Multiple Linear and 1D Regression

David J. Olive
Southern Illinois University
Department of Mathematics
Mailcode 4408
Carbondale, IL 62901-4408
dolive@math.siu.edu

July 18, 2009

Contents

Preface

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

$$Y \perp\!\!\!\perp \boldsymbol{x}|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) \text{ or } \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}.$$

Many of the most used statistical methods are 1D models, including generalized linear models such as multiple linear regression, logistic regression, and Poisson regression. Single index models, response transformation models and many survival regression models are also included. The class of 1D models offers a unifying framework for these models, and the models can be presented compactly by defining the population model in terms of the sufficient predictor SP $= \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ and the estimated model in terms of the estimated sufficient predictor **ESP** $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$. In particular, the **response plot** or estimated sufficient summary plot of the ESP versus Y is used to visualize the conditional distribution $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$. The residual plot of the ESP versus the residuals is used to visualize the conditional distribution of the residuals given the ESP. The goal of this text is to present the applications of these models in a manner that is accessible to undergraduate and beginning graduate students.

Response plots are heavily used in this text. With the response plot the presentation for the $p > 1$ case is about the same as the $p = 1$ case. Hence the text immediately covers models with $p \geq 1$, rather than spending 100 pages on the $p = 1$ case and then covering multiple regression models with $p \geq 2$.

The literature on multiple linear regression is enormous. See Stigler (1986) and Harter (1974ab, 1975abc, 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), Comstock (1895), Bartlett (1900), Merriman (1910), Weld (1916), Leland (1921), Ezekial (1930), Bennett and Franklin (1954), Ezekial and Fox (1959) and Brownlee (1965).

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 Price (1977), Draper and Smith (1998), Fox (2008), Hamilton (1992), Kutner, Nachtsheim, Neter and Li (2004), Montgomery, Peck and Vining (2006), Mosteller and Tukey (1977), Ryan (2009) and Weisberg (2005). Cook and Weisberg (1999a) was a breakthrough because of its use of response plots.

Other texts of interest include Abraham and Ledolter (2006), Harrell (2006), Pardoe (2006), Mickey, Dunn and Clark (2004), Cohen, Cohen, West and Aiken (2003), Kleinbaum, Kupper, Muller and Nizam (1997), Mendenhall and Sinich (2003), Vittinghoff, Glidden, Shiblski and McCulloch (2005) and Berk (2003).

The author's hope is that this text's use of the response plot will make other regression texts obsolete much as Draper and Smith (1966) made earlier texts obsolete by using residual plots. The response plot is much more important than a residual plot since 1D regression is the study of the $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$, and the response plot is used to visualize this conditional distribution. The response plot emphasizes model goodness of fit and can be used to complement or even replace goodness of fit tests, while the residual plot of the ESP versus the residuals emphasizes model lack of fit. In this text the response plot is used to explain multiple linear regression, logistic regression, Poisson regression, single index models and models for experimental design. The response plot can also be used to explain and complement the ANOVA F and deviance tests for $\boldsymbol{\beta} = \mathbf{0}$.

This text provides an introduction to several of the most used 1D regression models. Chapter 1 reviews the material to be covered in the text and can be skimmed and then referred to as needed. Concepts such as interpretation of coefficients and interactions, goodness and lack of fit diagnostics, and variable selection are all presented in terms of the SP and ESP. The next few chapters present the multiple linear regression model. Then the one and two way ANOVA, logistic and Poisson regression models are easy to learn. Generalized linear models, single index models and general 1D models are also presented. Several important survival regression models are 1D models, but the sliced survival plot is used instead of the response plot to visualize the model.

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

- How can the conditional distribution $Y|(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$ be visualized?
- How can α and $\boldsymbol{\beta}$ be estimated?
- How can variable selection be performed efficiently?
- How can Y be predicted?
- What happens if a parametric 1D model is unknown or misspecified?

The author's research on 1D regression models includes visualizing the models, outlier detection, and extending least squares software, originally meant for multiple linear regression, to 1D models. Some of the applications in this text using this research are listed below.

- 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.
- It is shown how to use the response plot to detect outliers and to assess the adequacy of very general regression models of the form $Y = m(\boldsymbol{x}) + e$.
- A graphical method for selecting a response transformation for linear models is given. Linear models include multiple linear regression and many experimental design models.
- 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.

- Fast methods of variable selection for multiple linear regression, including an all subsets method, are extended to the 1D regression model. Plots for comparing a submodel with the full model after performing variable selection are also given.

- It is shown that least squares partial F tests, originally meant for multiple linear regression, are useful for exploratory purposes for a much larger class of 1D regression models.

- Asymptotically optimal prediction intervals for a future response Y_f are given for general regression models of the form $Y = m(\boldsymbol{x}) + e$ where the errors are iid, unimodal and independent of $\boldsymbol{x}$.

- Rules of thumb for selecting predictor transformations are given.

- 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.

- Graphical aids, including plots for overdispersion, for binomial regression models such as logistic regression are given.

- Graphical aids, including plots for overdispersion, for Poisson regression models such as loglinear regression are given.

- Graphical aids for survival regression models, including the Cox proportional hazards regression model and Weibull regression model, are given.

- 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 (www.math.siu.edu/olive/regbk.htm) for this book provides 28 data sets for *Arc*, and 40 *R/Splus* programs in the file *regpack.txt*. The students should save the data and program files on a disk. Chapter 17 discusses how to get the data sets and programs into the software, but the commands below will work for *R/Splus*.

Downloading the book's R/Splus functions *regpack.txt* into *R* or *Splus*:

Download *regpack.txt* onto a disk. Enter *R* and wait for the cursor to appear. Then go to the *File* menu and drag down *Source R Code*. A window should appear. Navigate the *Look in* box until it says *3 1/2 Floppy(A:)*. In the *Files of type* box choose *All files(*.*)* and then select *regpack.txt*. The following line should appear in the main *R* window.

```
> source("A:/regpack.txt")
```

If you use *Splus*, the command

```
> source("A:/regpack.txt")
```

will enter the functions into *Splus*. Creating a special workspace for the functions may be useful.

Type *ls()*. The *R/Splus* functions from *regpack.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 you have the functions on your disk).

Similarly, to download the text's *R/Splus* data sets, save *regdata.txt* on a disk and use the following command.

```
> source("A:/regdata.txt")
```

This text is an introduction to 1D regression models for undergraduates and beginning graduate students, and the prerequisites for this text are linear algebra and a calculus based course in statistics at the level of Hogg and Craig (1995), Hogg and Tanis (2005), Rice (2006), Wackerly, Mendenhall and Scheaffer (2008), or Walpole, Myers, Myers and Ye (2002). The student should be familiar with vectors, matrices, confidence intervals, expectation, variance, the normal distribution and hypothesis testing. This text may not be easy reading for nonmathematical students. Lindsey (2004) and Bowerman and O'Connell (1990) attempt to present regression models to students who have not had calculus or linear algebra. Also see Kachigan (1982, ch. 3–5) and Allison (1999).

This text will help prepare the student for the following courses.

1) Categorical data analysis: Agresti (2002, 2007) and Simonoff (2003).

2) Econometrics: see Greene (2007), Judge, Griffiths, Hill, Lütkepohl and Lee (1985), Kennedy (2008), and Woolridge (2008).

3) Experimental design: see Box, Hunter and Hunter (2005), Cobb (1998), Kirk (1982), Kuehl (1994), Ledolter and Swersey (2007), Maxwell and Delaney(2003), Montgomery (2005) and Oehlert (2000).

4) Exploratory data analysis: this text could be used for a course in exploratory data analysis, but also see Chambers, Cleveland, Kleiner and Tukey (1983) and Tukey (1977).

5) Generalized linear models: this text could be used for a course in generalized linear models, but also see Dobson and Barnett (2008), Fahrmeir and Tutz (2001), Hoffmann (2004), McCullagh and Nelder (1989) and Myers, Montgomery and Vining (2002).

6) Large sample theory for linear and econometric models: see White (1984).

7) Least squares signal processing: see Porat (1993).

8) Linear models: see Christensen (2002), Graybill (2000), Rao (1973), Ravishanker and Dey (2002), Scheffé (1959), Searle (1971) and Seber and Lee (2003).

9) Logistic regression: see Collett (2003) or Hosmer and Lemeshow (2000).

10) Poisson regression: see Cameron and Trivedi (1998) or Winkelmann (2008).

11) Numerical linear algebra: see Gentle (1998), Datta (1995), Golub and Van Loan (1989) or Trefethen and Bau (1997).

12) Regression graphics: see Cook (1998) and Li (2000).

13) Robust statistics: see Olive (2009a).

14) Survival Analysis: see Klein and Moeschberger (2003), Allison (1995), Collett (2003), or Hosmer, Lemeshow and May (2008).

15) Time Series: see Brockwell and Davis (2002), Chatfield (2003), Cryer and Chan (2008) and Shumway and Stoffer (2006).

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. For example, David (2006-7) 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 (2006-7) 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, type I and type II errors. Neyman also coined the term confidence interval.

David (2006-7) 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.

Acknowledgements

This work has been partially supported by NSF grant DMS 0202922 and DMS 0600933. Collaborations with Douglas M. Hawkins and R. Dennis Cook were extremely valuable. I am very grateful to the developers of useful mathematical and statistical techniques and to the developers of computer software and hardware. Teaching material from this text has been invaluable. Some of the material in this text has been used in two Math 484 multiple linear regression courses, two Math 485 categorical data courses, a Math 473 survival analysis course, a Math 583 regression graphics course, a Math 583 experimental design course and a Math 583 robust statistics course. Chapters 1 to 9 will be used in the Fall 2009 M484 course.