Reserve Services

Northwestern
University
Library

WARNING

CONCERNING COPYRIGHT RESTRICTIONS

The Copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or reproduction.

One of three specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research.

If electronic transmission of reserve material is used for purposes in excess of what constitutes “fair use”, that user may be liable for copyright infringement.

This policy is in effect for the following document:

Ratner, Bruce
Ordinary Regression: The Workhorse of Profit Modeling (Chapter 9) / from Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data
Boca Raton, FL: CRC Press, 2012. pp. 153-176.

NO FURTHER TRANSMISSION OR DISTRIBUTION OF THIS MATERIAL IS PERMITTED

Statistical and Machine-Learning Data Mining

Techniques for Better Predictive Modeling and Analysis of Big Data

Second Edition

Bruce Ratner

CRC Press
Taylor & Francis Group
Boca Raton London New York

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

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

© 2012 by Taylor and Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

International Standard Book Number: 978-1-4398-6091-5 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Ratner, Bruce.
Statistical and machine-learning data mining : techniques for better predictive modeling and analysis of big data / Bruce Ratner. -- 2nd ed.
p. cm.
Rev. ed. of: Statistical modeling and analysis for database marketing. c2003.
Includes bibliographical references and index.
ISBN 978-1-4398-6091-5
1. Database marketing--Statistical methods. 2. Data mining--Statistical methods. I. Ratner, Bruce. II. Ratner, Bruce. Statistical modeling and analysis for database marketing. III. Title.

HF5415.126.R38 2012
658.8'72--dc22 2011014298

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

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

9

Ordinary Regression: The Workhorse of Profit Modeling

9.1 Introduction

Ordinary regression is the popular technique for predicting a quantitative outcome, such as profit and sales. It is considered the workhorse of *profit* modeling as its results are taken as the gold standard. Moreover, the ordinary regression model is used as the benchmark for assessing the superiority of new and improved techniques. In a database marketing application, an individual's profit* to a prior solicitation is the quantitative dependent variable, and an ordinary regression model is built to predict the individual's profit to a future solicitation.

I provide a brief overview of ordinary regression and include the SAS program for building and scoring an ordinary regression model. Then, I present a mini case study to illustrate that the data mining techniques presented in Chapter 8 carry over with minor modification to ordinary regression. Model builders, who are called on to provide statistical support to managers monitoring expected revenue from marketing campaigns, will find this chapter an excellent reference for profit modeling.

9.2 Ordinary Regression Model

Let Y be a quantitative dependent variable that assumes a continuum of values. The ordinary regression model, formally known as the ordinary least squares (OLS) regression model, predicts the Y value for an individual based on the values of the predictor (independent) variables $X_1, X_2, \ldots, X_n$ for that individual. The OLS model is defined in Equation (9.1):

* Profit is variously defined as any measure of an individual's valuable contribution to the bottom line of a business.

$$Y = b_0 + b_1{*}X_1 + b_2{*}X_2 + \ldots + b_n{*}X_n \tag{9.1}$$

An individual's predicted Y value is calculated by "plugging in" the values of the predictor variables for that individual in Equation (9.1). The b's are the OLS regression coefficients, which are determined by the calculus-based method of least squares estimation; the lead coefficient b_0 is referred to as the intercept.

In practice, the quantitative dependent variable does not have to assume a progression of values that vary by minute degrees. It can assume just several dozens of discrete values and work quite well within the OLS methodology. When the dependent variable assumes only two values, the logistic regression model, not the ordinary regression model, is the appropriate technique. Even though logistic regression has been around for 60-plus years, there is some misunderstanding over the practical (and theoretical) weakness of using the OLS model for a binary response dependent variable. Briefly, an OLS model, with a binary dependent variable, produces typically some probabilities of response greater than 100% and less than 0% and does not typically include some important predictor variables.

9.2.1 Illustration

Consider dataset A, which consists of 10 individuals and three variables (Table 9.1): the quantitative variable PROFIT in dollars (Y), INCOME in thousands of dollars (X1), and AGE in years (X2). I regress PROFIT on INCOME and AGE using dataset A. The OLS output in Table 9.2 includes the ordinary regression coefficients and other "columns" of information. The "Parameter Estimate" column contains the coefficients for INCOME and AGE variables, and the intercept. The coefficient b0 for the intercept variable is used as a

TABLE 9.1

Dataset A

Profit ($)	Income ($000)	Age (years)
78	96	22
74	86	33
66	64	55
65	60	47
64	98	48
62	27	27
61	62	23
53	54	48
52	38	24
51	26	42

TABLE 9.2

OLS Output: PROFIT with INCOME and AGE

Source	df	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	460.3044	230.1522	6.01	0.0302
Error	7	268.0957	38.2994		
Corrected total	9	728.4000			
		Root MSE	6.18865	R-square	0.6319
		Dependent mean	62.60000	Adj R-Sq	0.5268
		Coeff var	9.88602		

Variable	df	Parameter Estimate	Standard Error	t Value	Pr > |t|
INTERCEPT	1	52.2778	7.7812	7.78	0.0003
INCOME	1	0.2669	0.2669	0.08	0.0117
AGE	1	-0.1622	-0.1622	0.17	0.3610

"start" value given to all individuals, regardless of their specific values of the predictor variables in the model.

The estimated OLS PROFIT model is defined in Equation (9.2):

$$\text{PROFIT} = 52.2778 + 0.2667*\text{INCOME} - 0.1622*\text{AGE} \qquad (9.2)$$

9.2.2 Scoring an OLS Profit Model

The SAS program (Figure 9.1) produces the OLS profit model built with dataset A and scores the external dataset B in Table 9.3. The SAS procedure REG produces the ordinary regression coefficients and puts them in the "ols_coeff" file, as indicated by the code "outest = ols_coeff." The ols_coeff file produced by SAS is in Table 9.4.

The SAS procedure SCORE scores the five individuals in dataset B using the OLS coefficients, as indicated by the code "score = ols_coeff." The procedure appends the predicted Profit variable in Table 9.3 (called pred_Profit as indicated by "pred_Profit" in the second line of code in Figure 9.1) to the output file B_scored, as indicated by the code "out = B_scored."

9.3 Mini Case Study

I present a "big" discussion on ordinary regression modeling with the mini dataset A. I use this extremely small dataset not only to make the discussion of data mining techniques tractable but also to emphasize two aspects

```
/****** Building the OLS Profit Model on dataset A ************/
PROC REG data = A outest = ols_coeff;
pred_Profit: model Profit =
Income Age;
run;

/****** Scoring the OLS Profit Model on dataset B ************/
PROC SCORE data = B predict type = parms score = ols_coeff
out = B_scored;
var Income Age;
run;
```

FIGURE 9.1
SAS program for building and scoring OLS profit model.

TABLE 9.3
Dataset B

Income ($000)	Age (years)	Predicted Profit ($)
148	37	85.78
141	43	82.93
97	70	66.81
90	62	66.24
49	42	58.54

TABLE 9.4
OLS_Coeff File

OBS	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	Income	Age	Profit
1	est_Profit	PARMS	Profit	6.18865	52.52778	0.26688	-0.16217	-1

of data mining. First, data mining techniques of great service should work as well with small data as with big data, as explicitly stated in the definition of data mining in Chapter 1. Second, every fruitful effort of data mining on small data is evidence that big data are not always necessary to uncover structure in the data. This evidence is in keeping with the EDA (exploratory data analysis) philosophy that the data miner should work from simplicity until indicators emerge to go further. If predictions are not acceptable, then increase data size.

The objective of the mini case study is as follows: to build an OLS Profit model based on INCOME and AGE. The ordinary regression model (celebrating over 200 years of popularity since the invention of the method of least squares on March 6, 1805) is the quintessential linear model, which implies the all-important assumption: The underlying relationship between a given predictor variable and the dependent variable is linear. Thus, I use the method of smoothed scatterplots, as described in Chapter 2, to determine whether the linear assumption holds for PROFIT with INCOME and with AGE. For the mini dataset, the smoothed scatterplot is defined by 10 slices, each of size 1. Effectively, the smooth scatterplot is the simple scatterplot of 10 paired (PROFIT, Predictor Variable) points. (Contrasting note: the logit plot as discussed with the logistic regression in Chapter 8 is neither possible nor relevant with OLS methodology. The quantitative dependent variable does not require a transformation, like converting logits into probabilities as found in logistic regression.)

9.3.1 Straight Data for Mini Case Study

Before proceeding with the analysis of the mini case study, I clarify the use of the bulging rule when analysis involves OLS regression. The bulging rule states that the model builder should try reexpressing the predictor variables as well as the dependent variable. As discussed in Chapter 8, it is not possible to reexpress the dependent variable in a logistic regression analysis. However, in performing an ordinary regression analysis, reexpressing the dependent variable is possible, but the bulging rule needs to be supplemented. Consider the illustration discussed next.

A model builder is building a profit model with the quantitative dependent variable Y and three predictor variables X_1, X_2, and X_3. Based on the bulging rule, the model builder determines that the powers of ½ and 2 for Y and X_1, respectively, produce a very adequate straightening of the Y-X_1 relationship. Let us assume that the correlation between the square root Y (sqrt_Y) and square of X_1 (sq_X_1) has a reliable $r_{\text{sqrt\_Y, sq\_X1}}$ value of 0.85.

Continuing this scenario, the model builder determines that the powers of 0 and ½ and -½ and 1 for Y and X_2 and Y and X_3, respectively, also produce very adequate straightening of the Y-X_2 and Y-X_3 relationships, respectively. Let us assume the correlations between the log of Y (log_Y) and the square root of X_2 (sq_X_2) and between the negative square root of Y (negsqrt_Y) and X_3 have reliable $r_{\text{log\_Y, sq\_X1}}$ and $r_{\text{negsqrt\_Y, X3}}$ values of 0.76 and 0.69, respectively. In sum, the model builder has the following results:

1. The best relationship between square root Y (p = ½) and square of X_1 has $r_{\text{sqrt\_Y, sq\_X1}} = 0.85$.
2. The best relationship between log of Y (p = 0) and square root of X_2 has $r_{\text{log\_Y, sq\_X2} =}$ 0.76.
3. The best relationship between the negative square root of Y (p = -½) and X_3 has $r_{\text{neg\_sqrt\_Y, X3}} = 0.69$.

In pursuit of a good OLS profit model, the following guidelines have proven valuable when several reexpressions of the quantitative dependent variable are suggested by the bulging rule.

1. If there is a small range of *dependent-variable powers* (powers used in reexpressing the dependent variable), then the best reexpressed dependent variable is the one with the noticeably largest correlation coefficient. In the illustration, the best reexpression of Y is the square root Y: Its correlation has the largest value: $r_{\text{sqrt\_Y, sq\_X1}}$ equals 0.85. Thus, the data analyst builds the model with the square root Y and the square of X_1 and needs to reexpress X_2 and X_3 again with respect to the square root of Y.
2. If there is a small range of dependent-variable powers and the correlation coefficient values are comparable, then the best reexpressed dependent variable is defined by the *average power* among the dependent variable powers. In the illustration, if the model builder were to consider the r values (0.85, 0.76, and 0.69) comparable, then the average power would be 0, which is one of the powers used. Thus, the modeler builds the model with the log of Y and square root of X_2 and needs to reexpress X_1 and X_3 again with respect to the log of Y.

   If the average power were not one of the dependent-variable powers used, then all predictor variables would need to be reexpressed again with the newly assigned reexpressed dependent variable, Y raised to the "average power."
3. When there is a large range of dependent-variable powers, which is likely when there are many predictor variables, the practical and productive approach to the bulging rule for building an OLS profit model consists of initially reexpressing only the predictor variables, leaving the dependent variable unaltered. Choose several handfuls of reexpressed predictor variables, which have the largest correlation coefficients with the unaltered dependent variable. Then, proceed as usual, invoking the bulging rule for exploring the best reexpressions of the dependent variable and the predictor variables. If the dependent variable is reexpressed, then apply steps 1 or 2.

Meanwhile, there is an approach considered the most desirable for picking out the best reexpressed quantitative dependent variable; it is, however, neither practical nor easily assessable. It is outlined in Tukey and Mosteller's *Data Analysis and Regression* ("Graphical Fitting by Stages," pages 271–279). However, this approach is extremely tedious to perform manually as is required because there is no commercially available software for its calculations. Its inaccessibility has no consequence to the model builders' quality of model, as the approach has not provided noticeable improvement over the procedure in step 3 for marketing applications where models are implemented at the decile level.

Now that I have examined all the issues surrounding the quantitative dependent variable, I return to a discussion of reexpressing the predictor variables, starting with INCOME and then AGE.

9.3.1.1 Reexpressing INCOME

I envision an underlying positively sloped straight line running through the 10 points in the PROFIT-INCOME smooth plot in Figure 9.2, even though the smooth trace reveals four severe kinks. Based on the general association test with the test statistic (TS) value of 6, which is *almost* equal to the cutoff score of 7, as presented in Chapter 2, I conclude there is an *almost noticeable* straight-line relationship between PROFIT and INCOME. The correlation coefficient for the relationship is a reliable $r_{\text{PROFIT, INCOME}}$ of 0.763. Notwithstanding these indicators of straightness, the relationship could use some straightening, but clearly, the bulging rule does not apply.

An alternative method for straightening data, especially characterized by nonlinearities, is the GenIQ Model, a machine-learning, genetic-based data mining method. As I extensively cover this model in Chapters 29 and 30, suffice it to say that I use GenIQ to reexpress INCOME. The genetic structure, which represents the reexpressed INCOME variable, labeled gINCOME, is defined in Equation (9.3):

$$\text{gINCOME} = \sin(\sin(\sin(\sin(\text{INCOME})*\text{INCOME}))) + \log(\text{INCOME}) \quad (9.3)$$

The structure uses the nonlinear reexpressions of the trigonometric sine function (four times) and the log (to base 10) function to loosen the "kinky"

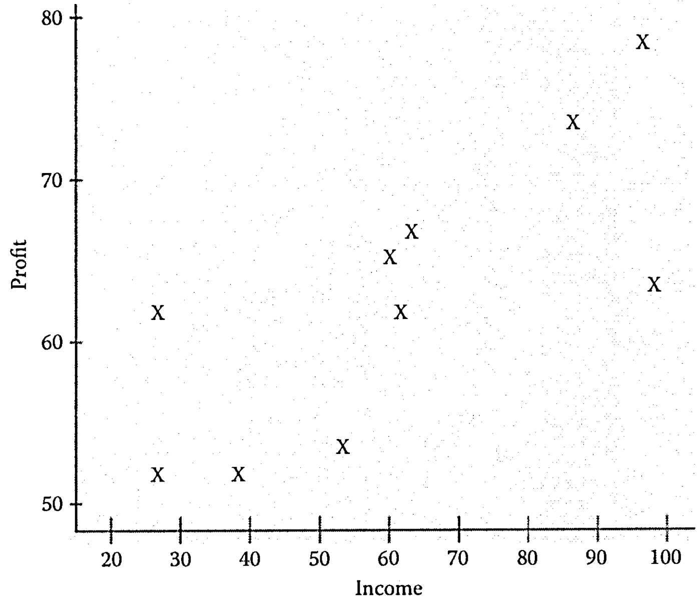

FIGURE 9.2
Plot of PROFIT and INCOME.

PROFIT-INCOME relationship. The relationship between PROFIT and INCOME (via gINCOME) has indeed been smoothed out, as the smooth trace reveals no serious kinks in Figure 9.3. Based on TS equal to 6, which again is almost equal to the cutoff score of 7, I conclude there is an almost noticeable straight-line PROFIT-gINCOME relationship, a nonrandom scatter about an underlying positively sloped straight line. The correlation coefficient for the reexpressed relationship is a reliable $r_{\mathrm{PROFIT,\,gINCOME}}$ of 0.894.

Visually, the effectiveness of the GenIQ procedure in straightening the data is obvious: the sharp peaks and valleys in the original PROFIT smooth plots versus the smooth wave of the reexpressed smooth plot. Quantitatively, the gINCOME-based relationship represents a noticeable improvement of 7.24% (= (0.894 - 0.763)/0.763) increase in correlation coefficient "points" over the INCOME-based relationship.

Two points are noteworthy: Recall that I previously invoked the statistical factoid that states a dollar-unit variable is often reexpressed with the log function. Thus, it is not surprising that the genetically evolved structure gINCOME uses the log function. With respect to logging the PROFIT variable, I concede that PROFIT could not benefit from a log reexpression, no doubt due to the "mini" in the dataset (i.e., the small size of the data), so I chose to work with PROFIT, not log of PROFIT, for the sake of simplicity (another EDA mandate, even for instructional purposes).

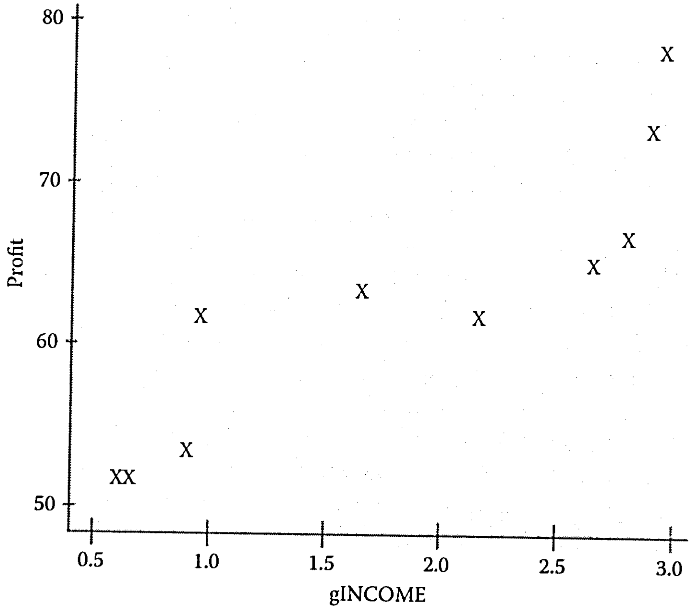

FIGURE 9.3
Plot of PROFIT and gINCOME.

9.3.1.2 Reexpressing AGE

The stormy scatter of the 10-paired (PROFIT, AGE) points in the smooth plot in Figure 9.4 is an exemplary plot of a nonrelationship between two variables. Not surprisingly, the TS value of 3 indicates there is no noticeable PROFIT-AGE relationship. Senselessly, I calculate the correlation coefficient for this nonexistent linear relationship: $r_{PROFIT, AGE}$ equals -0.172, which is clearly not meaningful. Clearly, the bulging rule does not apply.

I use GenIQ to reexpress AGE, labeled gAGE. The genetically based structure is defined in Equation (9.4):

$$gAGE = \sin(\tan(\tan(2*AGE) + \cos(\tan(2*AGE)))) \tag{9.4}$$

The structure uses the nonlinear reexpressions of the trigonometric sine, cosine, and tangent functions to calm the stormy-nonlinear relationship. The relationship between PROFIT and AGE (via gAGE) has indeed been smoothed out, as the smooth trace reveals in Figure 9.5. There is an almost noticeable PROFIT-gAGE relationship with TS = 6, which favorably compares to the original TS of 3. The reexpressed relationship admittedly does not portray an exemplary straight line, but given its stormy origin, I see a beautiful positively sloped ray, not very straight, but trying to shine through.

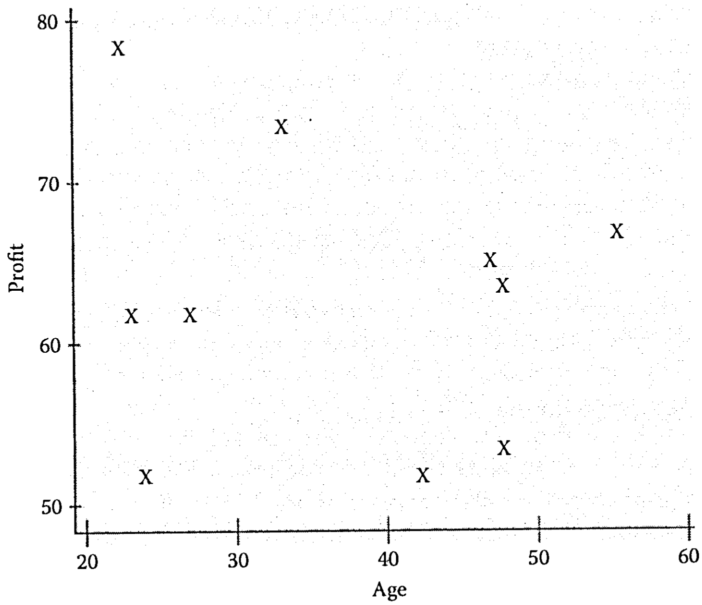

FIGURE 9.4
Plot of PROFIT and AGE.

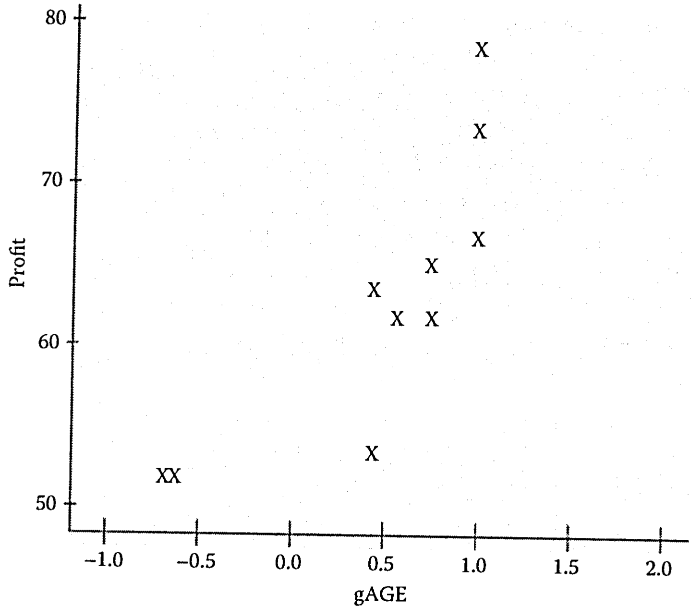

FIGURE 9.5
Plot of PROFIT and gAGE.

I consider the corresponding correlation coefficient $r_{\text{PROFIT, gAGE}}$ value of 0.819 as reliable and remarkable.

Visually, the effectiveness of the GenIQ procedure in straightening the data is obvious: the abrupt spikes in the original smooth plot of PROFIT and AGE versus the rising counterclockwise wave of the second smooth plot of PROFIT and gAGE. With enthusiasm and without quantitative restraint, the gAGE-based relationship represents a noticeable improvement—a whopping 376.2% (= (0.819 - 0.172)/0.172; disregarding the sign) improvement in correlation coefficient points over the AGE-based relationship. Since the original correlation coefficient is meaningless, the improvement percentage is also meaningless.

9.3.2 Plot of Smooth Predicted versus Actual

For a closer look at the detail of the strength (or weakness) of the gINCOME and gAGE structures, I construct the corresponding plots of PROFIT smooth predicted versus actual. The scatter about the 45° lines in the smooth plots for both gINCOME and gAGE in Figures 9.6 and 9.7, respectively, indicate a reasonable level of certainty in the reliability of the structures. In other words, both gINCOME and gAGE should be important variables for predicting PROFIT. The correlations between gINCOME-based predicted and

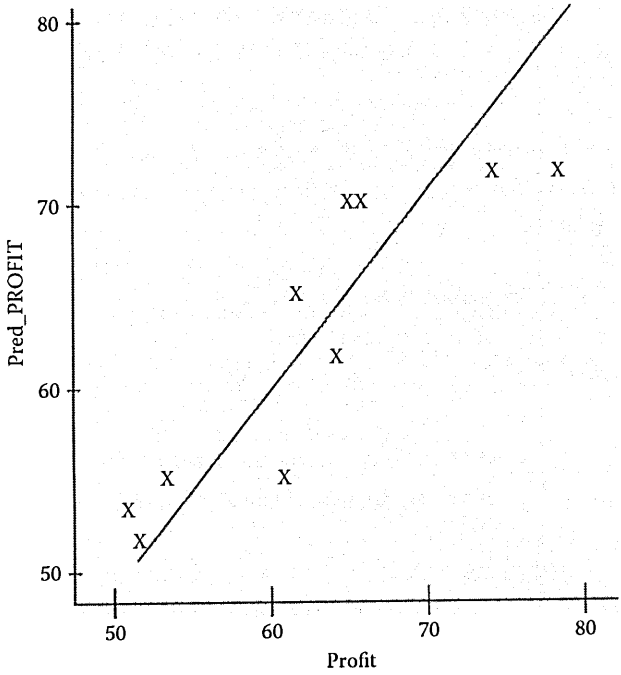

FIGURE 9.6
Smooth PROFIT predicted versus actual based on gINCOME.

actual smooth PROFIT values and between gAGE-based predicted and actual smooth PROFIT values have $r_{sm.PROFIT,\ sm.gINCOME}$ and $r_{sm.PROFIT,\ sm.gAGE}$ values equal to 0.894 and 0.819, respectively. (Why are these r values equal to $r_{PROFIT,\ INCOME}$ and $r_{PROFIT,\ AGE}$, respectively?)

9.3.3 Assessing the Importance of Variables

As in the correlating section of Chapter 8, the classical approach of assessing the statistical significance of a variable for model inclusion is the well-known null hypothesis-significance testing procedure,* which is based on the reduction in prediction error (actual PROFIT minus predicted PROFIT) associated with the variable in question. The only difference between the discussions of the logistic regression in Chapter 8 is the apparatus used. The statistical apparatus of the formal testing procedure for ordinary regression consists of the sum of squares (total; due to regression; due to error), the F statistic, degrees of freedom (df), and the p value. The procedure uses

* "What If There Were No Significance Testing?" (on author's Web site, http://www.geniq.net/res/What-If-There-Were-No-Significance-Testing.html).

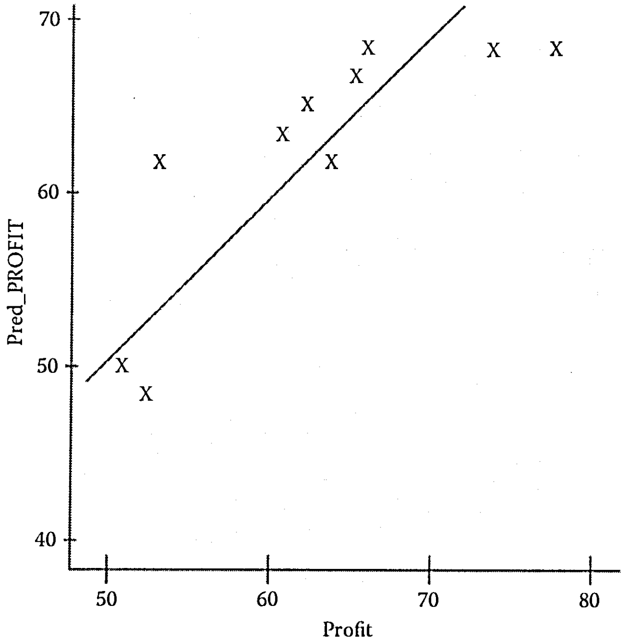

FIGURE 9.7
Smooth PROFIT predicted versus actual based on gAGE.

the apparatus within a theoretical framework with weighty and untenable assumptions, which, from a purist's point of view, can cast doubt on findings of statistical significance. Even if findings of statistical significance are accepted as correct, it may not be of practical importance or have noticeable value to the study at hand. For the data miner with a pragmatist slant, the limitations and lack of scalability of the classical system of variable assessment cannot be overlooked, especially within big data settings. In contrast, the data mining approach uses the F statistic, R-squared, and degrees of freedom in an informal data-guided search for variables that suggest a noticeable reduction in prediction error. Note that the informality of the data mining approach calls for suitable change in terminology, from declaring a result as statistically significant to worthy of notice or noticeably important.

9.3.3.1 Defining the F Statistic and R-Squared

In data mining, the assessment of the importance of a subset of variables for predicting profit involves the notion of a noticeable reduction in prediction error due to the subset of variables. It is based on the F statistic, R-squared, and degrees of freedom, which are always reported in the ordinary regression output. For the sake of reference, I provide their definitions and relationship with each other in Equations (9.5), (9.6), and (9.7).

$$F = \frac{\text{Sum of squares due to regression/df due to regression model}}{\text{Sum of squares due to error/df due to error in regression model}} \tag{9.5}$$

$$\text{R-squared} = \frac{\text{Sum of squares due to regression}}{\text{Total Sum of squares}} \tag{9.6}$$

$$F = \frac{\text{R-squared/number of variables in model}}{(1-\text{R-squared})/(\text{sample size} - \text{number of variables in model} - 1)} \tag{9.7}$$

For the sake of completion, I provide an additional statistic: the adjusted R-squared. R-squared is affected, among other things, by the ratio of the number of predictor variables in the model to the size of the sample. The larger the ratio, the greater the overestimation of R-squared is. Thus, the adjusted R-squared as defined in Equation (9.8) is not particularly useful in big data settings.

Adjusted R-squared =

$$1-(1-\text{R - squared})\frac{(\text{sample size}-1)}{(\text{sample size}-\text{number of variables in model}-1)} \tag{9.8}$$

In the following sections, I detail the decision rules for three scenarios for assessing the importance of variables (i.e., the likelihood the variables have some predictive power). In brief, the larger the F statistic, R-squared, and adjusted R-squared values, the more important the variables are in predicting profit.

9.3.3.2 Importance of a Single Variable

If X is the only variable considered for inclusion into the model, the decision rule for declaring X an important predictor variable in predicting profit is if the F value due to X is greater than the *standard F value 4*, then X is an important predictor variable and should be considered for inclusion in the model. Note that the decision rule only indicates that the variable has some importance, not how much importance. The decision rule implies that a variable with a greater F value has a greater *likelihood of some importance* than a variable with a smaller F value, not that it has greater importance.

9.3.3.3 Importance of a Subset of Variables

When subset A consisting of k variables is the only subset considered for model inclusion, the decision rule for declaring subset A important in predicting profit is as follows: If the average F value per number of variables (the degrees of freedom) in the subset A—F/df or F/k—is greater than standard F value 4, then subset A is an important subset of predictor variable and should be considered for inclusion in the model. As before, the decision rule only indicates that the subset has some importance, not how much importance.

9.3.3.4 Comparing the Importance of Different Subsets of Variables

Let subsets A and B consist of k and p variables, respectively. The number of variables in each subset does not have to be equal. If the number of variables is equal, then all but one variable can be the same in both subsets. Let F(k) and F(p) be the F values corresponding to the models with subsets A and B, respectively.

The decision rule for declaring which of the two subsets is more important (greater likelihood of some predictive power) in predicting profit is

1. If F(k)/k is greater than F(p)/p, then subset A(k) is the more important predictor variable subset; otherwise, B(p) is the more important subset.
2. If F(k)/k and F(p)/p are equal or have comparable values, then both subsets are to be regarded tentatively as of comparable importance. The model builder should consider additional indicators to assist in the decision about which subset is better. It clearly follows from the decision rule that the model defined by the more important subset is the better model. (Of course, the rule assumes that F/k and F/p are greater than the standard F value 4.)

Equivalently, the decision rule either can use R-squared or adjusted R-squared in place of F/df. The R-squared statistic is a friendly concept in that its values serve as an indicator of the percentage of variation explained by the model.

9.4 Important Variables for Mini Case Study

I perform two ordinary regressions, regressing PROFIT on gINCOME and on gAGE; the outputs are in Tables 9.5 and 9.6, respectively. The F values are

TABLE 9.5

OLS Output: PROFIT with gINCOME

Source	df	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	582.1000	582.1000	31.83	0.0005
Error	8	146.3000	18.2875		
Corrected total	9	728.4000			
		Root MSE	4.2764	R-square	0.7991
		Dependent mean	62.6000	Adj R-sq	0.7740
		Coeff var	6.8313		

Variable	df	Parameter Estimate	Standard Error	t Value	Pr > |t|
INTERCEPT	1	47.6432	2.9760	16.01	<.0001
gINCOME	1	8.1972	1.4529	5.64	0.0005

TABLE 9.6

OLS Output: PROFIT with gAGE

Source	df	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	488.4073	488.4073	16.28	0.0038
Error	8	239.9927	29.9991		
Corrected total	9	728.4000			
		Root MSE	5.4771	R-Square	0.6705
		Dependent mean	62.6000	Adj R-sq	0.6293
		Coeff var	8.7494		

Variable	df	Parameter Estimate	Standard Error	t Value	Pr > |t|
INTERCEPT	1	57.2114	2.1871	26.16	< .0001
gAGE	1	11.7116	2.9025	4.03	0.0038

31.83 and 16.28, respectively, which are greater than the standard F value 4. Thus, both gINCOME and gAGE are declared important predictor variables of PROFIT.

9.4.1 Relative Importance of the Variables

Chapter 8 contains the same heading (Section 8.12), with only a minor variation with respect to the statistic used. The *t statistic* as posted in ordinary regression output can serve as an indicator of the relative importance of a variable and for selecting the best subset, which is discussed next.

9.4.2 Selecting the Best Subset

The decision rules for finding the best subset of important variables are nearly the same as those discussed in Chapter 8; refer to Section 8.12.1. Point 1 remains the same for this discussion. However, the second and third points change, as follows:

1. Select an initial subset of important variables.
2. For the variables in the initial subset, generate smooth plots and straighten the variables as required. The most noticeable handfuls of original and reexpressed variables form the starter subset.
3. Perform the preliminary ordinary regression on the starter subset. Delete one or two variables with absolute t-statistic values less than the *t cutoff value* 2 from the model. This results in the first incipient subset of important variables. Note the changes to points 4, 5, and 6 with respect to the topic of this chapter.
4. Perform another ordinary regression on the incipient subset. Delete one or two variables with t values less than the t cutoff value 2 from the model. The data analyst can create an illusion of important variables appearing and disappearing with the deletion of different variables. The remainder of the discussion in Chapter 8 remains the same.
5. Repeat step 4 until all retained predictor variables have comparable t values. This step often results in different subsets as the data analyst deletes judicially different pairings of variables.
6. Declare the best subset by comparing the relative importance of the different subsets using the decision rule in Section 9.3.3.4.

9.5 Best Subset of Variables for Case Study

I build a preliminary model by regressing PROFIT on gINCOME and gAGE; the output is in Table 9.7. The two-variable subset has an F/df value of 7.725 (= 15.45/2), which is greater than the standard F value 4. But, the t value for gAGE is 0.78 less than the t cutoff value (see bottom section of Table 9.7). If I follow step 4, then I would have to delete gAGE, yielding a simple regression model with the lowly, albeit straight, predictor variable gINCOME. By the way, the adjusted R-squared is 0.7625 (after all, the entire minisample is not big).

Before I dismiss the two-variable (gINCOME, gAGE) model, I construct the smooth residual plot in Figure 9.8 to determine the quality of the predictions

TABLE 9.7
OLS Output: PROFIT with gINCOME and gAGE

Source	df	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	593.8646	296.9323	15.45	0.0027
Error	7	134.5354	19.2193		
Corrected total	9	728.4000			
		Root MSE	4.3840	R-squared	0.8153
		Dependent mean	62.6000	Adj R-sq	0.7625
		Coeff var	7.0032		

Variable	df	Parameter Estimate	Standard Error	t Value	Pr > |t|
INTERCEPT	1	49.3807	3.7736	13.09	< .0001
gINCOME	1	6.4037	2.7338	2.34	0.0517
gAGE	1	3.3361	4.2640	0.78	0.4596

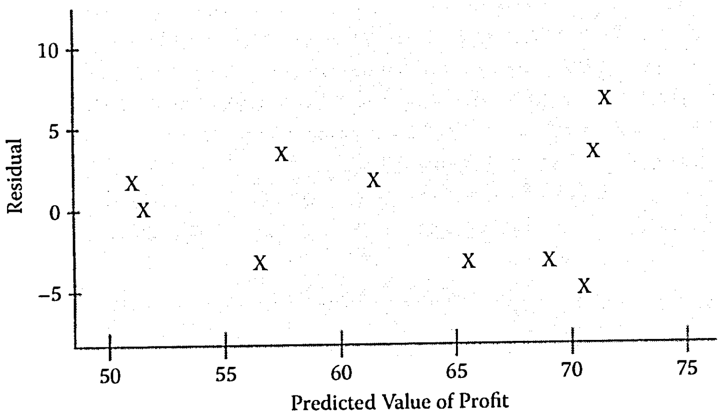

FIGURE 9.8
Smooth residual plot for (gINCOME, gAGE) model.

of the model. The smooth residual plot is declared to be equivalent to the null plot-based general association test (TS = 5). Thus, the overall quality of the predictions is considered good. That is, on average, the predicted PROFIT is equal to the actual PROFIT. Regarding the descriptive statistics for the smooth residual plot, for the smooth residual, the minimum and maximum values and range are −4.567, 6.508, and 11.075, respectively; the standard deviation of the smooth residuals is 3.866.

9.5.1 PROFIT Model with gINCOME and AGE

With a vigilance in explaining the unexpected, I suspect the reason for the relative nonimportance of gAGE (i.e., gAGE is not important in the presence of gINCOME) is the strong correlation of gAGE with gINCOME: $r_{gINCOME, gAGE}$ = 0.839. This supports my contention but does not confirm it.

In view of the foregoing, I build another two-variable model regressing PROFIT on gINCOME and AGE; the output is in Table 9.8. The (gINCOME, AGE) subset has an F/df value of 12.08 (= 24.15/2), which is greater than the standard F value 4. Statistic happy, I see the t values for both variables are greater than the t cutoff value 2. I cannot overlook the fact that the raw variable AGE, which by itself is not important, now has relative importance in the presence of gINCOME. (More about this "phenomenon" at the end of the chapter.) Thus, the evidence is that the subset of gINCOME and AGE is better than the original (gINCOME, gAGE) subset. By the way, the adjusted R-squared is 0.8373, representing a 9.81% (= (0.8373 - 0.7625)/0.7625) improvement in "adjusted R-squared" points over the original-variable adjusted R-squared.

The smooth residual plot for the (gINCOME, AGE) model in Figure 9.9 is declared to be equivalent to the null plot-based general association test with TS = 4. Thus, there is indication that the overall quality of the predictions is good. Regarding the descriptive statistics for the two-variable smooth residual plot, for the smooth residual, the minimum and maximum values and range are −5.527, 4.915, and 10.442, respectively, and the standard deviation of the smooth residual is 3.200.

To obtain further indication of the quality of the predictions of the model, I construct the plot of the smooth actual versus predicted for the (gINCOME,

TABLE 9.8

OLS Output: PROFIT with gINCOME and AGE

Source	df	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	636.2031	318.1016	24.15	0.0007
Error	7	92.1969	13.1710		
Corrected total	9	728.4000			
		Root MSE	3.6291	R-squared	0.8734
		Dependent mean	62.6000	Adj R-sq	0.8373
		Coeff var	5.79742		

Variable	df	Parameter Estimate	Standard Error	t Value	Pr > |t|
INTERCEPT	1	54.4422	4.1991	12.97	< .0001
gINCOME	1	8.4756	1.2407	6.83	0.0002
AGE	1	-0.1980	0.0977	-2.03	0.0823

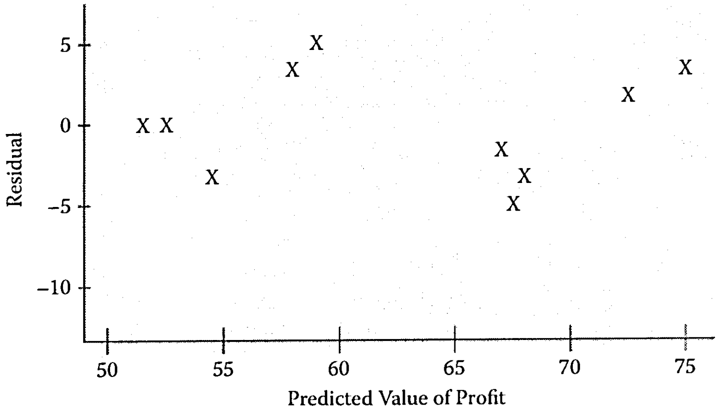

FIGURE 9.9
Smooth residual plot for (gINCOME, AGE) model.

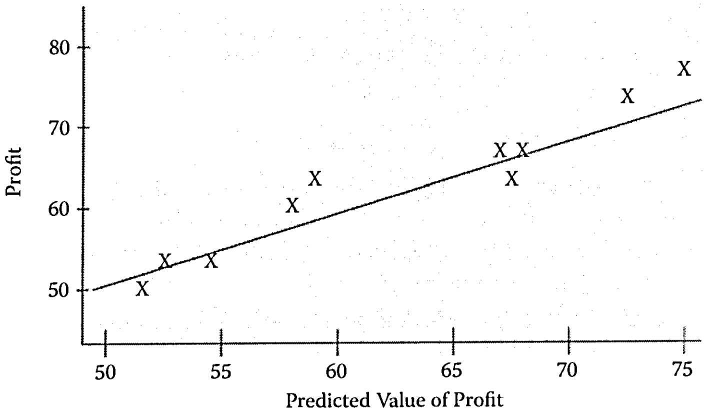

FIGURE 9.10
Smooth actual versus predicted plot for (gINCOME, AGE) model.

AGE) model in Figure 9.10. The smooth plot is acceptable with minimal scatter of the 10 smooth points about the 45 line. The correlation between smooth actual versus predicted PROFIT based on gINCOME and AGE has an $r_{sm.\,gINCOME,\,sm.AGE}$ value of 0.93. (This value is the square root of R-squared for the model. Why?)

For a point of comparison, I construct the plot of smooth actual versus predicted for the (gINCOME, gAGE) model in Figure 9.11. The smooth plot is acceptable with minimal scatter of the 10 smooth points about the 45° line, with a noted exception of some wild scatter for PROFIT values greater than $65. The correlation between smooth actual versus predicted PROFIT based

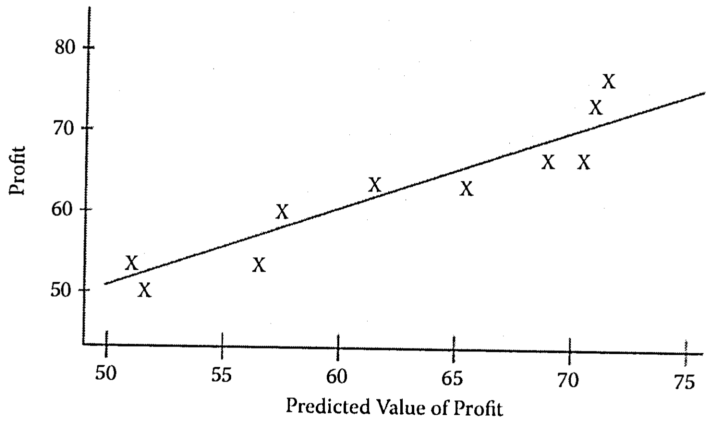

FIGURE 9.11
Smooth actual versus predicted plot for (gINCOME, gAGE) model.

on gINCOME and gAGE has an $r_{\text{sm.gINCOME, sm.gAGE}}$ value of 0.90. (This value is the square root of R-squared for the model. Why?)

9.5.2 Best PROFIT Model

To decide which of the two PROFIT models is better, I put the vital statistics of the preceding analyses in Table 9.9. Based on the consensus of a committee of one, I prefer the gINCOME-AGE model to the gINCOME-gAGE model because there are noticeable indications that the predictions of the former model are better. The former model offers a 9.8% increase in the adjusted R-squared (less bias), a 17.2% decrease in the smooth residual standard deviation (more stable), and a 5.7% decrease in the smooth residual range (more stable).

9.6 Suppressor Variable AGE

A variable whose behavior is like that of AGE—poorly correlated with the dependent variable Y, but becomes important by its inclusion in a model for predicting Y—is known as a *suppressor variable.* [1, 2] The consequence of a suppressor variable is that it increases the R-squared of the model.

I explain the behavior of the suppressor variable within the context of the mini case study. The presence of AGE in the model removes or suppresses the information (variance) in gINCOME that is not related to the variance in PROFIT; that is, AGE suppresses the unreliable noise in gINCOME. This

TABLE 9.9

Comparison of Vital Statistics for Two PROFIT Models
Smooth Residual

Model	Predictor Variables	F Value	t Value	Range	StdDev	Adjusted R-Square
First	gINCOME, gAGE	Greater than cutoff value	Greater than cutoff value for only gINCOME	11.075	3.866	0.7625
Second	gINCOME, AGE	Greater than cutoff value	Greater than cutoff value for both variablest	10.442	3.200	0.8373
Indication	Improvement of 2nd model over 1st model	NA	Because t value for gAGE is less than t cutoff value, gAGE contributes "noise" in model, as evidenced in range, stdDev and adjusted R-square	-5.7%	-17.2%	9.8%

renders the AGE-adjusted variance in gINCOME more reliable or potent for predicting PROFIT.

I analyze the paired correlations among the three variables to clarify exactly what AGE is doing. Recall that squaring the correlation coefficient represents the "shared variance" between the two variables under consideration. The paired bits of information are in Table 9.10. I know from the prior analysis that PROFIT and AGE have no noticeable relationship; their shared variance of 3% confirms this. I also know that PROFIT and gINCOME do have a noticeable relationship; their shared variance of 80% confirms this as well. In the presence of AGE, the relationship between PROFIT and gINCOME, specifically, the relationship between PROFIT and gINCOME adjusted for AGE has a shared variance of 87%. This represents an improvement of 8.75% (= (0.87 - 0.80/0.80) in shared variance. This "new" variance is now available for predicting PROFIT, increasing the R-squared (from 79.91% to 87.34%).

It is a pleasant surprise in several ways that AGE turns out to be a suppressor variable. First, suppressor variables occur most often in big data settings, not often with small data, and are truly unexpected with minidata. Second, the suppressor variable scenario serves as object lessons for the EDA

TABLE 9.10

Comparison of Pairwise Correlations among PROFIT, AGE, and gINCOME

Correlation Pair	Correlation Coefficient	Shared Variance
PROFIT and AGE	-0.172	3%
PROFIT and gINCOME	0.894	80%
PROFIT and AGE in the presence of gINCOME	0.608	37%
PROFIT and gINCOME in the presence of AGE	0.933	87%

paradigm: Dig, dig, dig into the data, and you will find gold or some reward for your effort. Third, the suppressor variable scenario is a small reminder of a big issue: The model builder must not rely solely on predictor variables that are highly correlated with the dependent variable, but also must consider the poorly correlated predictor variables as they are a great source of latent predictive importance.

9.7 Summary

The ordinary regression model is presented as the workhorse of profit modeling as it has been in steady use for almost 200 years. As such, I illustrated in an orderly and detailed way the essentials of ordinary regression. Moreover, I showed the enduring usefulness of this popular analysis and modeling technique as it works well within the EDA/data mining paradigm of today.

I first illustrated the rudiments of the ordinary regression model by discussing the SAS program for building and scoring an ordinary regression model. The program is a welcome addition to the tool kit of techniques used by model builders working on predicting a quantitative dependent variable.

Then, I discussed ordinary regression modeling with minidata. I used this extremely small dataset not only to make the discussion of the data mining techniques tractable but also to emphasize two aspects of data mining. First, data mining techniques of great service should work as well with big data as with small data. Second, every fruitful effort of data mining on small data is evidence that big data are not always necessary to uncover structure in the data. This evidence is in keeping with the EDA philosophy that the data miner should work from simplicity until indicators emerge to go further: If predictions are not acceptable, then increase data size. The data mining techniques discussed are those introduced in the logistic regression framework of Chapter 8 and carry over with minor modification to ordinary regression.

Before proceeding with the analysis of the mini case study, I supplemented the bulging rule when analysis involves ordinary regression. Unlike in logistic regression, for which the logit dependent variable cannot be reexpressed, in ordinary regression the quantitative dependent variable can be reexpressed. The bulging rule as introduced within the logistic regression framework in Chapter 8 can put several reexpressions of the quantitative dependent variable up for consideration. For such cases, I provided additional guidelines to the bulging rule, which will prove valuable:

1. If there is a small range of dependent-variable powers, then the best reexpressed dependent variable is the one with the noticeably largest correlation coefficient.
2. If there is a small range of dependent-variable powers and the correlation coefficient values are comparable, then the best reexpressed dependent variable is defined by the average power among the dependent-variable powers.
3. When there is a large range of dependent-variable powers, choose several handfuls of reexpressed predictor variables, which have the largest correlation coefficients with the unaltered quantitative dependent variable. Then, proceed as usual, invoking the bulging rule for exploring the best reexpressions of the dependent variable and the predictor variables. If the dependent variable is reexpressed, then apply steps 1 or 2.

With the minidataset selected for regressing PROFIT on INCOME and AGE, I introduced the alternative GenIQ Model, a machine-learning, genetic-based data mining method for straightening data. I illustrated the data mining procedure of smoothing and assessing the smooth with the general association test and determined that both predictor variables need straightening, but the bulging rule does not apply. The GenIQ Model evolved reasonably straight relationships between PROFIT and each reexpressed predictor variable, gINCOME and gAGE, respectively. I generated the plots of PROFIT smooth predictive versus actual, which provided further indication that gINCOME and gAGE should be important variables for predicting PROFIT.

Continuing with the mini case study, I demonstrated an ordinary regression-specific data mining alternative approach to the classical method of assessing. (The data mining techniques discussed are those introduced in the logistic regression framework of Chapter 8 and carry over with minor modification to ordinary regression.) This alternative approach involves the importance of individual predictor variables, as well as the importance of a subset of predictor variables and the relative importance of individual predictor variables, in addition to the goodness of model predictions. Additional methods that set out specific to ordinary regression included selecting the best subset of predictor variables and comparing the importance between two subsets of predictor variables.

Within my illustration of the case study is a pleasant surprise—the existence of a suppressor variable (AGE). A variable whose behavior is poorly correlated with the dependent variable but becomes important by its inclusion in a model for predicting the dependent variable is known as a suppressor variable. The consequence of a suppressor variable is that it increases the R-squared of the model. A suppressor variable occurs most often in big data settings, not often with small data, and is truly unexpected with minidata. The suppressor variable scenario served as an object lesson for the EDA paradigm: Dig deeply into the data, and you will find a reward for your effort. And, the suppressor variable scenario is a small reminder of a bigger issue: The model builder must not rely solely on predictor variables that are highly correlated with the dependent variable but also should consider the poorly correlated predictor variables as they are a great source of latent predictive importance.

References

1. Horst, P., The role of predictor variables which are independent of the criterion, *Social Science Research Bulletin*, 48, 431–436, 1941.
2. Conger, A.J., A revised definition for suppressor variables: A guide to their identification and interpretation, *Educational and Psychological Measurement*, 34, 35–46, 1974.