.

First, we consider computation of *Pearson residuals*:

$$r_{pi} = (y_i - \mu_i(\hat{\boldsymbol{B}}_w))\sqrt{\frac{w_i}{V(\hat{\mu}_i)}}.$$

In this notation, $\mu_i$ refers to the expected value of the outcome ($y$) for sampled case $i$, computed as a function of the weighted estimates of the regression parameters. The term $w_i$ refers to the survey weight for the $i$-th case based on the complex sample design, and the term $V(\mu_i)$ refers to the *variance function* for the outcome, which partly defines the variance of the outcome variable in a generalized linear model as a function of the expected value of the outcome.

The *hat matrix* is computed as

$$\mathrm{H} = \mathrm{W}^{1/2}\mathrm{X}(\mathrm{X}'\mathrm{W}\mathrm{X})^{-1}\mathrm{X}'\mathrm{W}^{1/2},$$

where:

$$\boldsymbol{W} = diag\left\{\frac{w_1}{V(\mu_1)[g'(\mu_1)]^2}, \cdots, \frac{w_n}{V(\mu_n)[g'(\mu_n)]^2}\right\}.$$

This matrix is used for the computation of various diagnostic statistics, and specifically, measures of leverage. This matrix is an $n \times n$ diagonal matrix, with zeroes off the diagonal and diagonal elements defined by the survey weights divided by a term that is a function of the variance function for the observation and the derivative of the *link function* $g$ for the specific generalized linear model (in a linear regression model, the link function is the identity function). Note that the derivative of the link function is computed for the diagonal elements and is evaluated as a function of the expected value of the outcome

according to the model. If a *canonical link* is used to define the generalized linear model (which is often the case in practice), the diagonal elements simplify to

$$\mathbf{W} = diag(w_1 V(\mu_1), \ldots, w_n V(\mu_n)).$$

Diagonal elements of the hat matrix (denoted by $h_{ii}$) updated with the survey weights are used to identify influential cases with high leverage, and a common rule of thumb is to identify diagonal elements larger than $2(p+1)/n$, or to identify large gaps in leverage values. Removing cases with high leverage will generally have little impact on estimates of regression parameters, but will have a large impact on the uncertainty of the estimates (i.e., standard errors).

Next, we consider computation of *Cook's Distance (D) statistic* in the complex sample design setting. Cook's D statistic can be useful for identifying observations that have a large impact on estimates of the regression parameters when they are removed from the estimation, taking the precision of the estimates into account. Cook's D statistic is computed for an individual sample observation $i$ as follows:

$$c_i = \frac{w_i^* w_i e_i^2}{p\phi V(\hat{\mu}_i)(1-h_{ii})^2}\, \mathbf{x}_i'[\hat{Var}(U_w(\hat{\mathbf{B}}_w))]^{-1}\mathbf{x}_i,$$

where:

$w_i^*$ = survey weight.
$w_i$ = remainder of the diagonal element in the hat matrix (e.g., $V(\hat{\mu}_i)$ for a canonical link).
$e_i$ = residual.
$p$ = number of parameters in the regression model.
$\phi$ = dispersion parameter in the generalized linear model.
$\hat{Var}(U_w(\hat{\mathbf{B}}_w))$ = linearized variance estimate of the *score equation*, which is used for pseudo MLE in generalized linear models fitted to complex sample survey data (Chapter 9).

Once the value of Cook's D has been determined for an individual sample element, the following test statistic can be computed to assess the significance of the D statistic:

$$\frac{(df-p+1)\cdot c_i}{df} \sim F(p, df-p),$$

where:
$df$ = number of PSUs—number of strata (design-based degrees of freedom).

This test statistic is approximately $F$-distributed, and can be used to identify any unusual elements that are having a significant impact on the estimates of the regression parameters when taking the complex design features into account.

When computation of these various diagnostic statistics becomes available in the statistical software packages discussed in this book, we will provide updates on the book web site.

concept of a finite population regression model and a broader super-population model, we recommend using standard residual analysis to evaluate regression models that are fitted to complex sample survey data. Li and Valliant (2015) describe recent advances in the study of residual diagnostics for regression models fitted to data from complex samples. These advances have been implemented in a new contributed package for the R software (`svydiags`), and we illustrate use of the functions in this package later in this chapter (Section 7.5.4).

3. *Model specification and homogeneity of variance*: Two-way scatter plots of the residuals for the estimated model against the predicted values, $\hat{y}$, and the independent variables, $x_j$, can identify problems with lack of correct functional form (e.g., omitting a squared term for age when modeling blood pressure) or where a moderating variable or interaction has not been correctly included in the model.

   These same plots may be used to diagnose a problem with heterogeneity of residual variances. A pattern of residuals that spreads out in a fan-shaped pattern with increasing values for the predicted $y$ or increasing values of an independent variable is a common observation. To address the problem of heterogeneity of variance and reduce standard errors for the estimated $\beta$ coefficients, a standard approach in regression analysis is to employ WLS, weighting each observation's contribution to the SSE inversely proportional to its residual variance, that is, $w_i^{(hv)} = 1/\sigma^2_{y \cdot x_i}$. With complex sample survey data, this becomes complicated, because weighted estimating equations for finite population regression parameters, $\boldsymbol{B}$, already include the survey weights, say $w_i^{(survey)}$. While it is possible to create a composite weight, $w_i^* = w_i^{(survey)} \cdot w_i^{(hv)}$ (see Little, 2004 for a suggested approach), in cases of serious heterogeneity of variance it may be possible to identify a transformation of the dependent or independent variables

(see below) that eliminates much of the residual variance heterogeneity but still permits the use of the survey weights in the estimation of the regression function.

4. *Normality of the residual errors*: This assumption can be assessed using standard diagnostic plots for the model-based residuals (e.g., normal quantile–quantile [Q–Q] plots or histograms). The Q–Q plot is a plot of quantiles for the observed residuals against those computed from a theoretical normal distribution having the same mean and variance as the distribution of observed residuals. A straight 45° line in this plot would therefore suggest that normality is a reasonable assumption for the random errors in the model. We present examples of these Q–Q plots in the application later in this chapter.

   In large survey data sets, formal tests of the normality of the residuals (e.g., the Kolmogorov–Smirnov and Shapiro–Wilks tests) tend to be extremely "powerful" and the null hypothesis of normality will be rejected due to the slightest deviation from normality. Our recommendation is to use the more informal visual methods illustrated in Section 7.5. Empirical research has provided evidence that as long as the residuals display symmetry about $E(e_i) = 0$, the regression estimates are quite robust against failure of strict normality. If the residual distribution is highly skewed or irregular (e.g., bimodal), the analyst should first determine that the model has been correctly specified (Section 7.4.1), and no important predictor variables or interaction terms have been omitted in error.

   Transformation of the dependent variable is a common method to address serious problems of non-normality of residuals and also heterogeneity of residual variances. Analysts should be careful when making transformations, however, because they can destroy the straightforward interpretation of the parameters discussed above. A common transformation that is often used when violations of normality are apparent and residual distributions appear to be right-skewed is the natural (base $e$) log transformation of the response variable:

$$\ln(y) = B_0 + B_1 x + e. \tag{7.28}$$

When this particular transformation is used, the regression parameters still have a somewhat straightforward interpretation:

$$\frac{E(y \mid x = x_0 + 1)}{E(y \mid x = x_0)} = \frac{e^{(B_0 + B_1 x_0 + B_1)}}{e^{(B_0 + B_1 x_0)}} = e^{B_1}. \tag{7.29}$$

That is, a one-unit change in a given predictor variable will *multiply* the expected response by $\exp(B_1)$. The important issue to keep

in mind when transforming the dependent variable is that predicted values on the response variable need to be back-transformed to the original scale of the response. For example, square root transformations are often used to stabilize variance of the residuals, and predicted values based on this type of model would need to be squared to return to the original scale of the response.

5. *Collinearity diagnostics*: When specifying linear regression models, analysts need to be careful not to include predictor variables that are highly correlated with each other in the same model. This type of model misspecification can lead to problems of *multicollinearity*, which can have an adverse effect on the properties of the estimated regression parameters: standard errors of the estimated coefficients become inflated, due to *variance inflation factors* (which are a function of the amount of variance in a given predictor explained by the other predictors), and the signs of estimated coefficients may even change artificially relative to what their values would be in the absence of multicollinearity. Analysts should always first examine the correlations of the different predictor variables that they are considering, and make sure that none of them are excessively high in absolute value (say, greater than 0.7). In addition, *condition indices* for fitted linear regression models should not exceed 30 in value, and variance inflation factors greater than 2 warrant consideration and further analysis. In these cases, analysts should consider either combining correlated predictors (e.g., using principal components analysis or factor analysis) or omitting redundant predictors. See Faraway (2014) for more discussion of problems with multicollinearity.

   Recent research in this area has considered the computation of these important diagnostic statistics for assessing multicollinearity when fitting linear regression models to complex sample survey data. Liao and Valliant (2012a,b) have led the way in this area, developing the underlying theory for determining these diagnostic statistics in a manner that accounts for the complex sampling features and illustrating practical examples of their implementation. Unfortunately, at the time of this writing, these diagnostic tools have not yet been implemented in any major statistical software packages. Liao and Valliant are currently developing a contributed package for the R software (mentioned earlier) that will include these collinearity diagnostics. We will provide updates in this area on the web page for the second edition of this book.

6. *Outliers and influence statistics*: Finally, analysts should determine if the sample data include *outlier values* (observations poorly fitted by a given model) or *influential points* (observations that have a strong impact on the fit of a model). Tools such as standard residual plots, *studentized residuals*, *hat statistics* (measures of leverage), and

*Cook's Distance (D) statistics* have been developed for this purpose. Regression models can be fitted with and without potential outliers and influential points, in order to assess whether the fit of the model (i.e., the significance of regression parameters) is changing substantially depending on the points in question. For more detail on these diagnostic methods in the standard linear regression case, we refer readers to Neter et al. (1996) or Faraway (2014).

Influential points should still be investigated for their impact on the fit of a model when fitting linear regression models to complex sample survey data sets. What changes is the way that complex sampling features are incorporated when calculating the various statistics (e.g., Cook's D statistics) used for residual diagnostics. Li and Valliant (2009, 2011a) were the first to discuss the identification of influential points when sampling weights are involved in the estimation of linear regression models, and these authors also discussed how to adapt the forward search method for identifying influential cases to the complex sampling context (Li and Valliant, 2011b). More recent work by Li and Valliant (2015) discussed extensions of the adjustments to diagnostic statistics for samples involving stratification and cluster sampling. Collectively, the current literature in this area suggests that the survey weights are most important for assessing the influence of individual observations on model fit.

Unfortunately, these promising methods for producing weighted versions of commonly used diagnostic statistics have not yet been widely implemented in general-purpose statistical software packages containing procedures for linear regression analysis of complex sample survey data. At the time of this writing, Rick Valliant has communicated that these tools will eventually be available in the aforementioned contributed R package `svydiags`, which is currently under development. We demonstrate use of the functions for regression diagnostics that are currently available in this package in the example analyses later in this chapter. It is our hope that the very near future will bring additional software advances in this area, and we will continue to provide readers with updates in this regard on the book web site.

This evaluation step of regression analysis may lead to a modification of the model structure. One then cycles back through the model-fitting and model diagnostics process, with careful consideration of model structure and parsimony.

### 7.3.4 Step 4: Inference

After following the three steps above, the final step in the model-building process is making inferences about the regression parameters in the finite population of interest. This section describes that process.

### 7.3.4.1 Inference Concerning Model Parameters

After rigorously evaluating the fit of a model, an analyst can use the estimated parameters and their standard errors to characterize or infer about the conditional distribution of $y$ given the predictor variables $x$. The analyst can perform a variety of hypothesis tests concerning the parameters being estimated, ranging from tests for a single regression parameter to tests for multiple regression parameters. We begin this section by considering tests for single regression parameters.

In the standard linear regression context, when the residuals follow a normal distribution, hypothesis tests for a single regression parameter associated with predictor variable $k$ employ a $t$-test statistic:

$$t = \frac{\hat{\beta}_k - \beta_k}{se(\hat{\beta}_k)} \sim t_{n-p}. \tag{7.30}$$

Therefore, when performing a test of the null hypothesis $H_0$: $\beta_k = 0$ versus the alternative hypothesis $H_A$: $\beta_k \neq 0$, one can calculate a $t$-statistic by dividing the estimate of the regression parameter $\beta_k$ by its standard error. For reasonably large samples, when $H_0$ is true, this test statistic is distributed as a variate from a Student $t$ distribution with $n - p$ degrees of freedom, where $p$ is the number of parameters being estimated in the regression model (including the intercept). The analyst (or their software) refers the computed $t$-statistic to a Student $t$ distribution with $n - p$ degrees of freedom. If the absolute value of the test statistic, $t$, exceeds a critical value of the Student $t$ distribution (e.g., $t_{1-\alpha/2,\, n-p}$ for a two-sided test), $H_0$ is rejected with a Type I error probability of $\alpha$. Pivoting on the value of $t$ for the two-sided test, a $100(1 - \alpha)\%$ confidence interval for the true model parameter can be constructed as $\hat{\beta}_k \pm t_{1-\alpha/2,df} \cdot se(\hat{\beta}_k)$.

When constructing the confidence interval (or the pivotal hypothesis test statistic) for a single regression parameter estimated from complex sample survey data, two aspects of the inferential process change: (1) $se(\hat{B})$, or the correct standard error of the estimated regression parameter, is *estimated* using a nonparametric technique like TSL, BRR, or JRR; and (2) the degrees of freedom for the Student $t$ reference distribution must be adjusted to reflect the reduced degrees of freedom for the complex sample estimate of $se(\hat{B})$. Recall from Chapter 3 that the design degrees of freedom are approximated as $df = \Sigma_h a_h - H$, or the number of primary stage ultimate clusters minus the number of primary stage strata. For example, in the National Comorbidity Survey Replication (NCS-R) data set, the approximation to the design degrees of freedom is $84 - 42 = 42$. The correct estimate of the standard error and degrees of freedom for the Student $t$ reference distribution can be used to develop a design-based $100(1 - \alpha)\%$ confidence interval (corresponding to a Type I error rate of $\alpha$) for the regression parameter of interest, as follows:

$$\hat{B} \pm t_{1-\alpha/2,df} \cdot se(\hat{B}). \tag{7.31}$$

The $t$-statistic for the comparable two-sided hypothesis test of $H_0$: $B = 0$ can be developed as: $t = \hat{B}/se(\hat{B})$. This is the form of the $t$-test statistic that is routinely printed in tables of regression model output. The "$p$-values" generally printed alongside the test statistics are the probability that $t_{df} \geq t$.

In standard linear regression, $F$-tests are often used to test hypotheses about multiple parameters in the model. The *overall F-test* typically reported in the analysis of variance (ANOVA) table output generated by software procedures for fitting regression models using standard OLS methods tests the null hypothesis $H_0$: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, that is, the fitted model predicts $E(y|x)$ no better than a model that only includes the intercept ($\beta_0 = \bar{Y}$). *Partial F-tests* can be used to test whether selected subsets of parameters in the model are not significantly different from 0. In this case, a "full" model is compared with a nested "reduced" model that contains a subset of the predictor variables in the "full" model. This type of hypothesis test is essentially a test of the null hypothesis that multiple parameters (the parameters omitted in the "reduced" model) are all equal to 0. This multiparameter test can be extremely useful for testing hypotheses about categorical predictor variables represented by several indicator variables in a regression model (Section 7.4.2.1). More formally, we can use the following notation to indicate the $p = p_1 + p_2$ predictor variables of interest:

$$x = (x_1, x_2)$$
$$x_1 = p_1 \text{ predictors}$$
$$x_2 = p_2 \text{ predictors.}$$

Then, we can write the full model as follows:

$$full: y = x\beta + \varepsilon. \tag{7.32}$$

The reduced model then omits the $p_2$ predictor variables:

$$reduced: y = x_1\beta_1 + \varepsilon. \tag{7.33}$$

Then, to test the null hypothesis that the regression parameters associated with the $p_2$ predictor variables are all equal to 0, that is, $H_0$: $\beta_2 = 0$, we can calculate the following partial $F$-test statistic:

$$F = \frac{\dfrac{SSE_{reduced} - SSE_{full}}{(n-p_1)-(n-p_1-p_2)}}{\dfrac{SSE_{full}}{n-p_1-p_2}} = \frac{\dfrac{SSE_{reduced} - SSE_{full}}{p_2}}{\dfrac{SSE_{full}}{n-p}}. \tag{7.34}$$

Under the null hypothesis and assuming normally distributed residuals, this *F*-statistic follows an $\mathcal{F}$ distribution with numerator degrees of freedom equal to $p_2$ and denominator degrees of freedom equal to $n - p$. This general result allows one to perform a variety of multiparameter tests when comparing nested linear regression models, at least in the simple random sample setting. This *F*-statistic is also fairly robust to slight deviations from normality in the residuals.

In the complex sample survey data setting, these multiparameter tests must be adapted to the complex design features of the sample. *Wald test statistics* (Judge et al., 1985) replace the overall *F*-test and the partial *F*-test. The equivalent multiparameter Wald test statistics can be calculated as follows:

$$\begin{aligned} \textit{Overall: Modified } X^2_{W,overall} &= \frac{\hat{\boldsymbol{B}}^T \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{B}})^{-1} \hat{\boldsymbol{B}}}{p} = F_{W,overall} \\ \textit{Partial: Modified } X^2_{W,partial} &= \frac{\hat{\boldsymbol{B}}_2^T \boldsymbol{\Sigma}(\hat{\boldsymbol{B}}_2)^{-1} \hat{\boldsymbol{B}}_2}{p_2} = F_{W,partial}, \end{aligned} \tag{7.35}$$

where:

$\hat{\boldsymbol{B}}, \hat{\boldsymbol{B}}_2$ are vectors of estimated regression parameters.

$\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{B}}), \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{B}}_2)$ are the estimated variance–covariance matrices.

Under the null hypothesis $H_0$: $\boldsymbol{B} = \boldsymbol{0}$, the overall modified Wald test statistic, $F_{W,overall}$, follows an $\mathcal{F}$ distribution with numerator degrees of freedom equal to $p$ and denominator degrees of freedom equal to the design degrees of freedom (*df*). Likewise, to test $H_0$: $\boldsymbol{B}_2 = \boldsymbol{0}$, or the null hypothesis that the $p_2$ parameters are all equal to 0 in the nested model, the modified Wald partial test statistic is referred to the critical value of the *F* distribution with $p_2$ and *df* degrees of freedom.

Wald tests can also be used to test more general hypotheses regarding linear combinations of regression model parameters. Consider the null hypothesis $H_0$: $\boldsymbol{CB} = \boldsymbol{0}$, where $\boldsymbol{C}$ is a matrix that defines specific linear combinations of the regression parameters in the vector $\boldsymbol{B}$. In this case, a version of the Wald test statistic that follows a chi-square distribution with degrees of freedom equal to the rank of the matrix $\boldsymbol{C}$ under the specified null hypothesis can be written as follows:

$$X^2_W = [\boldsymbol{C}\hat{\boldsymbol{B}}]'[\boldsymbol{C}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{B}})\boldsymbol{C}']^{-1}[\boldsymbol{C}\hat{\boldsymbol{B}}] \approx \frac{\text{contrast "squared"}}{\text{variance of contrast}}. \tag{7.36}$$

We consider one example of this more general type of hypothesis test for linear combinations of multiple parameters. First, suppose that a specified linear regression model includes three parameters of interest, that is, $\boldsymbol{B}' = [B_1, B_2, B_3]$. Using this more general framework for the Wald test statistic, if one

wished to test the null hypothesis $H_0$: $B_2 - B_3 = 0$ (or equivalently $H_0$: $B_2 = B_3$), the $C$ matrix would take the following form:

$$C = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Then, the Wald test statistic specified in Equation 7.36 above would follow a chi-square distribution with one degree of freedom (because the rank of the $C$ matrix is 1). These more general hypothesis tests are available in a variety of software packages that can fit regression models to complex sample survey data sets, including SAS and Stata. Dividing these more general chi-square test statistics by the number of parameters being tested will result in test statistics that follow $F$ distributions, similar to Equation 7.35. We consider examples of how to specify these tests for multiple parameters in Section 7.5.

### *7.3.4.2 Prediction Intervals*

Predicting expected outcomes for populations and single individuals is an important scientific application of regression modeling (Neter et al., 1996). Even in the social and health sciences, where regression models are more often used to explore relationships among dependent and independent variables, prediction from fitted regression models still has a role. In the standard linear regression case, given a set of predictor values, $x_{obs,i}$, the estimated regression model can be used to calculate a predicted value for $y$, in addition to *confidence intervals* and *prediction intervals* for the predicted value. First, we consider the expected value of the response variable $y$ given an estimated model and a known vector of values on the predictor variables, $x_{obs,i}$:

$$\hat{E}(y_i \mid x_{obs,i}) = x_{obs,i}\hat{\beta}. \tag{7.37}$$

Given this expected value, we can calculate a confidence interval for the expected value once the variance of the expected value is calculated:

$$var(\hat{E}(y_i \mid x_{obs,i})) = x'_{obs,i} cov(\hat{\beta}) x_{obs,i}. \tag{7.38}$$

A $100(1 - \alpha)\%$ confidence interval for the expected value of $y$ given $x_{obs,i}$ (i.e., the average expected outcome for a population of cases with covariates $x_{obs,i}$) can then be calculated as follows:

$$x_{obs,i}\hat{\beta} \pm t_{1-\alpha/2,n-p}\sqrt{var(\hat{E}(y_i \mid x_{obs,i}))}. \tag{7.39}$$

Note that the confidence interval above does not take into account the variance of the random errors that are also a part of the linear regression model. A *prediction interval* for a *single future value* of $y$ does take this estimated variance $\left(\hat{\sigma}^2_{y \cdot x}\right)$ into account:

$$x_{obs,i}\hat{\beta} \pm t_{1-\alpha/2,n-p}\sqrt{var(\hat{E}(y_i \mid x_{obs,i})) + \hat{\sigma}^2_{y \cdot x}}. \tag{7.40}$$

Prediction intervals, therefore, are wider than more standard confidence intervals for the expected value because they also include variance in the prediction due to random error. Both intervals can give analysts a notion of the precision of the predicted values based on the fitted model if prediction of future values is an important objective of the modeling process.

Confidence intervals for predicted values can also be computed in the context of regression models for complex sample survey data. Standard errors for the predicted values can be computed using the *delta method* (essentially linearization), which is a technique that can accommodate a wide variety of general predictions, based on fixed values of the predictors and the estimated regression parameters. For computational details on this technique, which is an approximate method based on large samples, interested readers can refer to the Stata (Version 14) Survey Data Reference Manual (SVY; StataCorp, 2015). The intervals change in that the degrees of freedom used to calculate the critical $t$-statistic are now based on the design degrees of freedom, and the variance–covariance matrix of the parameter estimates is estimated using approximate methods like TSL or replicated methods like JRR and BRR. Correct computation of these confidence intervals for predicted values when fitting regression models to complex sample survey data is currently implemented in Stata's `predictnl` postestimation command.

One can also compute different forms of *marginal predicted values* based on regression models fitted to complex samples (along with confidence intervals for the marginal predicted values). Heuristically, the marginal predicted value of a dependent variable for a given value of a covariate of interest is computed by first obtaining predicted values of the dependent variable using design-based estimates of the regression parameters of interest, and assuming that every case in the data set has the same value of the covariate of interest. The values of the other covariates (aside from the covariate of interest) are either left unchanged or fixed to constant values (e.g., their means) when computing these predicted values for each case. These predicted values are then averaged across all cases to compute the marginal predicted value, and a standard error for this marginal predicted value can be computed using the delta method. *Average marginal effects* of specific changes in the covariate of interest on the dependent variable can then be computed for any given values of the other covariates by comparing marginal predicted values for different values of the covariate of interest with selected

other covariates fixed to particular values (Bauer, 2015). These approaches are currently implemented in the Stata software via the `margins` command, and marginal predicted values computed using this command (and corresponding confidence intervals for the predicted values) can then be plotted using the subsequent `marginsplot` command. We illustrate the use of these commands in our illustration later in this chapter.

## 7.4 Some Practical Considerations and Tools

In this section, we discuss important considerations for data analysts fitting regression models to complex sample survey data sets in practice, and provide practical guidance on steps to avoid potential pitfalls when fitting regression models and making inferences based on the fitted models.

### 7.4.1 Distribution of the Dependent Variable

We specifically focus our discussion in this chapter on linear regression models for *continuous dependent variables* (e.g., weight in kilograms, blood pressure in millimeters of mercury to the second decimal place, weekly household expenditures on food items). Surprisingly, many survey data sets include very few variables that are measured on a truly continuous scale. Many response variables may be *semicontinuous, censored,* or *grouped* (or "coarsened") in nature. We do not consider models for semicontinuous, censored, or grouped dependent variables in this chapter; Tobit regression models, Heckman selection models, and other forms of latent variable models might be considered by analysts for these types of response variables (Skrondal and Rabe-Hesketh, 2004). The Stata software (Version 14) currently provides versions of commands designed to handle dependent variables of this type in the setting of complex sample survey data.

Many survey variables of interest are measured as ordinal scale variables. Examples include age in years and education in years. Survey questions such as "On a scale of 1–5, where 1 is excellent and 5 is poor, please rate your overall health" produce a response on an ordinal scale. Over the years, it has been common practice to fit linear regression models to ordinal scale-dependent variables. DeMaris (2004) describes an ordinal scale variable as *approximately continuous* if it meets the following conditions: the number of sample observations, $n$, is large; measurement is at least on an ordinal scale; the response has at least five-ordered levels; and the distribution of responses to the ordered categories is not skewed and ideally is approximately normal in appearance. We agree that there are obvious cases where it may be acceptable to apply linear regression to an ordinal dependent variable; however, analysts will generally have a difficult time satisfying the

underlying statistical assumptions for the models discussed in this chapter when working with ordinal outcomes. This could lead to highly inefficient or faulty inferences. Especially for ordinal variables with small numbers of levels, a better choice is to choose a regression model that is more appropriate for the measurement scale of the dependent variable. Several regression models appropriate for ordinal and categorical response variables are discussed in detail in Chapter 9.

### 7.4.2 Parameterization and Scaling for Independent Variables

An extremely important aspect of fitting linear regression models is the treatment and coding of categorical *predictor* variables, which can definitely be considered in the linear regression models discussed in the chapter. See also Theory Box 7.5 for a discussion of fitting analysis of variance (ANOVA) and analysis of covariance (ANCOVA) models as linear regression models for categorical predictors. When analysts consider nominal categorical predictor variables (e.g., race/ethnicity, region of the country) in linear

**THEORY BOX 7.5 ANOVA AND ANCOVA AS LINEAR REGRESSION ANALYSIS**

Statistical texts on linear models often include separate chapters for linear regression models for continuous outcomes, ANOVA models (where all predictors are categorical) and analysis of covariance (ANCOVA) models (involving a mix of categorical and continuous predictors). ANOVA and ANCOVA models for normal data are generally discussed in the context of experimental designs (e.g., full factorial, randomized block) and experimental hypotheses are tested using $F$-statistics and multiple comparisons that are functions of expected mean squares. Historically, the distinction between ANOVA, ANCOVA, and linear regression analysis was reinforced in statistical software systems that included separate programs adapted for ANOVA and standard linear regression modeling. Other programs such as SAS PROC GLM integrated these two analyses in a linear model framework.

In fact, ANOVA- and ANCOVA-type analyses can be performed through proper specification of a linear regression model (e.g., main effects, interactions, nesting); see Neter et al. (1996). Analysts who wish to apply ANOVA-type procedures to complex sample survey data can do so using regression analysis programs with indicator variable parameterization of categorical independent variables and interactions appropriate for the ANOVA-type model that they wish to fit to the data (e.g., indicator variables for levels of the main effects and interactions for a full factorial model).

regression models, they often generate *indicator variables* (a.k.a. "dummy" variables) to represent levels of the categorical predictor variables in the models. Regression parameters associated with these indicator variables (which are equal to 1 for cases falling into a specific category and 0 otherwise) represent changes in the expected value of the continuous outcome for a specific category relative to a *reference category*, which does *not* have an indicator variable included in the model. Alternative dichotomous specifications of these indicator variables are possible (e.g., 1/–1 "effect" coding, where estimated regression parameters represent changes in the expected value of the outcome relative to the *overall mean*), but we focus on the (1, 0) coding in this book for ease of interpretation.

Consider a categorical predictor variable measuring race/ethnicity with three possible values in a survey of a human population: 1 = Caucasian, 2 = African-American, and 3 = Other race/ethnicity. Analysts need to choose one of these three categories to be a reference category, and this choice is generally guided by contrasts of interest and research objectives (e.g., comparing African-Americans and other groups to Caucasians). In cases where the choice of the reference category is not clear, choosing the most prevalent group in the sample data will suffice. In the case of the race/ethnicity variable, an analyst choosing "Caucasian" to be the reference category would need to create two indicator variables to include as predictors in a regression model: a variable indicating African-Americans (1 = African-American, 0 = Caucasian or Other race/ethnicity), and a variable indicating the Other race/ethnicity group (1 = Other race/ethnicity, 0 = Caucasian/African-American). Most modern statistical software capable of fitting regression models to survey data will perform this "dummy" coding automatically for categorical predictors, requiring the analyst to simply choose the reference category for the analysis.

Continuing with the race/ethnicity example, suppose that a regression model was fitted to a continuous response variable $y$, where race/ethnicity was the only predictor variable. The dummy coding described above would lead to the following regression function:

$$E(y \mid x) = B_0 + B_1 x_1 + B_2 x_2. \tag{7.41}$$

In this model, $x_1 = 1$ for African-Americans and 0 otherwise, while $x_2 = 1$ for other racial/ethnic groups and 0 otherwise. The expected value on the continuous outcome variable $y$ for African-Americans would therefore be calculated as

$$E(y \mid x) = B_0 + B_1 \times 1 + B_2 \times 0 = B_0 + B_1, \tag{7.42}$$

and the expected value for individuals in other racial/ethnic groups would be calculated as

$$E(y \mid x) = B_0 + B_1 \times 0 + B_2 \times 1 = B_0 + B_2. \tag{7.43}$$

Because both indicator variables would be equal to 0 for Caucasians, the expected value on the outcome variable for Caucasians would simply be $B_0$. The regression parameters $B_1$ and $B_2$ therefore represent differences in the expected outcomes between African-Americans or Other racial/ethnic groups and Caucasians, and hypothesis tests about differences between the groups could therefore be conducted by testing whether these parameters are equal to 0. Similarly, the difference in expected outcome values between the nonreference groups (African-Americans and Others) would be equal to $B_1 - B_2$, or the difference in expected outcomes between these two groups.

When using statistical software to fit regression models to complex sample survey data, analysts have two choices: they can create indicator variables manually, or use special options in the different procedures to have the software automatically create indicator variables for the levels of categorical variables, given knowledge about a reference category. Users of the Stata software can define the reference category of a given categorical predictor variable when specifying a regression command, using *factor variable coding*. For example, one could fit a linear regression model using the command:

```
regress depvar ib1.catvar,
```

where `depvar` refers to the dependent variable in a linear regression model (fitted using the `regress` command), `catvar` refers to a categorical predictor variable, and `ib1.` refers to the specific value of `catvar` that is to be set as the reference category (in this case, the indicator variable (`i`) that will be omitted from the model as the baseline (`b`) category is 1). In general, when fitting any type of regression model, Stata users can include `i.` before the names of any categorical predictor variables, and Stata will by default set the *lowest alphanumeric category* as the reference category (if the user does not specify some form of `ib#.`). The following code presents an example of this, once again considering the race/ethnicity example (where OUTCOME is the continuous outcome variable and ETHNIC is the race/ethnicity variable):

```
regress outcome i.ethnic
```

We will consider several additional examples of factor variable coding in Stata later in this chapter.

Analysts should also exercise caution when interpreting intercept parameters ($\beta_0$) in linear regression models. The intercept, or the expected value of the response variable $y$ when all of the predictor variables are fixed at value 0, is often not of much interest, because it may represent an expectation that is well outside the range of the collected data on the predictor variables. As a result, reformulation (or *centering*) of the continuous predictor variables in the model can make the intercept more interpretable:

$$y = \beta_0^* + \beta_1(x - \bar{x}) + \beta_2(z - \bar{z}) + \varepsilon. \tag{7.44}$$

By subtracting the means of the predictor variables $x$ and $z$ from the observed values on each variable, and then using the "centered" predictors in the model, the reformulated intercept (denoted by an asterisk) now has the interpretation of the expected value on the response variable $y$ when the predictors are equal to their *means*. One can think of this term as representing a *centercept*, or the overall average value on the response variable $y$.

We note that the choice of omitting the intercept in a regression model forces the expected value of $y$ to be 0 when all of the predictor variables in the model are set to 0. This might be done when preliminary knowledge of the subject matter being studied dictates that the expected value of the response be 0 when all predictor variables are set to 0.

### 7.4.3 Standardization of the Dependent and Independent Variables

Analysts can also *standardize* all of the variables being considered in a linear regression model (including indicator variables). One can standardize any variable by subtracting the mean for the variable, similar to centering, and then also dividing by the standard deviation of the variable:

$$x_{i,std.} = \frac{x_i - \bar{x}}{sd(x)}. \tag{7.45}$$

This is often done in practice when predictor variables are measured on very different scales (e.g., income and grade point average), and rescales all of the variables so that they are on the same scale (changes of one unit in standardized variables correspond to changes of one standard deviation in the unstandardized variables). The estimates of the regression parameters in a model where *all* variables have been standardized are often referred to as *standardized regression coefficients*, and these can be used to determine which predictor variable has the largest relative impact on the expected value of the response variable. Nothing changes about this process when analyzing complex sample survey data, but an analyst should note whether weighted estimates or unweighted sample estimates of the mean and standard deviation were used to standardize the variables.

### 7.4.4 Specification and Interpretation of Interactions and Nonlinear Relationships

Analysts should also exercise caution when interpreting regression parameters associated with predictor variables that define nonlinear relationships between predictors and the outcome, or predictor variables that define interactions between two or more predictors. Consider the following linear

regression model defining a nonlinear (and specifically quadratic) relationship between $x$ and $y$:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon. \tag{7.46}$$

In this model, the regression parameters $\beta_1$ and $\beta_2$ are *linked*, because they involve different transformations of the same predictor $x$, and $\beta_1$ alone in the presence of $\beta_2$ has no interpretation. Instead, the parameter $\beta_2$ measures the extent of the nonlinearity in the relationship between $x$ and $y$. If the estimate of $\beta_2$ suggests that the parameter is not different from 0, one can assume (unless higher order polynomial terms are significant) that the relationship between $x$ and $y$ is linear.

The same issue arises when considering linear regression models that involve interactions between predictors. Consider the following linear regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \varepsilon. \tag{7.47}$$

In this model, the three regression parameters $\beta_1$, $\beta_2$, and $\beta_3$ are once again *linked*, and one cannot interpret the regression parameter $\beta_1$ as being the relationship of the predictor variable $x$ with the response variable $y$. The full relationship of $x$ with $y$ depends on the values of $z$ (and therefore the value of $\beta_3$), and different values of $z$ will result in different relationships of $x$ with $y$.

In general, interactions between two or more predictor variables are computed for entry into a regression model by saving the product of two or more variables as a new variable. Interactions can be easily computed for two or more continuous predictor variables, for two or more categorical predictor variables, or for combinations of both types of variables. When working with categorical predictor variables, relevant products of *all* dummy variables to be included in the model for a given categorical variable must be computed for all other predictor variables that are specified to interact with the categorical predictor. For example, to include the interaction of a three-category predictor with a continuous variable, two product terms must be computed and included in the regression model: the product of the continuous variable with each of the indicators for the two nonreference categories of the categorical predictor. Most software procedures will perform these tasks automatically, but analysts need to be extremely careful when including interaction terms in a regression model.

Plotting predicted values in linear regression models with significant interactions can be extremely helpful when attempting to interpret significant regression parameters (i.e., regression parameters statistically different from zero) associated with the interactions. Social scientists often think of one of the variables involved in an interaction (e.g., $z$ in the example above)

as a *moderator variable,* because that variable moderates the relationship of the other variable(s) involved in the interaction with the response variable (i.e., the relationship of $x$ with the response variable depends on the value of $z$). The plot in Figure 7.2 illustrates the predicted values of a continuous response variable $y$ based on the parameter estimates in two different regression models, showing two possible interactions between a continuous predictor variable $x$ and a binary moderator variable $z$, taking on values of 1 and 0 for two different groups.

In the left panel of Figure 7.2, there appears to be an interaction of $x$ with $z$; the relationship of $x$ with the continuous outcome $y$ clearly depends on the value of $z$. The right panel shows no interaction, because the relationship of $x$ with $y$ in both groups defined by $z$ is essentially the same. Analysts often make the mistake of interpreting regression parameters associated with single predictors in models that include interactions between that predictor and other predictors as "main effects." For example, in the first model fitted in Figure 7.2 (left panel), an analyst may be tempted to interpret the regression parameter associated with the predictor $x$ as being the "main effect" of $x$. In truth, the relationship of $x$ depends on the value of $z$, even if

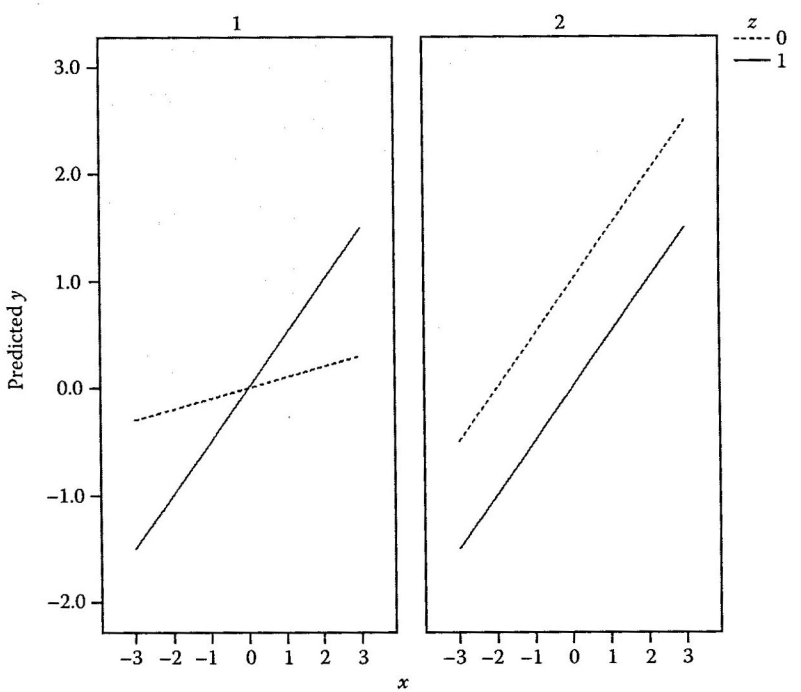


**FIGURE 7.2**
Hypothetical fits of two linear regression models, one with a significant interaction between $x$ and $z$ (left panel), and one without (right panel).

the interaction between $x$ and $z$ *is not significant*, so there is no "main effect" of $x$. This problem can be eliminated by eliminating regression parameters representing interactions from the model if they are not significantly different from zero, which will be discussed in the next section. Nothing about the interpretation of interactions changes when fitting models to complex sample survey data.

### 7.4.5 Model-Building Strategies

A universal set of model-building steps that is widely acknowledged and will always lead to the best fitting linear regression model does not exist. Many statisticians have proposed practical guidelines for model fitting, keeping in mind the four steps in regression modeling discussed in Section 7.3. Out of many quality choices, we consider the model-building steps proposed by Hosmer et al. (2013). These steps are summarized below:

1. Conduct exploratory bivariate analyses (e.g., two-sample $t$-tests, chi-square tests, tests of correlations, one-way ANOVA, etc.) to get a sense of candidate predictors that appear to have a significant relationship with the response variable.
2. Include those predictor variables that are *scientifically relevant* and have a bivariate relationship of significance $p < 0.25$ with the response variable in the initial multivariate model, and possibly consider variable selection techniques (e.g., backward selection), with discretion. Be wary of *multicollinearity*, which could arise from including predictors that are highly correlated with each other in the same model (Section 7.3.3).
3. Verify the importance of the predictor variables retained in the model, using $t$-tests for individual coefficients and Wald tests for multiple coefficients (Section 7.3.4), and assess whether or not the coefficients of *all* of the predictor variables change substantially in the multivariate model (relative to the bivariate case); this represents the preliminary "main" model.
4. Examine the forms of the predictor variables: if they are categorical, are sample sizes in each category large enough to use the categories as they are in the model? If they are continuous, do they have linear relationships with the response variable? Or do the relationships appear to be nonlinear? Residual diagnostics are useful as a part of this step.
5. Consider adding *scientifically relevant* interactions between the predictor variables to the model, one at a time, and do not retain them if they are not significant.
6. If any continuous or ordinal predictor variable has a large number of zeroes, include an indicator variable that is equal to 1 for nonzero

values and 0 for zero values in the model, in addition to the predictor variable in question, and see if the fit of the model has been improved.

The idea behind this last step (6) is that if we have a semicontinuous predictor with many zero values and also continuous values, it is unlikely that we would have a true linear relationship passing through the mass of points at 0 and continuing through the range of nonzero values. If we create the indicator variable and also include the predictor variable in the model, this introduces a discontinuity, modeling the effect of being zero or nonzero and then for nonzero values the linear relationship of the predictor with the response variable.

We remind readers that there are many possible model-fitting strategies that one can follow; in particular, Harrell (2015) also provides a comprehensive critique of alternative strategies.

## 7.5 Application: Modeling Diastolic Blood Pressure with the 2011–2012 NHANES Data

In this practical application of linear regression analysis for complex sample survey data, we consider building a predictive model of diastolic BP (DBP) (a continuous response variable) based on the sample of data collected from the U.S. adult population (age ≥18) in the 2011–2012 National Health and Nutrition Examination Survey (NHANES). After exploring the bivariate relationships of the predictors of interest with DBP, we perform a naïve linear regression analysis that completely ignores the complex design features of the NHANES sample. Next, we perform a weighted regression analysis that ignores the stratification and cluster sampling of the NHANES sample design. Finally, we take all of the important design features of the NHANES sample (stratification, cluster sampling, and weighting for unequal probability of selection, nonresponse, and poststratification) into account at each step of the model-building process.

Specifically, the design variables that the online documentation for the 2011–2012 NHANES data set states should be used for variance estimation* include SDMVPSU (which contains masked versions, or approximations, of the true primary sampling unit codes for each respondent for the purposes of variance estimation; Section 4.3.1) and SDMVSTRA (which contains the "approximate" sampling stratum codes for each respondent, for variance estimation purposes). In addition, the appropriate survey weight to be used to generate finite population estimates of the regression parameters

* http://www.cdc.gov/nchs/data/nhanes/analytic_guidelines_11_12.pdf

for the U.S. adult population for the years of 2011 and 2012 is WTMEC2YR. This survey weight variable was selected for analysis purposes instead of WTINT2YR because variables that will be used in the regression analyses were collected as a part of the physical examination, and the NHANES physical examination was performed on a *subsample* of all respondents (which required adjustments to the survey weights to account for the subsampling and nonresponse to the mobile examination center (MEC) follow-up phase of the NHANES data collection).

### 7.5.1 Exploring the Bivariate Relationships

In this application, we follow the regression modeling strategies recommended by Hosmer et al. (2013) to build a model for DBP (Section 7.4.5). We will describe each of the steps explicitly as a part of the example. First, we consider a set of predictors of DBP that are scientifically relevant: age, gender, ethnicity, and marital status. We begin by identifying the relevant design variables for the NHANES sample in Stata, requesting TSL for variance estimation, and indicating that identified strata with single PSUs should result in missing standard errors (see Chapters 3 and 4 for more on this issue):

```
svyset sdmvpsu [pweight = WTMEC2YR], strata(sdmvstra) ///
vce(linearized) singleunit(missing)
```

An initial descriptive summary of the DBP variable in the NHANES data set (BPXDI1) revealed several values of 0, and we set these values to missing in Stata before proceeding with the analysis:

```
gen bpxdi1_1 = BPXDI1
replace bpxdi1_1 = . if BPXDI1 == 0
```

We also generate an indicator variable for the subpopulation of adults (respondents with age greater than or equal to 18), for use in the analyses:

```
gen age18p = 1 if age >= 18 & age != .
replace age18p = 0 if age < 18
```

With the subpopulation indicator defined, we now consider a series of simple bivariate regression analyses, to get an initial exploratory sense of the relationships of the candidate predictor variables with DBP. We make use of the `svy: regress` command to take the survey weights, stratum codes, and cluster codes into account when fitting these simple initial regression models, so that parameter estimates will be unbiased and variance estimates will reflect the complex design features of the NHANES sample. We first compute a weighted estimate of the mean age for the adult subpopulation,

and then center the AGE variable at the weighted mean age based on the NHANES sample (46.36):

```
svy, subpop(age18p): mean age
gen agec = age - 46.36
```

Next, the continuous dependent variable, BPXDI1_1, is regressed separately on each of the candidate predictors. For the categorical predictor variables (i.e., race/RIDRETH1, gender/RIAGENDR, and marital status/MARCAT), we consider multiparameter Wald tests in Stata (Section 7.3.4) to assess the significance of the bivariate relationships. The Stata software allows users to perform these multiparameter Wald tests by using `test` commands immediately after the models have been estimated:

```
svy, subpop(age18p): regress bpxdi1_1 i.RIDRETH1
test 2.RIDRETH1 3.RIDRETH1 4.RIDRETH1 5.RIDRETH1

svy, subpop(age18p): regress bpxdi1_1 i.marcat
test 2.marcat 3.marcat

svy, subpop(age18p): regress bpxdi1_1 i.riagendr
test 2.riagendr

svy, subpop(age18p): regress bpxdi1_1 agec
test agec
```

Note in the four Stata commands above how the indicator for the adult subpopulation (AGE18P) is explicitly specified for the analysis, via the use of the `subpop()` option. This ensures that Stata will perform an unconditional subclass analysis, treating the adult subpopulation sample size as a random variable and taking the full complex design of the NHANES sample into account.

We note also that after the dependent variable has been specified first following the `regress` command, the categorical predictor variables in the regression models are identified with the `i.` modifier, initiating dummy variable (or indicator) coding. In the example commands above, we use the default factor variable coding, where the lowest alphanumeric value of a given categorical variable will be set as the reference category, with no corresponding indicator variable included in the model. The terms indicated in the `test` commands above (e.g., `2.RIDRETH1`) then refer to the specific regression parameters for each of the nonreference indicator variables temporarily generated by Stata to fit the model, and the labels for the parameters can be clearly identified in the model output. The `test` commands are used to test the null hypothesis that all of the listed regression parameters are equal to zero (or, in other words, the mean DBP does not vary as a function of a given predictor variable). Table 7.1 presents the results of these initial bivariate analyses.

**TABLE 7.1**

Initial Design-Based Bivariate Regression Analysis Results Assessing Potential Predictors of Diastolic Blood Pressure for the 2011–2012 NHANES Adult Sample

| Predictor Variable | Parameter Estimate (Linearized SE) | Test Statistic | *p*-Value |
|---|---|---|---|
| **Race/ethnicity ($n$ = 5,112)** | | Wald $F(4,14) = 3.93$ | 0.02 |
| Mexican-American | – | – | – |
| Other Hispanic | −0.15 (1.46) | $t(17) = -0.11$ | 0.92 |
| Non-Hispanic White | 2.18 (0.74) | $t(17) = 2.94$ | <0.01 |
| Non-Hispanic Black | 2.29 (0.70) | $t(17) = 3.26$ | <0.01 |
| Other race | 1.31 (0.70) | $t(17) = 1.85$ | 0.08 |
| **Age (cent.) ($n$ = 5,112)** | 0.04 (0.02) | $t(17) = 2.09$ | 0.05 |
| **Gender ($n$ = 5,112)** | | Wald $F(1,17) = 15.01$ | <0.01 |
| Male | – | – | – |
| Female | −2.20 (0.57) | $t(17) = -3.87$ | <0.01 |
| **Marital status ($n$ = 4,845)** | | Wald $F(2,16) = 0.85$ | 0.45 |
| Married | – | – | – |
| Previously married | −0.15 (0.70) | $t(17) = -0.21$ | 0.84 |
| Never married | −1.12 (0.84) | $t(17) = -1.33$ | 0.20 |

*Note:* – denotes reference category.

Stata presents *adjusted Wald tests* for the parameters in each of these models by default, where the standard Wald $F$-statistic (Section 7.3.4) is multiplied by $(df - k + 1)/df$, with $df$ = the design-based degrees of freedom, and $k$ = the number of parameters being tested (Korn and Graubard, 1990). Under the null hypothesis, the resulting test statistic follows an $F$ distribution with $k$ and $df - k + 1$ degrees of freedom. For example, in the Wald test for the ethnicity predictor, there are $k = 4$ parameters being tested, and the design-based degrees of freedom are equal to 31 (ultimate clusters) minus 14 (strata), or 17. The denominator degrees of freedom for the adjusted test statistic are therefore $17 - 4 + 1 = 14$.

Note the different subpopulation sample sizes in Table 7.1; over 250 of the adult cases appear to have missing data on the marital status variable. The design-based multiparameter Wald tests and $t$-tests for the single parameters suggest that race/ethnicity, age, and gender each have potentially significant relationships with the response variable (DBP). Specifically, investigating the weighted parameter estimates in these simple models, males, non-Hispanic whites, non-Hispanic blacks, and older adults appear to have the highest DBPs at first glance, while marital status does not appear to be related to DBP. Following the guidelines of Hosmer et al., we therefore include the first three predictors in an initial model for the response variable measuring DBP.

### 7.5.2 Naïve Analysis: Ignoring Sample Design Features

In the first regression analysis, we *ignore* the sample weights, stratification, and cluster sampling inherent to the NHANES sample design, we do not consider any interactions between the predictors, and we use standard OLS estimation to calculate the parameter estimates for the adult subpopulation:

```
regress bpxdi1_1 i.ridreth1 i.riagendr agec if age18p == 1
```

When fitting regression models in Stata, the first variable listed after the main command is the response variable (BPXDI1_1), and the variables listed after the response variable represent the predictor variables in the model. The variable list is then generally followed by options (after a comma). In this example, we do not include any options; however, we do restrict the analysis conditionally to those subjects with age ≥18 by using the `if` modifier. We also once again use the `i.` modifiers to have Stata automatically generate indicator variables for selected levels of the categorical predictor variables (recall that Stata, by default, treats the lowest valued level of a categorical predictor as the reference category; see Section 7.4.2 for syntax to manually choose the reference category). Table 7.2 presents OLS estimates of the regression parameters in this preliminary model, along with their standard errors and associated test statistics.

**TABLE 7.2**

Unweighted OLS Estimates of the Regression Parameters in the Initial Diastolic Blood Pressure Model

| Predictor | Parameter Estimate | Standard Error | *t*-Statistic (*df*) | *p*-Value | 95% CI |
|---|---|---|---|---|---|
| Intercept | 70.784 | 0.548 | 129.20 (5105) | <0.001 | (69.709, 71.858) |
| *Ethnicity* | | | | | |
| Other Hispanic | 0.255 | 0.738 | 0.35 (5105) | 0.730 | (–1.192, 1.702) |
| Non-Hispanic White | 1.193 | 0.597 | 2.00 (5105) | 0.046 | (0.021, 2.364) |
| Non-Hispanic Black | 2.205 | 0.615 | 3.58 (5105) | <0.001 | (0.999, 3.412) |
| Other race | 2.013 | 0.662 | 3.04 (5105) | 0.002 | (0.716, 3.310) |
| Mexican-American | – | – | – | – | – |
| *Gender* | | | | | |
| Female | –2.404 | 0.331 | –7.25 (5105) | <0.001 | (–3.054, –1.754) |
| Male | – | – | – | – | – |
| **Age (centered)** | 0.041 | 0.009 | 4.58 (5105) | <0.001 | (0.024, 0.059) |

*Note:* $n = 5{,}112$, $R^2 = 0.018$, $F$-test of null hypothesis that all parameters are 0: $F(6, 5105) = 15.58$, $p < 0.001$. – denotes the reference category.

These initial parameter estimates suggest that age has a positive linear relationship with DBP, females tend to have significantly lower mean DBP, and Mexican-American respondents tend to have the lowest mean BPs (significantly lower than whites, blacks, and other ethnicities). These parameter estimates may be biased, however, because the NHANES survey weights for respondents given a physical examination were not used to calculate nationally representative finite population estimates. In addition, the standard errors are likely understated, because the weights and the stratified cluster sample design of the NHANES sample were not taken into account. We therefore consider these results only for illustration purposes.

### 7.5.3 Weighted Regression Analysis

Next, we consider WLS estimation for calculating the parameter estimates in the initial model. Note that we explicitly indicate in the Stata command (with the `pweight` option) that the NHANES survey weights for respondents given a physical examination (WTMEC2YR) should be included in the estimation to calculate unbiased estimates of the regression parameters:

```
regress bpxdi1_1 i.RIDRETH1 i.riagendr agec ///
if age18p == 1 [pweight = WTMEC2YR]
```

Table 7.3 below presents weighted estimates of the regression parameters, in addition to *robust standard errors* automatically calculated by Stata's standard

**TABLE 7.3**

Weighted Least Squares (WLS) Estimates of the Regression Parameters in the Initial Diastolic Blood Pressure Model

| Predictor | Parameter Estimate | Robust Standard Error | *t*-Statistic (*df*) | *p*-Value | 95% CI |
|---|---|---|---|---|---|
| Intercept | 71.149 | 0.566 | 125.66 (5105) | <0.001 | (70.039, 72.259) |
| *Ethnicity* | | | | | |
| Other Hispanic | –0.141 | 0.721 | –0.20 (5105) | 0.845 | (–1.555, 1.272) |
| Non-Hispanic White | 1.904 | 0.611 | 3.12 (5105) | 0.002 | (0.707, 3.102) |
| Non-Hispanic Black | 2.302 | 0.645 | 3.57 (5105) | <0.001 | (1.037, 3.567) |
| Other race | 1.262 | 0.705 | 1.79 (5105) | 0.074 | (–0.121, 2.644) |
| Mexican-American | – | – | – | – | – |
| *Gender* | | | | | |
| Female | –2.291 | 0.432 | –5.31 (5105) | <0.001 | (–3.138, –1.444) |
| Male | – | – | – | – | – |
| **Age (centered)** | 0.037 | 0.012 | 3.17 (5105) | 0.002 | (0.014, 0.060) |

*Note:* $n = 5{,}112$, weighted $R^2 = 0.017$, $F$-test of null hypothesis that all parameters are 0: $F(6, 5105) = 10.41$, $p < 0.001$. – denotes the reference category.

regression command (`regress`) when sampling weights are explicitly specified with the `pweight` option. Recall from earlier chapters that this option generally refers to a "probability" weight; the full NHANES survey weights are the base "probability" weights, adjusted for nonresponse and calibrated to population control totals (Section 2.7). The "robust" standard errors are "sandwich-type" standard errors (see Freedman, 2006, for an introduction) that are considered robust to possible misspecification of the correlation structure of the observations. In this part of the example, there is some misspecification involved: we have once again *ignored* the stratification and cluster sampling inherent to the NHANES sample design when calculating the standard errors, meaning that they will likely be understated. Stata's automatic calculation of robust standard errors for the parameter estimates in the presence of analysis weights is therefore an effective type of "safeguard" against this failure to incorporate the sample design features in the analysis (meaning that standard errors will not be understated), but we do not recommend following this approach in practice. Readers should be aware that not all software packages capable of survey data analysis perform this type of calculation automatically when standard regression commands are used with survey weights specified. We only present this part of the analysis for illustrative purposes.

In Table 7.3, we note fairly large differences in some of the parameter estimates relative to the OLS case (Table 7.2), especially in terms of the race/ethnicity parameters. When failing to incorporate the survey weights (Table 7.2), the differences between the ethnic groups were either being overstated (note that the difference in means between those with other ethnicities and Mexican-Americans is smaller and no longer significant at the 0.05 level) or understated (the difference between Mexican-Americans and whites is larger when accounting for the weights, with stronger evidence of a significant difference). The estimates in Table 7.3 are nationally representative parameter estimates, and use of the OLS estimates in Table 7.2 would have painted an incorrect picture of the relationships of these variables with DBP; the weights definitely appear to be informative about these relationships (see Theory Box 7.3 for more discussion of whether to use weights in regression modeling). We also note that the robust standard errors tend to be larger than the understated standard errors from Table 7.2, where no adjustments to the standard errors were made to account for the complex design features of the NHANES sample.

To emphasize the differences that analysts might see when specifying the survey weights but failing to specify the sampling error codes (stratum and cluster codes) correctly in specialized software procedures for regression analysis of survey data, we include output from a similar analysis using SAS PROC GLM with a WEIGHT statement below:

```
proc glm data = nhanes1112;
   class ridreth1 (ref = "1") riagendr (ref = "1");
```

```
   model bpxdi1_1 = ridreth1 riagendr agec / solution;
   where age18p = 1;
   weight wtmec2yr;
run;
quit;
```

| Parameter | Estimate | Standard Error | *t*-Value | Pr > \|*t*\| |
|---|---|---|---|---|
| Intercept | 71.14869684 | 0.59156886 | 120.27 | <0.0001 |
| RIDRETH1 2 | –0.14141197 | 0.84161219 | –0.17 | 0.8666 |
| RIDRETH1 3 | 1.90419899 | 0.60771827 | 3.13 | 0.0017 |
| RIDRETH1 4 | 2.30195323 | 0.73448786 | 3.13 | 0.0017 |
| RIDRETH1 5 | 1.26178602 | 0.80523230 | 1.57 | 0.1172 |
| RIDRETH1 1 | 0.00000000 | . | . | . |
| RIAGENDR 2 | –2.29113574 | 0.31815963 | –7.20 | <0.0001 |
| RIAGENDR 1 | 0.00000000 | . | . | . |
| agec | 0.03682344 | 0.00928946 | 3.96 | <0.0001 |

Readers should note in the SAS output above that the weighted parameter estimates are identical to those found in Stata, but the standard errors are computed differently (using standard WLS calculations). A more appropriate approach for SAS users would be to use PROC SURVEYREG and specify the NHANES stratum and cluster variables, enabling appropriate variance estimation. We now consider this more appropriate form of the analysis.

### 7.5.4 Appropriate Analysis: Incorporating All Sample Design Features

We now use the `svy: regress` command in Stata to fit the initial finite population regression model to the adult subpopulation and take *all* of the NHANES complex sample design features into account, calculating weighted estimates of the regression parameters and linearized estimates of the standard errors for the parameter estimates (incorporating the stratification and cluster sampling of the NHANES sample). Note how an unconditional subpopulation analysis is requested by specifying the binary AGE18P indicator in the `subpop()` option, similar to the bivariate analyses performed previously:

```
svyset sdmvpsu [pweight = WTMEC2YR], strata(sdmvstra) ///
vce(linearized) singleunit(missing)

svy, subpop(age18p): regress bpxdi1_1 i.RIDRETH1 ///
i.riagendr agec

estat effects, deff
```

We also use the postestimation command `estat effects, deff` to request calculation of design effects for the estimated regression parameters. Table 7.4 presents the estimated parameters in this initial "main" model:

**TABLE 7.4**

Design-Based Estimates of the Regression Parameters in the Initial "Main" Model for Diastolic Blood Pressure, Linearized Standard Errors for the Estimates, Design-Adjusted Test Statistics and Confidence Intervals for the Parameters, and Design Effects for the Parameter Estimates

| Predictor | Est. | Linearized SE | *t*-Statistic (*df*) | *p*-Value | 95% CI | DEFF |
|---|---|---|---|---|---|---|
| Intercept | 71.149 | 0.518 | 137.4 (17) | <0.001 | (70.056, 72.241) | 1.06 |
| *Ethnicity* | | | | | | |
| Other Hispanic | –0.141 | 1.375 | –0.10 (17) | 0.919 | (–3.042, 2.759) | 3.86 |
| Non-Hispanic White | 1.904 | 0.809 | 2.35 (17) | 0.031 | (0.197, 3.611) | 2.47 |
| Non-Hispanic Black | 2.302 | 0.665 | 3.46 (17) | 0.003 | (0.900, 3.704) | 1.00 |
| Other race | 1.262 | 0.707 | 1.79 (17) | 0.092 | (–0.229, 2.753) | 1.09 |
| Mexican-American | – | – | – | – | – | – |
| *Gender* | | | | | | |
| Female | –2.291 | 0.548 | –4.18 (17) | <0.001 | (–3.448, –1.134) | 3.97 |
| Male | – | – | – | – | – | – |
| **Age (centered)** | 0.037 | 0.021 | 1.77 (17) | 0.095 | (–0.007, 0.081) | 6.02 |

*Note:* Subclass $n = 5{,}112$, weighted $R^2 = 0.017$, adjusted Wald test for all parameters: $F(6,12) = 10.13$, $p < 0.001$. – denotes the reference category.

The estimated parameters and tests of significance presented in Table 7.4 confirm most of the simple relationships observed in the initial design-based bivariate analyses, and suggest that the relationships remain similar when taking other predictor variables into account in a multivariate analysis (with the exception of the linear relationship of age with DBP). When holding the other predictor variables in this model fixed, non-Hispanic Whites and Blacks have significantly higher expected DBP values than Mexican-Americans, females have significantly lower DBP than males, and interestingly, age does not appear to have a significant linear relationship with DBP. Age is, therefore, the only predictor that does not appear to be important, but we have only considered a linear relationship thus far. None of the sample sizes for the groups defined by the categorical variables appear to be extremely small, so we do not consider further recoding of these variables. Readers should note that the weighted parameter estimates in Table 7.4 are exactly equal to those in Table 7.3; differences arise in how the estimated standard errors for the parameter estimates are being calculated.

There are several important observations regarding the test statistics for the regression parameters in Table 7.4. First, the degrees of freedom for the *t*-statistics based on the complex sample design of the NHANES (17) are calculated by subtracting the number of strata (14) from the number of sampling error computation units, or ultimate clusters (31). These degrees of freedom are substantially different from those noted in Tables 7.2 and 7.3 ($df = 5105$), where the complex design was not taken into account when performing the

estimation; this shows how the primary sampling units (rather than the unique respondents) are providing the independent contributions to the estimation of distributional variance when one accounts for the complex sample design. In addition, Stata presents an adjusted Wald test for all of the parameters in the model (see the discussion of the Table 7.1 results). The numerator degrees of freedom for this adjusted statistic are equal to $k$ (6 in this example, because six parameters are being tested; the "null" or "reduced" model still contains the intercept parameter), and the denominator degrees of freedom are calculated as $df - k + 1$ ($17 - 6 + 1 = 12$ in this example). This adjusted Wald test definitely suggests that a null hypothesis that all of the regression parameters are equal to 0 would be strongly rejected.

The design effects presented in Table 7.4 (DEFF) are nearly all greater than 1, suggesting that the complex design of the NHANES sample is generally resulting in a decrease in the precision of the parameter estimates relative to the precision that would have been achieved under a simple random sampling design with the same sample size (Section 2.4). The effects of the complex design on the standard errors are apparent.

We now consider some initial model diagnostics to assess the fit of this preliminary model. We start by saving the residuals in a new variable (RESIDS) in Stata, and then plotting the residuals against the values of the continuous mean-centered age (AGEC) variable. The left-hand panel of Figure 7.3 presents this plot.

```
predict resids, resid
scatter resids agec
```

The first plot in Figure 7.3 indicates a fairly well-defined curvilinear pattern of the residuals as a function of age, suggesting that the structure of the model has been misspecified; there is evidence that age actually has a quadratic relationship with DBP that has not been adequately captured by including a linear relationship of age with the response variable. We therefore add a squared version of age (AGECSQ) to the model to capture this relationship:

```
gen agecsq = agec * agec

svy, subpop(age18p): regress bpxdi1_1 i.RIDRETH1 ///
i.riagendr agec agecsq
estat effects

predict resids2, resid
scatter resids2 agec
```

In the new model (Table 7.5), the regression parameters for both the centered age predictor and the squared version of the age predictor are significantly different from 0 ($p < 0.001$), confirming that the relationship of age with DBP is in fact nonlinear and quadratic in nature. The weighted $R$-squared of the

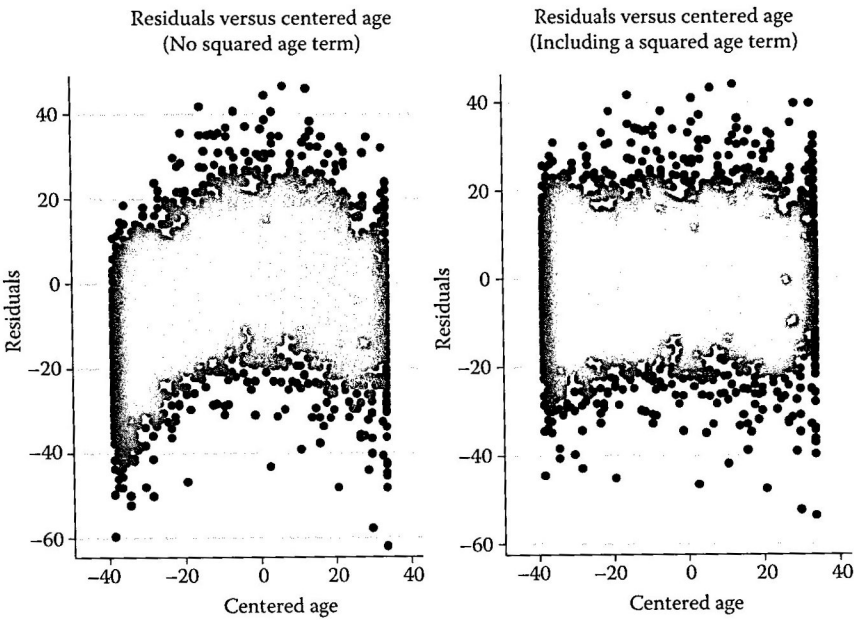


**FIGURE 7.3**
Plots of residuals versus centered age for the diastolic blood pressure application, before and after the addition of the squared AGE variable to the model.

**TABLE 7.5**

Estimates of the Regression Parameters in the Intermediate Model for the Diastolic Blood Pressure Response Variable, Prior to Inclusion of Interaction Terms

| Predictor | Est. | Linearized SE | *t*-Statistic (*df*) | *p*-Value | 95% CI | DEFF |
|---|---|---|---|---|---|---|
| Intercept | 74.462 | 0.565 | 131.73 (17) | <0.001 | (73.270, 75.655) | 0.65 |
| *Ethnicity* | | | | | | |
| Other Hispanic | 0.218 | 1.217 | 0.18 (17) | 0.860 | (−2.350, 2.786) | 1.22 |
| Non-Hispanic White | 2.084 | 0.857 | 2.43 (17) | 0.026 | (0.276, 3.893) | 1.31 |
| Non-Hispanic Black | 2.511 | 0.734 | 3.42 (17) | 0.003 | (0.963, 4.059) | 1.16 |
| Other race | 1.410 | 0.687 | 2.05 (17) | 0.056 | (−0.041, 2.860) | 1.20 |
| Mexican-American | – | – | – | – | – | – |
| *Gender* | | | | | | |
| Female | −2.169 | 0.489 | −4.43 (17) | <0.001 | (−3.202, −1.137) | 1.16 |
| Male | – | – | – | – | – | – |
| **Age (centered)** | 0.075 | 0.016 | 4.80 (17) | <0.001 | (0.042, 0.108) | 1.59 |
| **Age (cent.) squared** | −0.012 | 0.001 | −16.28 (17) | <0.001 | (−0.013, −0.010) | 1.77 |

*Note:* Subclass $n = 5{,}112$, weighted $R^2 = 0.114$, adjusted Wald test for all parameters: $F(7,11) = 159.86$, $p < 0.001$. – denotes the reference category.

new model becomes 0.114, suggesting an improved fit by allowing the relationship of age with DBP to be nonlinear. The right-hand panel of Figure 7.3 shows the improved distribution of the residuals as a function of age after adding the squared term, where there is no pattern evident in the residuals as a function of age.

Now, we consider testing specific interactions of interest, one at a time: the interactions between age (both predictors) and race/ethnicity, and the interactions between age (both predictors) and gender. This step essentially allows for testing whether the nonlinear relationship of age with DBP tends to be moderated by these two demographic factors; for example, is the quadratic trend in DBP as a function of age flatter (i.e., more stable) for certain ethnic groups than others? We first add the interaction between age and ethnicity to the model, and then investigate a design-adjusted Wald test of the null hypothesis that all eight parameters associated with this first order interaction are simultaneously equal to zero:

```
svy, subpop(age18p): regress bpxdi1_1 i.RIDRETH1 ///
i.riagendr agec agecsq ///
i.RIDRETH1#c.agec i.RIDRETH1#c.agecsq

test 2.RIDRETH1#c.agec 3.RIDRETH1#c.agec ///
4.RIDRETH1#c.agec 5.RIDRETH1#c.agec 2.RIDRETH1#c.agecsq ///
3.RIDRETH1#c.agecsq 4.RIDRETH1#c.agecsq 5.RIDRETH1#c.agecsq
```

Note in the `svy: regress` command how the interactions are specified: the term `i.RIDRETH1#c.agec` indicates that an interaction (#) between the categorical predictor RIDRETH1 (`i.`) and the *continuous* predictor AGEC (indicated with `c.`) should be included in the model. The subsequent `test` command then simply lists all eight parameters defining the two interaction terms, exactly as they are labeled in the output (including the `c.` notation for the continuous predictors).

We remind readers that when using the `test` commands in conjunction with survey regression commands, Stata performs an adjusted Wald test by default. The `nosvyadjust` option can be added to a `test` command if a user does not desire the additional adjustment to the test statistic. The multiparameter Wald test for all of the newly added interaction parameters essentially amounts to a design-based test of *change in R-squared* for comparing nested models (where in this case, one model includes the interactions, and one does not). The adjusted Wald test performed by Stata indicates that we can reject this null hypothesis ($F(8,10) = 7.00$, $p < 0.01$), which suggests that adding the interactions between both age terms and ethnicity is significantly improving the fit of the model (we will return to their interpretation shortly).

Next, we add the two-way interactions between the two age terms and gender (RIAGENDR) to the model, and again test the associated parameters using an adjusted Wald test:

```
svy, subpop(age18p): regress bpxdi1_1 i.RIDRETH1 ///
i.riagendr agec agecsq ///
i.RIDRETH1#c.agec i.RIDRETH1#c.agecsq ///
i.riagendr#c.agec i.riagendr#c.agecsq

test 2.riagendr#c.agec 2.riagendr#c.agecsq
```

This Wald test once again suggests that at least one of the two regression parameters associated with the interactions between the age terms and gender is significantly different from zero ($F(2,16) = 4.83$, $p = 0.03$), so we have evidence in favor of the model including these interactions as well. Readers can use similar methods to test interactions between two (or more) categorical predictors.

After fitting this latest model including the new interaction terms, we can plot *marginal predicted values* for DBP based on the fitted model to assess variability in the relationship of age with DBP depending on race/ethnicity and gender. In order to do this, we first refit the model specifying an "interaction" between the original centered age variable and itself, so that Stata does not simply interpret the squared age term from the previous syntax as an arbitrary additional predictor, and instead knows that there is a curvilinear relationship between centered age and the outcome:

```
svy, subpop(age18p): regress bpxdi1_1 i.RIDRETH1 ///
i.riagendr agec c.agec#c.agec ///
i.RIDRETH1#c.agec i.RIDRETH1#c.agec#c.agec ///
i.riagendr#c.agec i.riagendr#c.agec#c.agec
```

While this syntax may seem tedious, it is essential for correctly plotting curvilinear relationships involving continuous predictor variables. We generate the plots of marginal predicted values for DBP using the following syntax:

```
margins RIDRETH1, at(agec=(-30(5)30))
marginsplot
margins riagendr, at(agec=(-30(5)30))
marginsplot
```

This `margins` syntax indicates that we wish to display marginal predicted values of DBP for each level of RIDRETH1 (plot 1) and RIAGENDR (plot 2), separately at all possible increments of five units in centered age (from –30 to 30; i.e., –30, –25, –20, ..., 25, 30). The marginal predicted values are computed by default in Stata by: (1) assuming that everyone in the data set has a particular value of RIDRETH1 (or RIAGENDR) and centered age, (2) computing the predicted values based on the fitted model for each case in the data set, and then (3) averaging the predictions. We then plot these marginal predicted values (including 95% confidence intervals for the predictions) using the `marginsplot` command. The resulting plots are displayed in Figure 7.4.

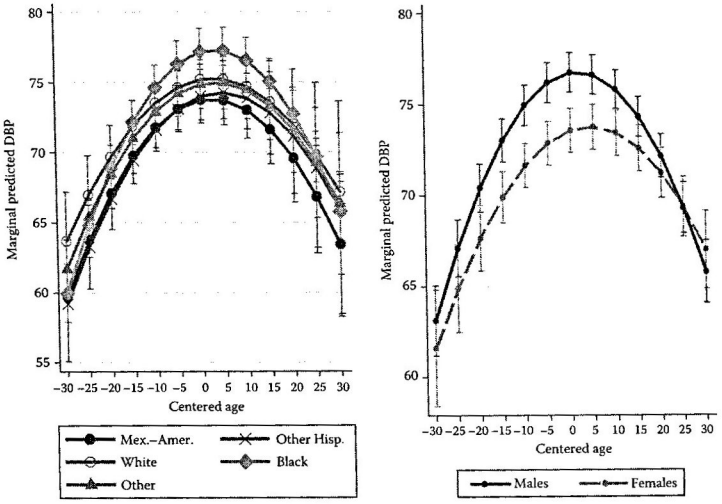


**FIGURE 7.4**
Plots of marginal predicted values based on the regression model, including interactions between age and race/ethnicity as well as age and gender, illustrating differences in the curvilinear relationship of centered age with diastolic blood pressure depending on race/ethnicity and gender.

Figure 7.4 shows that African-Americans have a higher acceleration in DBP as a function of age, up until about middle age. The same can be said for males, where the gap between males and females in marginal predicted values widens approaching middle age, and then narrows in older ages. While these differences are subtle, they are significant (as shown earlier).

We now carefully consider diagnostics for this latest model. We first refit the model, and then save variables containing residuals (EHAT1) and predicted values (YHAT1) based on the fitted model. Next, we generate a series of diagnostic plots to assess the assumptions underlying the model:

```
svy, subpop(age18p): regress bpxdi1_1 i.RIDRETH1 ///
i.riagendr agec c.agec#c.agec ///
i.RIDRETH1#c.agec i.RIDRETH1#c.agec#c.agec ///
i.riagendr#c.agec i.riagendr#c.agec#c.agec

predict ehat1, resid

symplot ehat1, name(sym_ehat1_1, replace) ///
title(Symplot of Residuals)

histogram ehat1, normal name(h_ehat1, replace) ///
title(Histogram of Residuals)

qnorm ehat1, name(qnorm_ehat1, replace) ///
```

```
title(Normal Q-Q Plot of Residuals)

predict yhat1, xb

scatter ehat1 yhat1, name(ehat1xyhat1, replace) ///
title(Residuals vs. Predicted Y)

graph combine sym_ehat1_1 h_ehat1 qnorm_ehat1 ///
ehat1xyhat1, rows(2)
```

We introduce one additional diagnostic plot that can be helpful for assessing model fit. The `symplot` command in Stata produces a symmetry plot, which is useful for determining whether the distribution of values on a given continuous variable appears to be symmetric in nature. Specifically, the distance below the median of the first-ordered value in the distribution is plotted against the distance above the median of the last-ordered value; the distance below the median of the second-ordered value is plotted against the distance above the median of the second-to-last-ordered value; and so forth. If the values in the symmetry plot lie on the straight diagonal line, there is evidence that the distribution is symmetric. The resulting plot in the upper-left panel of Figure 7.5 suggests that the distribution of the residuals based on the final fitted model is nearly symmetric around 0 (allowing for a handful of outliers), alleviating possible concerns about slight deviations from normality.

Collectively, the diagnostic plots presented in Figure 7.5 suggest that the residuals follow a symmetric distribution, and that assumptions of normality and constant variance for the residuals definitely seem reasonable. Given any apparent violations of normality in the distribution of the residuals, the symmetry of the residuals around the expected value of 0 gives us confidence in the inferences that we are making. However, the diagnostic assessments that we have performed so far have neglected to account for the complex sampling features.

We now examine the residuals and check for the possibility of influential points and outliers in a manner that accounts for the complex sampling features. We do this using state-of-the-art diagnostic techniques for linear regression models fitted to complex sample survey data, which were discussed earlier in Section 7.3.3. Several of these diagnostic tools have been implemented in the contributed R package `svydiags`, which at the time of this writing is available upon request in .zip format from Rick Valliant (rvalliant@survey.umd.edu). We will provide updates on the book web page when this contributed package is available on the Comprehensive R Archive Network (CRAN). First, we load the contributed `survey` package in R, assuming that it has already been installed from a CRAN mirror:

```
library(survey)
```

Next, we install the local .zip file containing the `svydiags` functions, which can be done using the Packages menu in the R GUI (given that the .zip

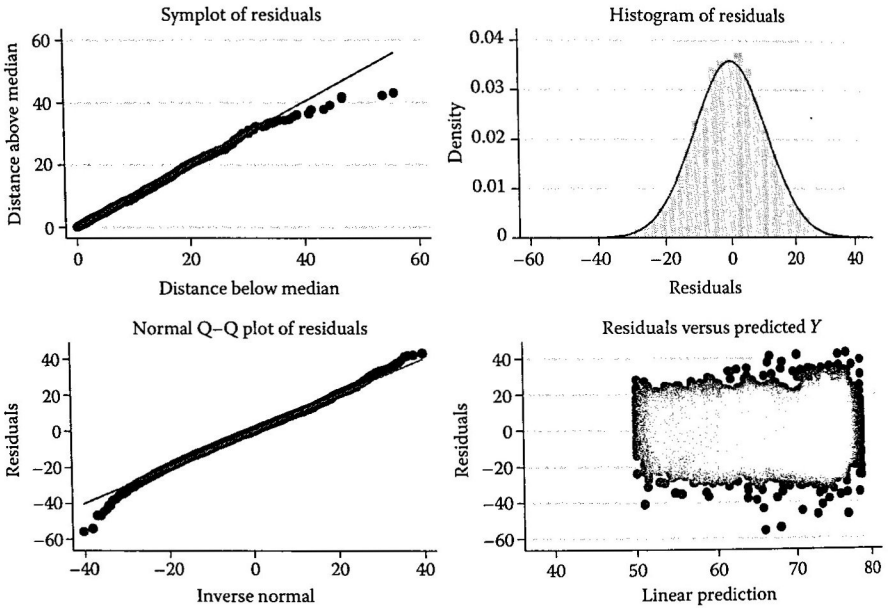


**FIGURE 7.5**
Diagnostic plots for the "final" regression model fitted to the diastolic blood pressure response variable in the 2011–2012 NHANES data set.

archive has been saved in some working directory), and load the `svydiags` package. We then load the 2011–2012 NHANES data, recode the dependent variable, compute the centered age variable, create a survey design object describing the complex sampling features, refit the final regression model discussed above using the `svyglm()` function, and then generate the parameter estimates, standard errors, and tests of significance for the parameters, using the same approximate degrees of freedom method (number of clusters minus number of strata) employed by Stata:

```
library(svydiags)

load("C:\\nhanes1112.rdata")

nhanes1112$bpxdi1.1 <- nhanes1112$BPXDI1
nhanes1112$bpxdi1.1[nhanes1112$BPXDI1 == 0] <- NA
nhanes1112$agec <- nhanes1112$age - 46.36
nhanes1112$agec2 <- nhanes1112$agec ^ 2

dnhanes <- svydesign(id =~ sdmvpsu, strata =~ sdmvstra,
weights =~ WTMEC2YR, nest = TRUE, data = nhanes1112)

finmod <- svyglm(bpxdi1.1 ~ as.factor(RIDRETH1) +
as.factor(riagendr) + agec + agec2 + as.factor(RIDRETH1):agec
```

```
+ as.factor(RIDRETH1):agec2 + as.factor(riagendr):agec +
as.factor(riagendr):agec2, subset = (age18p == 1), design =
dnhanes)

summary(finmod, df.resid = degf(dnhanes))
```

Now that we have replicated the model-fitting process in R and created an object named `finmod` containing all of the relevant information about the model, we can first employ the `svyCooksD()` function in the `svydiags` package to generate the modified Cook's D statistic for each case, measuring the influence of that case on the parameter estimates in the final model:

```
mcook <- svyCooksD(mobj = finmod, stvar = "sdmvstra", clvar =
"sdmvpsu", doplot = TRUE)
```

The plot of the modified Cook's D statistics for each case generated by the `doplot = TRUE` argument is shown in Figure 7.6. The two horizontal lines indicate rules of thumb for which cases have overly large influence on the parameter estimates in the model (values of 2 or 3). We see that several cases fall above the highest horizontal line, which warrants refitting the model excluding some of these extreme cases.

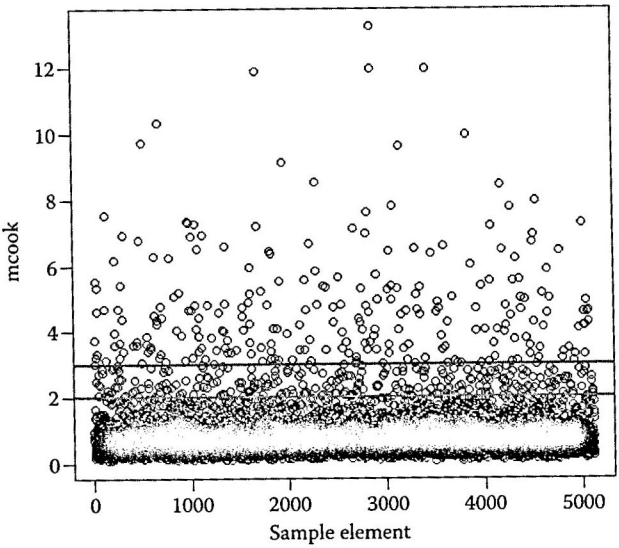


**FIGURE 7.6**
Plot of modified Cook's D statistics for the $n = 5112$ cases analyzed in the final model, demonstrating the influence of each individual case and accounting for the complex sampling features of the 2011–2012 NHANES.

We specifically identify the four cases with the largest modified Cook's D statistics, and create a temporary data frame object excluding these specific rows:

```
mcook[mcook > 11]
    3257     5529     5543     6624
11.88683 11.97656 13.26392 11.97656
nhanes1112a <- nhanes1112[-c(3257,5529,5543,6624),]
```

We then create a new survey design object using this new data frame, and refit the model excluding these cases:

```
dnhanes2 <- svydesign(id =~ sdmvpsu, strata =~ sdmvstra,
weights =~ WTMEC2YR, nest = TRUE, data = nhanes1112a)

finmod2 <- svyglm(bpxdi1.1 ~ as.factor(RIDRETH1) +
as.factor(riagendr) + agec + agec2 + as.factor(RIDRETH1):agec
+ as.factor(RIDRETH1):agec2 + as.factor(riagendr):agec +
as.factor(riagendr):agec2, subset = (age18p == 1), design =
dnhanes2)

summary(finmod2, df.resid = degf(dnhanes2))
```

The results of this analysis (not shown) suggest slight shifts in the resulting parameter estimates, but none large enough to cause us to change our inferences. One could explore the impact of removing additional cases with large values of Cook's statistic at this point, but we do not consider these further. In practice, if the exclusion of a small number of influential points results in different inferences, a sound argument would be needed for why those individual cases should be removed from the analysis (e.g., extreme values for particular variables).

Next, we examine design-based calculations of dfbetas, which indicate the influence of individual observations on the parameter estimates in the final model. Using the original model fit object, we compute these values using the following function in the `svydiags` package:

```
dfbetas <- svydfbetas(mobj = finmod, stvar = "sdmvstra", clvar
= "sdmvpsu")
```

The resulting object shows, for each case used to fit the model, the changes in each of the parameter estimates that would occur if a given case was excluded from the analysis. We examine these changes for the case in row 5543, which was found above to have the largest modified Cook's D statistic:

```
b <- data.frame(dfbetas$Dfbetas)

b$X5543
```

```
[1] -0.0012526839  0.0015667514  0.0015432959  0.0013492664
0.0014159274
[6] -0.0007269432 -0.0163536130 -0.0118780945  0.0125492120
0.0118395900
[11]  0.0162467031  0.0127952869  0.0055352874  0.0113366512
0.0099378641
[16]  0.0066420863  0.0019866268  0.0021669724
```

The two boldfaced values above correspond to the changes in the estimates of the centered age and centered age squared parameters (the 7th and 8th parameters in the model output) that would occur from dropping this one case. These examinations can therefore be used to quantify the impacts that individual cases are having on the parameter estimates, fully accounting for the complex sampling features in the calculations. We recommend performing these analyses alongside an examination of the Cook's D statistics to describe the impact that a given case is having on the parameter estimates if that case is being considered for exclusion from the analysis.

Finally, we compute standardized residuals based on the final fitted model, or residuals divided by the standard deviation of the residuals based on the fitted model (again accounting for the complex sampling features in the calculations). We can generate the standardized residuals by using the `svystdres()` function in the `svydiags` package:

```
st.resids <- svystdres(mobj = finmod, stvar = "sdmvstra",
clvar = "sdmvpsu", doplot = TRUE)
```

The plot of the standardized residuals that is automatically generated when using the `doplot = TRUE` argument (Figure 7.7) draws horizontal lines at values of –3 and 3 for the standardized residuals, where cases with residuals outside of this range would normally be considered outliers.

We identify those cases with standardized residuals less than –4:

```
st.resids$stdresids[st.resids$stdresids < -4]
     2735      3732      3138      1158      5815
-4.183308 -4.995978 -4.182579 -4.000169 -4.817674
```

These are cases that are not being fitted adequately by the final model. For example, the case in row 3732 was 80 years old and had a recorded DBP of 10 (!). Careful assessment of these outliers (accounting for the complex sampling features) can lead to the identification of situations like this, where there may have been data entry errors. If a value of 10 is unrealistic or not possible, this case could be dropped from the analysis moving forward.

After assessing the model diagnostics carefully (accounting for the complex sampling features if possible) and possibly making changes to the model specification or dropping overly influential cases or extreme outliers (with discretion), the focus can turn to interpretation of the parameter

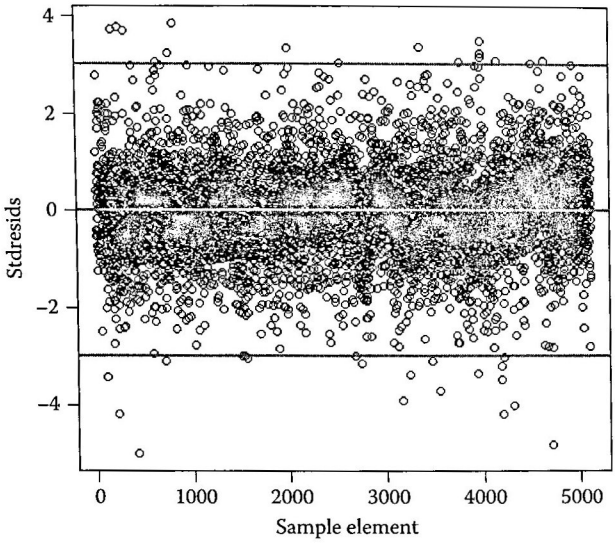


**FIGURE 7.7**
Plot of standardized residuals for the $n = 5112$ cases analyzed in the final model, accounting for the complex sampling features of the 2011–2012 NHANES.

estimates in the final model. Table 7.6 presents the parameter estimates and associated tests of significance in the "final" model for this example. The estimates in Table 7.6 provide strong evidence of the quadratic relationship of age with DBP when adjusting for other sociodemographic features, and also show how there are still strong ethnicity and gender effects on BP. We once again note the substantial improvement in the model $R$-squared values (from 0.017 to 0.120) due to the inclusion of the squared age term and the additional interactions. In addition, while most of the design effects are greater than 1, suggesting losses in the efficiency of the estimates given the complex sampling employed by NHANES, there are examples of estimates with *increased* efficiency due to the complex sampling features (likely due to effective stratification). Clear interpretation of the estimated parameters is essential, however, to understand these design effects further.

If we wish to describe our "final" interpretation of the parameter estimates in this model, we need to be very careful with the estimated interaction coefficients. Importantly, the estimated coefficients for ethnicity, gender, and the two age predictors are *not* main effects for the entire population. Instead, because of the interactions included in the final model, these coefficients correspond to subgroups defined by the *reference categories* (or "0" values) of other variables involved in the interaction. So, the four estimated coefficients for the nonreference race/ethnicity categories correspond to race/ethnicity effects *specifically for individuals with a value of 0 on the two age variables* (or individuals with age equal to the population mean, given the centered age variable). It would therefore be erroneous to say, based on Table 7.6,

**TABLE 7.6**

Estimates of the Regression Parameters in the Final Model for the Diastolic Blood Pressure Response Variable, Including Significant Interaction Terms (Estimates and Standard Errors Based on the "q-Weighted" Approach Outlined by Pfeffermann (2011) are in Parentheses)

| Predictor | Est. | Linearized SE | *t*-Statistic (*df*) | *p*-Value | 95% CI | DEFF |
|---|---|---|---|---|---|---|
| Intercept | 75.346 (75.41) | 0.819 (0.772) | 91.99 (17) | <0.001 | (73.618, 77.075) | 1.43 |
| *Ethnicity* | | | | | | |
| Other Hispanic | 0.271 (0.25) | 0.921 (0.954) | 0.29 (17) | 0.772 | (–1.672, 2.215) | 0.91 |
| Non-Hispanic White | 1.461 (1.50) | 0.910 (0.895) | 1.60 (17) | 0.127 | (–0.460, 3.382) | 1.70 |
| Non-Hispanic Black | 3.450 (3.57) | 0.961 (0.989) | 3.59 (17) | 0.002 | (1.422, 5.478) | 1.16 |
| Other race | 1.144 (1.24) | 0.895 (0.892) | 1.28 (17) | 0.218 | (–0.744, 3.032) | 0.98 |
| Mexican-American | – | – | – | – | – | – |
| *Gender* | | | | | | |
| Female | –3.195 (–3.43) | 0.759 (0.633) | –4.21 (17) | 0.001 | (–4.797, –1.593) | 4.19 |
| Male | – | – | – | – | – | – |
| **Age (centered)** | 0.039 (0.05) | 0.040 (0.040) | 0.99 (17) | 0.332 | (–0.045, 0.123) | 1.20 |
| **Age (cent.) squared** | –0.015 (–0.02) | 0.002 (0.002) | –8.43 (17) | <0.001 | (–0.019, –0.011) | 0.82 |
| *Age × ethnicity* | | | | | | |
| Age × Other Hisp. | 0.050 (0.05) | 0.050 (0.047) | 1.00 (17) | 0.332 | (–0.055, 0.154) | 1.07 |
| Age × White | –0.004 (–0.01) | 0.053 (0.051) | –0.08 (17) | 0.934 | (–0.117, 0.108) | 2.07 |
| Age × Black | 0.035 (0.036) | 0.039 (0.037) | 0.89 (17) | 0.385 | (–0.047, 0.116) | 0.78 |
| Age × Other | 0.015 (0.01) | 0.049 (0.046) | 0.30 (17) | 0.766 | (–0.089, 0.119) | 1.13 |
| *Age sq. × ethnicity* | | | | | | |
| Age sq. × Other Hisp. | 0.001 (0.001) | 0.003 (0.003) | 0.24 (17) | 0.811 | (–0.006, 0.008) | 1.57 |
| Age sq. × White | 0.003 (0.003) | 0.002 (0.002) | 1.55 (17) | 0.139 | (–0.001, 0.006) | 0.71 |
| Age sq. × Black | –0.002 (–0.01) | 0.002 (0.002) | –1.19 (17) | 0.249 | (–0.007, 0.002) | 0.66 |
| Age sq. × Other | 0.001 (0.001) | 0.001 (0.003) | 0.48 (17) | 0.634 | (–0.005, 0.008) | 1.31 |
| *Age × gender* | | | | | | |
| Age × female | 0.045 (0.034) | 0.023 (0.026) | 1.94 (17) | 0.069 | (–0.004, 0.095) | 2.37 |
| *Age Sq. × gender* | | | | | | |
| Age Sq. × female | 0.003 (0.003) | 0.002 (0.002) | 2.04 (17) | 0.058 | (–0.001, 0.007) | 3.46 |

*Note:* Subclass $n = 5{,}112$, weighted $R^2 = 0.120$, adjusted Wald test for all parameters: $F(17,1) = 176.18$, $p = 0.059$. – denotes the reference category.

that African-Americans have a mean DBP that is higher by 3.45 units than Mexican-Americans *in general;* this interpretation is only true *for individuals with age equal to the population mean.*

In the same manner, one would interpret the estimated coefficients for age and age-squared not as main effects for the entire population, but rather effects of age *for the reference categories of race/ethnicity and gender.* That is, for Mexican-American males, the relationship of age with DBP is defined by a quadratic equation with intercept 75.35, linear age coefficient 0.039 (not significantly different from zero), and quadratic age coefficient −0.015 (significantly different from zero). The estimated interaction coefficients then represent *changes* in these age coefficients corresponding to the nonreference categories of race/ethnicity and gender. So why did we find in the Wald test that the overall interaction terms involving race/ethnicity were significant, given that none of these "change" coefficients seem to be different from zero in Table 7.6? Keep in mind that we are looking at changes in relationships relative to *specific* reference categories. It would appear that the difference in the quadratic age coefficients for whites and African-Americans is largest, but we are not getting a test of this difference (given that Mexican-American is the reference category). If we were to change the reference category to white, the change coefficient for African-Americans would now be positive and significant (at the 5% level).

In general, if one finds significant interactions in a linear regression model (or any regression model for that matter), very interesting insights are possible regarding subgroup differences in the relationships of interest. Great care is simply needed in interpreting these relationships.

Finally, we conclude this illustration by applying the "q-weighted" approach proposed by Pfeffermann (2011) to our "final" model, to determine whether we can generate similar weighted estimates with increased efficiency. This approach proceeds as follows:

1. Fit a regression model to the final survey weights using the predictor variables in the regression model of substantive interest. [We note that under mild regularity conditions, this "q-weighted" approach will ultimately yield consistent estimates of the finite population regression parameters even if this initial regression model for the weights is misspecified in some way; see Pfeffermann (2011, p. 124).] In our example:

```
regress WTMEC2YR i.RIDRETH1 i.riagendr agec
```

2. Save the expected values (i.e., predicted values) of the survey weights for each case as a function of the predictor variables in the data set:

```
predict w_hat, xb
```

3. Divide the final analysis weights by their expected values as a function of the predictor variables:

```
gen q_WTMEC2YR = WTMEC2YR / w_hat
```

4. Use the new "q-weights" as the survey weights in estimating the final regression model:

```
svyset sdmvpsu [pweight = q_WTMEC2YR], ///
strata(sdmvstra) vce(linearized) singleunit(missing)

svy, subpop(age18p): regress bpxdi1_1 i.RIDRETH1 ///
i.riagendr agec c.agec#c.agec ///
i.RIDRETH1#c.agec i.RIDRETH1#c.agec#c.agec ///
i.riagendr#c.agec i.riagendr#c.agec#c.agec
```

We compare weighted estimates and standard errors based on this approach with those based on the original weights in Table 7.6. We note that these estimates tend to have higher efficiency than those based on the original weights, but the differences in this example are not substantial. Given the ease of implementing this approach and its demonstrated empirical benefits in terms of efficiency (Pfeffermann, 2011), we recommend that analysts fitting regression models to complex sample survey data consider applying it before presenting any final estimates.

## EXERCISES

7.1 **Regression Model Building:**

a. Briefly outline and describe the four model-building steps recommended in this chapter.

b. Describe how each of the four steps changes when analyzing complex sample survey data.

7.2 **Linear Regression Analysis Project Using ESS6 Russian Federation Data:** This exercise asks you to perform a linear regression analysis using the ESS Round 6 Russian Federation data. Exercise 7.2 contains many parts and emphasizes model-building and correct analysis of data collected from a complex sample survey. The data set, *chapter_exercises_ess6ru*, is available for download from the ASDA web site, in SAS or Stata format.

a. Perform data exploration of the Satisfaction with Government (STFGOV) variable through use of a weighted frequency table, weighted bar chart, or a weighted histogram. This will be the dependent variable in the linear regression exercises to follow. Discuss the distribution of the variable and explain how this variable can function as a continuous outcome despite consisting of 10 distinct levels.

b. Consider three possible predictors of the dependent variable Satisfaction with the Government (STFGOV): (1) a left-to-right

political scale collapsed into three categories, from the 0–10 scale LRSCALE (LR3CAT, 1 = left to moderately left, 2 = moderate, 3 = moderately right to right); (2) gender (GNDR, 1 = Male, 2 = Female); and (3) Self-rated health status (HEALTH, 1 = Very good, 2 = Good, 3 = Fair, 4 = Bad, 5 = Very bad). Describe how each predictor variable should be handled in the regression model (continuous or categorical).

c. Fit a set of bivariate models where STFGOV is predicted by each possible predictor listed in part b (one at a time), while incorporating the complex sample design features (STRATIFY and PSU) and the final survey weights (PSPWGHT). Based on these results, what are the *F*-test statistics, degrees of freedom, and *p*-values for each possible predictor variable?

d. Fit the "preliminary" model with the retained predictors (from part c), fully accounting for the complex sampling features. Request design effects for the estimated coefficients, and produce a residual versus predicted plot and a histogram with a superimposed normal curve of the residuals. Based on these results, answer the following questions:

   i. Do these plots indicate acceptable model fit?

   ii. What other techniques might be used to assess model fit?

   iii. What do the design effects indicate about the relationships of each predictor?

   iv. How many degrees of freedom are used for the significance tests, and how are they calculated?

   v. What is the *R*-squared for this model and what does this mean?

e. Test the interaction of gender and left/right political orientation (three category variables) in the "preliminary" model. Then, answer these questions:

   i. What are the *F*-test statistic, degrees of freedom, and *p*-value for the interaction?

   ii. Is the interaction term significant at the $\alpha = 0.05$ level?

   iii. Should the interaction be retained in your final model?

f. Fit your "final" model, and based on these results, prepare a table similar to Table 7.5. Be sure to either exclude or include your interaction terms depending on the decision you made in part e. Include variable labels, design effects, and descriptive columns in the table.

g. Fit your "final" model using the final survey weights, but without applying the complex sample design adjustments to the standard errors (i.e., ignoring PSU and STRATIFY in the variance

estimation). Based on these results, prepare a table similar to Table 7.3. Make sure to label the standard error column as either "robust standard error" or "standard error" depending on how your software of choice handles weighted but non-design-based standard error calculations. Do any of your inferences change? Would any of your inferences change if you ignored the weights as well?

h. Write a brief paragraph (as you might for a research article) explaining the methods that you used to estimate the model parameters and estimate the standard errors of the estimated parameters in a way that accounts for the complex sample design of the ESS6 Russian Federation data. Also, discuss how your conclusions might have changed had you not performed a correct design-based analysis incorporating the complex sample design features and the survey weights.