Reviews of Books and Teaching Materials

Applied Regression Analysis: A Second Course in Business and Economic Statistics (4th ed.).

Terry E. DIELMAN. Stamford, CT: Thomson Learning, 2005, xii+424 pp., \$137.95(H+CD + InfoTrac), ISBN: 0-534-46548-X.

The preface of this book (hereafter ARA) states that it can be used either for a one-semester course in regression for undergraduates or MBAs, or as a text in a second-semester statistics course. The regression part of the book is composed of eight chapters consisting of 334 pages. For the second stated purpose, short chapters on analysis of variance (ANOVA) and forecasting are included, with a very sparse chapter on discriminant analysis and logistic regression sandwiched in between; there is no quality control material. A "Using the Computer" section at the end of each chapter describes relevant capabilities of Excel, MINITAB, and SAS. Paraphrasing the author, ARA is directed at applied researchers or consumers of statistics, and the emphasis is on statistical analysis as a multistage process rather than on computational issues or theory (the matrix approach to regression is discussed only in a short appendix at the end of the book).

Chapter 1, "Introduction to Regression Analysis," occupies just three pages of text with nary a symbol, equation, or any data. I did appreciate the one-paragraph warning on the perils of using Excel. In contrast to some other regression texts (e.g., the recent Abraham and Ledolter 2006), ARA includes a chapter reviewing basic statistical concepts. Coverage here is necessarily selective and cursory. I personally would have liked to see boxplots and comparative boxplots to support their use in later chapters (sigh, no boxplots in the book). A page plus is devoted to p values, not enough in my judgment given their subsequent prominence. Here the author writes "That is, it is the probability of observing a t value or z value as extreme as, or more extreme than, the sample test statistic." No mention is made of the probability being calculated assuming that H_0 is true (a correct statement appears in the next chapter) or that the p value is a valid concept for F, χ^2, and various other tests and generally involves a tail-area under an appropriate reference curve.

The real story on regression begins in Chapter 3 with coverage of the usual topics in simple linear regression. Surprisingly, descriptive correlations or inference about the population correlation coefficient are absent; the only mention of r is as the appropriately signed square root of the coefficient of determination. On page 76, the author writes "To allow statistical inference from a sample to the population, some assumptions about the population regression line are necessary." But the subsequent assumptions are about the random disturbances in the model equation rather than the "line" itself. Section 3.3, "Inferences from Simple Linear Regression," rattles on for 10 pages of exposition covering various procedures before an example appears, and that example involves an artificial dataset consisting of just six observations with no context. The F test for a useful linear relationship is introduced without any preliminary discussion of F distributions, although a first statistics course typically does not include exposure to this topic. The exposition is generally technically correct, but I would have liked to see rationale for various procedures. For example, a confidence interval for the slope is given without first stating a general distributional result involving standardizing the estimated slope to obtain a t variable. Maybe this strategy would have prevented the author from giving an expression for the variance of a prediction which is actually the variance of a prediction error. I did like the subsection on assessing the quality of prediction in which the potentially deceptive nature of R^2 is revealed.

Chapter 4 covers the model and inferential procedures for multiple linear regression. The principle of least squares is introduced before any data or notation for data is given. Inferences about individual regression coefficients precede the coefficient of multiple determination and F test for model utility. I liked the author's emphasis on a correct interpretation of the null hypothesis $H_0 : \beta_k = 0$ and also his explanation of adjusted R^2. Section 4.6 discusses multicollinearity. Absence of the correlation coefficient in the previous chapter makes it difficult to understand the author's suggestion that pairwise correlation coefficients be examined. Variance inflation factors are briefly introduced along with mention of regressing each predictor against all the others. This section is devoid of examples.

Curve-fitting is briefly considered in Chapter 5 (curiously, a regression of y against $\ln(x)$ is considered but not the usual exponential regression model involving $\ln(y)$ versus x). A commendable characteristic of the book is the expansive treatment of assessing model assumptions in Chapter 6. I especially appreciated inclusion of suggested remedies for violations. There is plenty of output from software, including normal probability plots (unfortunately the author explains these in terms of normal scores whereas the included plots actually use an appropriate nonlinear probability scale on the vertical axis, and residuals rather than standardized residuals are plotted). The Ryan-Joiner test for normality is mentioned, but without a discussion of correlation no meaningful intuition can be conveyed. Section 6.7 includes brief mention of Cook's distances and deviation of fits (DFITS).

The last two chapters on regression cover the use of indicator and interaction predictors and the issue of variable selection. The first of these chapters does not include anything about the complete second-order model or graphs showing contours of regression functions with and without interaction and curvature. The explanation and use of C_p in variable selection is good. Chapter 8, unfortunately, has only five exercises.

The ANOVA chapter is comparable to what is usually included in the second half of a general business statistics book such as McClave, Benson, and Sincich (2008). Ditto for the last chapter on forecasting. Regression models for time series data are actually considered in Chapters 3 and 4 rather than in the forecasting chapter. The distinction between causal and extrapolative models is useful. But I take issue with the much earlier statement that "Most of the techniques discussed in this and subsequent chapters can be applied to either time-series or cross-sectional data." In Section 3.6, a linear trend is fit to quarterly data and the usual t ratio is used without adjustment for autocorrelation to test for a linear effect. These data are reconsidered in the chapter on assessing model assumptions, where it is asserted that no model violations are apparent. Unfortunately, examination of the sample autocorrelation coefficients of the standardized residuals shows a significant effect at lag four, something not detected by the Durbin-Watson test. Many regression exercises include time series data without scrutinizing autocorrelation (a sin also committed by various business statistics texts). In Chapter 11, forecasts via moving averages are briefly considered (but how does one forecast more than one period ahead?), followed by several variants of exponential smoothing. Only a brief allusion is made to the issue of selecting the smoothing parameter(s).

ARA is a reasonable book with a decent presentation of basic methodology and some nice examples and exercises. However, the bigger business statistics books cover most of this material, and using such a book for two semesters is obviously more cost effective than switching to a second book. As far as a stand alone regression course is concerned, my tastes tend toward books with a bit more attention to motivation and underlying theory such as the aforementioned Abraham and Ledolter (2006) or the classic Kutner, Nachstein, Neter, and Li (2004). These, however, are more ambitious and sophisticated than ARA, so the latter may be a sensible, conservative choice.

Jay DEVORE
California Polytechnic State University

REFERENCES

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

Kutner, M. H., Nachstein, C. J., Neter, J., and Li, W. (2004), *Applied Linear Statistical Models* (5th ed.), New York: McGraw-Hill/Irwin.

McClave, J., Benson, G., and Sincich, T. (2008), *Statistics for Business and Economics* (10th ed.), Upper Saddle River, NJ: Pearson Education.

Applied Regression Modeling: A Business Approach.

Iain PARDOE. Hoboken, NJ: Wiley, 2006, xx + 303 pp., \$105.00 (H), ISBN: 0-471-97033-6.

Pardoe's book is intended for undergraduate students for a course that follows an introductory probability and statistics course. The author claims that it would also be useful for nonstatistics major graduate students, including MBAs.

©2007 American Statistical Association DOI: 10.1198/000313007X243115

By design, no calculus is used and even algebra is kept to a minimum. Except for one isolated use on page 233, there are no matrices or vectors.

The attention to detail is impressive. The book is very well written and the author is extremely careful with his descriptions. My approaches to some of the examples disagree with the author's, however.

The examples are wonderful. Unfortunately, every problem is reduced to the same variable names: Y for the dependent variable; X_1, X_2, X_3, and so on for quantitative independent variables; D_1, D_2, D_3, and so on for dummy variables. We get to see models in formats like this on page 205:

$$\begin{aligned} \mathrm{E}(Y) \quad = \quad & b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 \\ & + b_5 X_3 X_4 + b_6 X_5 + b_7 X_5^2 + b_8 X_6 \\ & + b_9 D_7 + b_{10} D_8 + b_{11} D_9 + b_{12} D_{10} \\ & + b_{13} D_{11} + b_{14} D_{12}. \end{aligned}$$

The variable definitions are supplied, so the reader will be able to assemble the meanings correctly. The files that I checked on the book's Web site do not identify the original variable names; these are given as $Y, X_1, X_2, X_3, \ldots, D_1, D_2, D_3, \ldots$ as well.

A number of the illustrative examples use small sample sizes, and they work well for the intended purposes. Some of the larger sample size examples involve real data, but lack credibility after probing. The dataset on credit cards (p. 189), with a sample size of $n = 50$, regresses unpaid monthly credit card balance (Y) on (average monthly purchases, average monthly housing expense, renter [Y/N], gender [M/F]). The sample variances of the Y_i's is 58.42. The model developed next in the text (p. 190), gets the regression standard error down to 5.80. This strikes me as much too good (low) to believe for data of this variety. Moreover, the data are completely separated on the renter variable; specifically $\max(Y \mid \text{(renter)} = N) < \min(Y \mid \text{(renter)} = Y)$.

There are seven chapters: (1) foundations, (2) simple linear regression, (3) multiple linear regression, (4) model building I, (5) model building II, (6) case studies, and (7) extensions.

The foundations chapter covers random sampling, normal and t distributions, central limit theorem, interval estimation, and hypothesis testing. The notation choices are for the most part standard, but there are a few minor annoyances. For instance, m_Y is used for the sample mean (p. 11) where most people would use μ_Y. Page 15 has the statement $\Pr(\ \mathrm{E}(Y) > \ 258.492\ ) = 0.975$, which puts a probability on a statement without a random variable. Also, the author uses NH and AH (used first on p. 17) whereas the standard is H_0 and H_1.

The chapters on simple linear regression and multiple linear regression are standard and contain no surprises. The discussion of tests with nested models was most welcome.

Regression output summaries are given in two forms, "model summary" and "parameters." The "model summary" lists R, R^2, R^2_{adj} , and the regression standard error. The "parameters" section gives the coefficient estimates, standard errors of coefficient estimates, the t statistics, and the p values associated with the t statistics. These summaries do not provide the sample size, the F statistic, the p value associated with F, or the degrees of freedom accounting. Analysis of variance (ANOVA) tables would have helped. ANOVA is discussed in the multiple linear regression chapter but is not seen in the remainder of the book, except for exercise 4.6.

The first chapter on model building covers data transformations, interactions, and qualitative predictors. Predictor effect plots are introduced in this chapter. This topic is not commonly seen in regression texts, and Pardoe's inclusion of these plots is welcome. In the linear regression of Y on X, the predictor effect of X is simply the estimated coefficient. In a model of the form $\mathrm{E}(Y) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2$, the predictor effect of X_1 varies according to X_2, and Pardoe gives the plot of this predictor effect as a function of X_2. This plot is insightful for models that are quadratic in X_1 or in models with predictor interactions, including interactions with dummy variables. These plots are entertaining, but they are not automatic in any four of the computer environments (though hint 31 in Appendix A shows step-by-step details of their construction).

The second chapter on model building deals with outliers, collinearity, and variable selection; there are additional comments on overfitting, extrapolation, and missing data. The author does not use the model selection C_p statistic, nor does he use stepwise regression or best subsets regression. Variable selection is made through the p values associated with the individual t statistics. This chapter also covers autocorrelation, but the Durbin–Watson test is simply referenced out (p. 175). The example on autocorrelation uses the lagged Y as a predictor, but the dataset is confused by unequal time spacings (pp. 173–175).

The case studies chapter shows the methodical steps leading from data to final analysis. There were two instances in which additional discussion would have been appreciated. The first case, on home prices, involved a qualitative variable with six levels. One level was selected as the reference category, and five dummy variables were created for the regression. The p values for three of these five dummies were above 0.05, and these dummies were removed from the model. This is a controversial action; the variable selection problem is not invariant over the choices for the reference category.

The second case study dealt with model year 2004 automobiles and had dependent variable miles/gallon with quantitative independent variables weight, horsepower, engine size, cylinders, and wheelbase. The author finds similar-looking curved relationships between the dependent variable and each of the independent variables. The analysis then takes the reciprocal of all five of these independent variables. This comes at an enormous cost in interpretability. Why not just take the reciprocal of the miles/gallon dependent variable, which at least has an obvious interpretation?

The final chapter on extensions includes, among other things, logistic regression, and Bayesian inference.

There are many things to like about this book.

1. Computer hints are provided in Appendix A for four common environments: SPSS, Minitab, SAS, and R/S-PLUS. The hints have the same sequential numbers in each of the four subsections of this appendix. For instance, hint 11 in each subsection gives the method for creating boxplots.

2. Interpretations are carefully crafted. Consider, for example, this explanation on page 81: "The regression modeling described in this book can really only be used to quantify relationship and to identify whether a change in one variable is associated with a change in another variable, not to establish whether changing one variable 'causes' another to change." This careful attention to detail is done faithfully through the entire book.

3. The book is written with clerical decency. Cross-referencing is always done through page numbers, rather than through devices such as "see figure 4" or "compare to the first residual plot of the previous chapter." Each problem set is preceded by a short reminder that solutions to even-numbered problems are in an appendix. The book is free of typographical errors. The index lists all the datasets. The end material includes a list of formulas and glossary of terms.

4. All the datasets are available from the Web in ready-to-use form at address *http://lcb1.uoregon.edu/ipardoe/armaba/*. The preface page xiv suggests that these can be found at *www.wiley.com*. My search of this site, along with a parallel Google search, failed to find these. Fortunately, a kind soul at Wiley's Hoboken office was able to direct me to the correct location.

5. The exercises are detailed and useful.

Pardoe's book is a reminder that the practice of regression still contains topics on which practitioners will disagree. The book could be used as indicated for a second course in statistics, but the instructor would almost certainly need to add in discussion on other regression approaches to the problems of nonlinearity, expanding residuals, collinearity, variable selection, and outlier detection.

Gary Simon
New York University

Introduction to Mixed Modelling: Beyond Regression and Analysis of Variance.

N. W. Galwey. Chichester, West Sussex, U.K.: Wiley, 2006, x+366 pp. \$100.00 (H); \$90.00 (online). ISBN: (H) 0-470-01496-2; ISBN: (online) 0-470-03598-6.

When asked if I would review this recently published introductory text on mixed modeling, I accepted with pleasure, and with a sense of curiosity to see how the author tackled this rather challenging, broad subject. The term "mixed modeling" is used here to refer to linear models that have both fixed and random effects. The preface provides an excellent summary of the basic concepts of a mixed model, and how these models form a natural extension of regression and analysis of variance techniques. The introductory chapter then motivates the necessity (at times) of including an additional random effect term beyond the usual error term through the use of a regression example. One of the main strengths of the text is the bridge it provides between traditional analysis of variance (ANOVA) and regression models and the more recently developed class of mixed models.

The text covers a broad array of mixed models and associated topics over ten chapters in a very thorough and logical manner, ranging from the more familiar mixed effect ANOVA models (e.g., split plot models) to the inclusion of random effects in regression models. The final chapters touch upon more