Preface

Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response variable Y given the $p \times 1$ vector of predictors $\boldsymbol{x}$. In a **linear regression model**, $Y = \boldsymbol{\beta}^T\boldsymbol{x} + e$, and Y is conditionally independent of $\boldsymbol{x}$ given a single linear combination $\boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T\boldsymbol{x}.$$

Multiple linear regression and many experimental design models are special cases of the linear regression model, and the models can be presented compactly by defining the population model in terms of the sufficient predictor $SP = \boldsymbol{\beta}^T\boldsymbol{x}$ and the estimated model in terms of the estimated sufficient predictor $\mathbf{ESP} = \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$. In particular, the **response plot** or estimated sufficient summary plot of the ESP versus Y is used to visualize the conditional distribution $Y|\boldsymbol{\beta}^T\boldsymbol{x}$. The residual plot of the ESP versus the residuals is used to visualize the conditional distribution of the residuals given the ESP.

The literature on multiple linear regression is enormous. See Stigler (1986) and Harter (1974a,b, 1975a,b,c, 1976) for history. Draper (2002) is a good source for more recent literature. Some texts that were "standard" at one time include Wright (1884), Johnson (1892), Bartlett (1900), Merriman (1907), Weld (1916), Leland (1921), Ezekiel (1930), Bennett and Franklin (1954), Ezekiel and Fox (1959), and Brownlee (1965). Recent reprints of several of these texts are available from www.amazon.com.

Draper and Smith (1966) was a breakthrough because it popularized the use of residual plots, making the earlier texts obsolete. Excellent texts include Chatterjee and Hadi (2012), Draper and Smith (1998), Fox (2015), Hamilton (1992), Kutner et al. (2005), Montgomery et al. (2012), Mosteller and Tukey (1977), Ryan (2009), Sheather (2009), and Weisberg (2014). Cook and Weisberg (1999a) was a breakthrough because of its use of response plots.

Other texts of interest include Abraham and Ledolter (2006), Harrell (2015), Pardoe (2012), Mickey et al. (2004), Cohen et al. (2003), Kleinbaum et al. (2014), Mendenhall and Sincich (2011), Vittinghoff et al. (2012), and Berk (2003).

This text is an introduction to linear regression models for undergraduates and beginning graduate students in a mathematics or statistics department. The text is for graduate students in fields like quantitative psychology. The prerequisites for this text are linear algebra and a calculus-based course in statistics at the level of Chihara and Hesterberg (2011), Hogg et al. (2014), Rice (2006), or Wackerly et al. (2008). The student should be familiar with vectors, matrices, confidence intervals, expectation, variance, normal distribution, and hypothesis testing.

This text will not be easy reading for nonmathematical students. Lindsey (2004) and Bowerman and O'Connell (2000) attempt to present regression models to students who have not had calculus or linear algebra. Also see Kachigan (1991, ch. 3–5) and Allison (1999).

This text does not give much history of regression, but it should be noted that many of the most important ideas in statistics are due to Fisher, Neyman, E.S. Pearson, and K. Pearson. See Lehmann (2011). For example, David (2006–2007) says that the following terms were due to Fisher: analysis of variance, confounding, consistency, covariance, degrees of freedom, efficiency, factorial design, information, information matrix, interaction, level of significance, likelihood, location, maximum likelihood, null hypothesis, pivotal quantity, randomization, randomized blocks, sampling distribution, scale, statistic, Student's t, test of significance, and variance.

David (2006–2007) says that terms due to Neyman and E.S. Pearson include alternative hypothesis, composite hypothesis, likelihood ratio, power, power function, simple hypothesis, size of critical region, test criterion, test of hypotheses, and type I and type II errors. Neyman also coined the term confidence interval. David (2006–2007) says that terms due to K. Pearson include bivariate normal, goodness of fit, multiple regression, nonlinear regression, random sampling, skewness, standard deviation, and weighted least squares.

This text is different from the massive competing literature in several ways. First, response plots are heavily used in this text. With the response plot, the presentation for multiple linear regression is about the same as the presentation for simple linear regression. Hence the text immediately starts with the multiple linear regression model, rather than spending 100 pages on simple linear regression and then covering multiple regression.

Second, the assumption of iid normal $N(0, \sigma^2)$ errors is replaced by the assumption that the iid zero mean errors have constant variance σ^2. Then large sample theory can be used to justify hypothesis tests, confidence intervals, and prediction intervals.

Third, the *multivariate linear model* $\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, \ldots, n$ has $m \geq 2$ response variables $Y_1, \ldots, Y_m$ and p predictor variables $x_1, x_2, \ldots, x_p$.

Multivariate linear regression and MANOVA models are special cases. Recent results from Kakizawa (2009), Su and Cook (2012), Olive et al. (2015), and Olive (2016b) make the multivariate linear regression model (Chapter 12) easy to learn after the student has mastered the multiple linear regression model (Chapters 2 and 3). For the multivariate linear regression model, it is assumed that the iid zero mean error vectors have fourth moments.

Fourth, recent literature on plots for goodness and lack of fit, bootstrapping, outlier detection, response transformations, prediction intervals, prediction regions, and variable selection has been incorporated into the text. See Olive (2004b, 2007, 2013a,b, 2016a,b,c) and Olive and Hawkins (2005).

Chapter 1 reviews the material to be covered in the text and can be skimmed and then referred to as needed. Chapters 2 and 3 cover multiple linear regression, Chapter 4 considers generalized least squares, and Chapters 5 through 9 consider experimental design models. Chapters 10 and 11 cover linear model theory and the multivariate normal distribution. These chapters are needed for the multivariate linear regression model covered in Chapter 12. Chapter 13 covers generalized linear models (GLMs) and generalized additive models (GAMs).

The text also uses recent literature to provide answers to the following important questions:

How can the conditional distribution $Y|\boldsymbol{\beta}^T\boldsymbol{x}$ be visualized?

How can $\boldsymbol{\beta}$ be estimated?

How can variable selection be performed efficiently?

How can Y be predicted?

The text emphasizes prediction and visualizing the models. Some of the applications in this text using this research are listed below.

1) It is shown how to use the response plot to detect outliers and to assess the adequacy of linear models for multiple linear regression and experimental design.

2) A graphical method for selecting a response transformation for linear models is given. Linear models include multiple linear regression and many experimental design models. This method is also useful for multivariate linear regression.

3) A graphical method for assessing variable selection for the multiple linear regression model is described. It is shown that for submodels I with k predictors, the widely used screen $C_p(I) \leq k$ is too narrow. More good submodels are considered if the screen $C_p(I) \leq \min(2k, p)$ is used. Variable selection methods originally meant for multiple linear regression can be extended to GLMs. See Chapter 13. Similar ideas from Olive and Hawkins (2005) have been incorporated in Agresti (2013). Section 3.4.1 shows how to bootstrap the variable selection estimator.

4) Asymptotically optimal prediction intervals for a future response Y_f are given for models of the form $Y = \boldsymbol{\beta}^T\boldsymbol{x} + e$ where the errors are iid,

unimodal, and independent of $\boldsymbol{x}$. Asymptotically optimal prediction regions are developed for multivariate linear regression.

5) Rules of thumb for selecting predictor transformations are given.

6) The DD plot is a graphical diagnostic for whether the predictor distribution is multivariate normal or from some other elliptically contoured distribution. The DD plot is also useful for detecting outliers in the predictors and for displaying prediction regions for multivariate linear regression.

7) The multivariate linear regression model has m response variables. Plots, prediction regions, and tests are developed that make this model nearly as easy to use as the multiple linear regression model ($m = 1$), at least for small m.

Throughout the book, there are goodness of fit and lack of fit plots for examining the model. The response plot is especially important.

The website (http://lagrange.math.siu.edu/Olive/lregbk.htm) for this book provides *R* programs in the file *lregpack.txt* and several *R* data sets in the file *lregdata.txt*. Section 14.1 discusses how to get the data sets and programs into the software, but the following commands will work.

Downloading the book's R functions *lregpack.txt* and data files *lregdata.txt* into *R*: The commands

```
source("http://lagrange.math.siu.edu/Olive/lregpack.txt")
source("http://lagrange.math.siu.edu/Olive/lregdata.txt")
```

can be used to download the *R* functions and data sets into *R*. Type *ls()*. Over 65 *R* functions from *lregpack.txt* should appear. In *R*, enter the command *q()*. A window asking *Save workspace image?* will appear. Click on *No* to remove the functions from the computer (clicking on *Yes* saves the functions on *R*, but the functions and data are easily obtained with the source commands).

Chapters 2–7 can be used for a one-semester course in regression and experimental design. For a course in generalized linear models, replace some of the design chapters by Chapter 13. Design chapters could also be replaced by Chapters 12 and 13. A more theoretical course would cover Chapters 1,10, 11, and 12.

Acknowledgments

This work has been partially supported by NSF grants DMS 0202922 and DMS 0600933. Collaborations with Douglas M. Hawkins and R. Dennis Cook were extremely valuable. I am grateful to the developers of useful mathematical and statistical techniques and to the developers of computer software and hardware (including R Core Team (2016)). Cook (1998) and Cook and Weisberg (1999a) influenced this book. Teaching material from this text has been invaluable. Some of the material in this text has been used in a Math 583 regression graphics course, a Math 583 experimental design course, and a Math 583 robust statistics course. In 2009 and 2016, Chapters 2 to 7 were used in Math 484, a course on multiple linear regression and experimental design. Chapters 11 and 12 were used in a 2014 Math 583 theory of linear

models course. Chapter 12 was also used in a 2012 Math 583 multivariate analysis course. Chapter 13 was used for a categorical data analysis course.

Thanks also goes to Springer, to Springer's associate editor Donna Chernyk, and to several reviewers.

Carbondale, IL, USA David J. Olive

http://www.springer.com/978-3-319-55250-7

Linear Regression

Olive, D.

2017, XIV, 494 p. 57 illus., Hardcover
ISBN: 978-3-319-55250-7