# Chapter 6

# **Regression Diagnostics**

Chapter 10, "Departures from Assumptions: Diagnosis and Remedies" describes the use of residuals and influence measures in diagnosing situations in which the assumptions on the Gaussian linear model,  $\mathcal{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ , are called into question.

For the purposes of examining residuals, it is convenient to express the distribution,  $\mathcal{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta},\sigma^2 \mathbf{I}_n)$ , in a "signal plus noise" form. Although this is a useful separation in linear regression models, it is not universally applicable. In particular, it does not help to write a generalized linear model in this way — but that doesn't stop people from trying it.

Back to the Gaussian linear model, if we subtract the expected value of  $\mathcal{Y}$ , which is  $X\beta$ , we are left with a "white noise" vector,  $\epsilon$ . That is,

$$\mathcal{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n).$$

The distribution of  $\epsilon$  is described as "white noise" because it has no signal left it in. Its elements are independent, constant variance, mean zero, normally distributed random variables.

Given the observed data,  $\boldsymbol{y}$ , from which we estimate the coefficients,  $\hat{\boldsymbol{\beta}}$ , our estimate of the "signal" component is  $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$  and our estimate of the noise component is the *residual vector*,  $\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ . We want to check the Gauss-Markov assumptions

$$E[\boldsymbol{\epsilon}] = \boldsymbol{0}$$
$$\operatorname{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$$

Although not strictly part of the Gauss-Markov assumptions, we also usually check for a normal or Gaussian distribution. Together these provide the desired distribution,  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ .

As is frequently the case, checking for normality is either difficult or redundant. With a small sample size it is difficult to tell much about the distribution. With a large sample size, the Central Limit Theorem takes over, more-or-less, because the estimates of the coefficients,  $\hat{\beta}$ , which are linear combinations of the responses, tend towards a normal distribution. This is a rather loose statement and there should be some conditions imposed to ensure that the central limit effect will be in force but the general idea is that linear combinations behave somewhat like sample means.

# 6.1 Plots of the residuals

We use residual plots to check for the following possible violations of the assumptions:

- 1. Non-constant variance
- 2. Outliers
- 3. Non-normal distribution
- 4. Incorrect specification of the "signal" part,  $X\beta$ .

#### 6.1.1 Basic, "canned" residual plots

Plots that are always of interest are:

- 1. (Raw) residuals,  $\widehat{e} = y \widehat{y}$ , versus the fitted values,  $\widehat{y} = X\widehat{\beta}$ .
- 2. A normal quantile-quantile plot (also called a normal probability plot) of the standardized residuals.

Canned versions of these plots are available as plot(lm1, which=1), for the residuals versus fitted plot, and plot(lm1, which=2), for the normal quantile-quantile plot of the standardized residuals. You can obtain both plots in a horizontal layout by changing the default layout, calling the plot method with which=1:2, and restoring the layout. The idiom is

```
> opar <- par(mfrow=c(1,2), las=1)
> plot(lm1, which=1:2)
> par(opar)
```

The first call to **par** sets the layout to be 1 row and 2 columns and adjusts the label style for horizontal axis labels. The old values of the graphical parameters are saved for later restoration. It is a good idea always to restore graphical parameters to the default state after finishing a plot. That way you are always starting from a known state.

For the model,

```
> lm2 <- lm(log(Volume) ~ 1 + log(Girth) + log(Height), trees)
```

these plots are shown in Fig. 6.1

The plot of the residuals versus the fitted values should show points scattered within a horizontal band. In any model with an intercept term the residuals will sum to zero and we must have both positive and negative residuals. The reference line y = 0 is drawn on the plot as is a scatterplot smoother curve showing the general trend in the residuals as they depend on the fitted values. Extreme residuals are marked with the row number in the original data set or the row names, if available.

The foremost purpose of this plot is to evaluate whether the variability about the fitted model increases as the level of the response increases. Although this is only one way in which the variance of the noise term can fail to be constant, it is by far the most common violation of constant variance. Recall model av3



Figure 6.1: Residual plots for the model 1m2.

> summary(av3 <- aov(breaks ~ wool \* tension, warpbreaks))</pre>

Df Sum Sq Mean Sq F value Pr(>F)wool 1 450.7 450.67 3.7653 0.0582130 tension 2 2034.3 1017.13 8.4980 0.0006926 2 1002.8 4.1891 0.0210442 wool:tension 501.39 48 5745.1 119.69 Residuals

The residual plots for this model, Fig. 6.2, show such a pattern. All the extreme residuals are from the larger fitted values and the vertical spread of the points in the right hand side of the plot is noticeably greater than for those on the left.

This pattern of increasing variability with increasing fitted response indicates a need for transformation of the response using the type of transformation that squeezes large values together and stretches small values apart. The inverse transformation that we used in model av3a

```
> summary(av3a <- aov(1/breaks ~ wool * tension, warpbreaks))</pre>
              \mathtt{Df}
                               Mean Sq F value
                                                   Pr(>F)
                    Sum Sq
wool
               1 0.0002403 0.00024035
                                         0.9001 0.347511
               2 0.0033455 0.00167274
                                         6.2642 0.003826
tension
wool:tension
               2 0.0012088 0.00060442
                                         2.2635 0.114978
Residuals
              48 0.0128174 0.00026703
```



Figure 6.2: Residual plots for the model av3.

is such a transformation. John Tukey's "ladder of reexpression" gives successively more extreme versions of such transformations (when applied to positive values of the response, which is commonly the case). Writing  $y^*$  for the transformed value, this set of transformations is

original scale  $y^* = y$ square root  $y^* = \sqrt{y}$ logarithm  $y^* = \log(y)$ inverse square root  $y^* = (y)^{-1/2}$ 

inverse  $y^* = 1/y$ 

All of these transformations are contained within the *Box-Cox* family of transformations proposed by George Box and Sir David Cox. This is a parameterized family of transformations, depending on a parameter  $\lambda$  and defined as

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda} - 1}{\lambda} & \lambda \neq 0\\ \log(y) & \lambda = 0 \end{cases}$$

Not only is this family continuous in y > 0 for a fixed  $\lambda$ , it is also continuous in  $\lambda$  for a fixed y, even at  $\lambda = 0$ . The continuity at  $\lambda = 0$  can be verified using l'Hôpital's rule.

#### 6.1. PLOTS OF THE RESIDUALS

Given a model form,  $X\beta$ , we can evaluate the parameter estimates for various values of  $\lambda$ and determine which  $\lambda$  value maximizes the log-likelihood. The log-likelihood at the optimal  $\beta$ and  $\sigma$  for each value of  $\lambda$  is called the *profiled log-likelihood*. If we write the optimal value as  $\hat{\lambda}$ then a likelihood ratio test of  $H_0: \lambda = \lambda_0$  versus  $H_a: \lambda \neq \lambda_0$  consists of comparing the profiled log-likelihood at  $\hat{\lambda}$  to that at  $\lambda_0$ .

The boxcox function from the MASS package automates this procedure and, by default, creates a plot of the profiled log-likelihood, Fig. 6.3

> library(MASS)

> boxcox(av3)



Figure 6.3: Profiled log-likelihood for  $\lambda$  in the Box-Cox transformation family applied to model av3. The left panel shows the default range of  $\lambda$  from -2 to 2. In the right panel the range is restricted to -0.6 to 0.6.

In this figure the plot of the profiled log-likelihood is shown twice, once on the default range of (-2, 2) and then on the restricted range (-0.6, 0.6) which more closely bounds the 95% confidence interval on  $\lambda$ .

Interestingly, the optimal value  $\hat{\lambda}$  is very close to zero, indicating the logarithmic transformation will be a good choice. The inverse transformation, which we chose based on comparative boxplots, corresponding to  $\lambda = -1$ , is not in the confidence interval. Refitting the model with the logarithmic transformation,

```
> summary(av3e <- aov(log(breaks) ~ tension * wool, warpbreaks))</pre>
```

	$\mathtt{Df}$	Sum Sq	Mean Sq	F value	Pr(>F)
tension	2	2.1762	1.08808	7.7792	0.001185
wool	1	0.3125	0.31253	2.2344	0.141511
tension:wool	2	0.9131	0.45657	3.2642	0.046863
Residuals	48	6.7138	0.13987		

provides a p-value just under 5% for the interaction term. If we chose to strictly follow the 5% significance level we would keep the interaction term and, hence, each of the main effects.

Box and Cox noted that frequently the transformation that produces a stable variance also produces a simpler model form. In fact, we found that on the reciprocal scale the model can be reduced to two coefficients, the intercept and a linear term in tension. On the logarithmic scale it may not be possible to simplify to that extent.

#### 6.2 Verbatim notes from Brittany Schwefel

Check Gauss-Markov assumptions on error vector using graphical and numerical methods.

- $E(\epsilon) = 0$
- $Var(\epsilon) = \sigma^2 I_n$
- $\epsilon \sim N(0, \sigma^2 I_n)$

For large sample size, the Central Limit Theorem applies.

#### 6.3 Residuals

Regression function is linear, but with following possible violations

- Error terms do not have constant variance
- Outliers
- Error terms are not normally distributed

Plots to check:

- Residuals vs. Predictor Variables
- [Residuals] vs. Predictor Variables
- Residuals vs. Time (Or person collecting data, other systematic ways of collecting data, etc.)
- Residuals vs. Omitted predictor variables
- Boxplots of residuals

How to deal with violation of residual plot

#### 6.3. RESIDUALS

- Weighted least squares
- Transformation of variables (response or predictors)
- Changing linear model

QQ-plots - Deviation

Are errors independent?

- Sequence plot pattern
- Residuals vs. predictor

Omission of Important Predictor Variables - Residuals against variables omitted from model - Does not mean that the model is wrong, it can just be improved

#### 6.3.1 Added Variable (Partial Regression Plots)

$$\hat{Y}(X_1) = \hat{\beta}_0 + \hat{\beta}_1 X_1$$
$$\hat{\epsilon}(Y|X_1) = Y - \hat{Y}(X_1)$$
$$\hat{X}_2(X_1) = \hat{\beta}_0^* + \hat{\beta}_1^* X_1$$
$$\hat{\epsilon}(X_2|X_1) = X_2 - \hat{X}_2(X_1)$$

Would reveal whether we need  $X_2$  in the model.  $X_2$  would show pattern for the residuals  $Y \sim X_1$ vs  $X_2 \sim X_1$ . If there is a pattern, we must transform  $X_2$ 

#### 6.3.2 Overfitting

Overfitting - using too many parameters

$$\begin{split} Y &= X\beta + \epsilon \qquad X_j = j^{th} \text{variable} \qquad X_{(j)} \text{design matrix without } j^{th} \text{ variable} \\ \text{Examine } Y \sim X_{(j)} \widehat{\beta}_{(j)} \text{ vs. } X_j \sim \beta_{(j)} X_{(j)} \\ \qquad Y &= X_{(j)} \beta_{(j)} + \beta_j X_j + \epsilon \\ H_{(j)} &= X_{(j)} (X'_{(j)} X_{(j)})^{-1} X'_{(j)} \\ [I - H_{(j)}] Y &= [I - H_{(j)}] X_{(j)} \beta_{(j)} + [I - H_{(i)}] \beta_j X_j \Rightarrow \text{Residuals for} X_{(i)} = Y - X_{(i)} \beta_j X_j \end{split}$$

$$[I - H_{(j)}]Y = [I - H_{(j)}]X_{(j)}\beta_{(j)} + [I - H_{(j)}]\beta_j X_j \Rightarrow \text{Residuals for} X_{(j)} = Y - X_{(j)}\beta_{(j)}$$
$$[I - H_{(j)}]X_j\beta_j = X_j - X_{(j)}(X'_{(j)}X_{(j)})^{-1}X'_{(j)}X_j = X_j - X_{(j)}\beta_{(j)}$$

The slope of added variable plot is value for  $\beta_j$ 

### 6.4 Outliers and Influential Points

- 1. Regression outliers  $\Rightarrow$  unusual Y values
- 2. High leverage points  $\Rightarrow$  unusual X values
- 3. Influential points affect fit of data

Are outliers due to data collection or data entry?

#### 6.4.1 Leverages

$$P = H = X(X'X)^{-1}X'$$
 Projection Matrix

 $h_{ii} \rightarrow$  leverage for the  $i^{th}$  data point

$$Y_{i} = \beta_{0} + \beta_{1}X_{i1} + \dots + \beta_{p}X_{ip} + \epsilon_{i} \quad i = 1, \dots, n$$
$$= \alpha + \beta_{1}(X_{i1} - \bar{X}_{1}) + \dots + \beta_{p}(X_{ip} - \bar{X}_{p}) + \epsilon_{i}$$
$$\sum X_{i}$$

$$\bar{X}_j = \frac{\sum X_{ij}}{n}$$
  $\alpha = \beta_0 + \beta_1 \bar{X}_1 + \dots + \beta_p \bar{X}_p$ 

$$Y = \begin{bmatrix} 1 & \dots & X_c \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_c \end{bmatrix} + \epsilon \qquad \beta_c = \begin{bmatrix} \beta_1 & \dots & \beta_p \end{bmatrix}'$$

Proof.

$$X = \begin{bmatrix} 1 & X_{11} & \dots & X_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{bmatrix} \qquad X_c = (I - \frac{1}{n}J) \begin{bmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \vdots & \vdots \\ X_{n1} & \dots & X_{np} \end{bmatrix} \qquad J = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \vdots & \vdots \\ 1 & \dots & 1 \end{bmatrix} \\ H_c = X_c (X'_c X_c)^{-1} X'_C \qquad \hat{Y} - \bar{Y}_j + X_c (X'_c X_c)^{-1} X'_c Y = [\frac{1}{n}J + H_c] Y$$

**Theorem 10.** X is a n x (p+1) of rank(p+1) < n. X[,1] = j. Then  $h_{ij}$  of  $H = X(X^TX)^{-1}X^T$  has the following properties:

1.

$$\frac{1}{n} \le h_{ii} \le 1 \qquad i = 1, \dots, n$$

2.

$$-\frac{1}{2} \le h_{ij} \le \frac{1}{2} \quad \forall \ i \ \neq \ j$$

3.

$$h_{ii} = \frac{1}{n} + (X_{1i} - Z)' (X'_c X_c)^{-1} (X_{1i} - Z)$$

 $i^{th}$  row -  $X_{1i} = (X_{1i} \dots X_{ip})'$  with covariate vector for the  $i^{th}$  data point  $Z = (\bar{X}_1 \dots \bar{X}_p)$ 

Proof. Part 1: 
$$H = \frac{1}{n}J + X_c(X'_cX_c)^{-1}X'_c$$
 where  $X_c(X'_cX_c)^{-1}X'_c$  is positive definite  
 $\min(h_{ii}) = \frac{1}{n}$   $H = H^2$   $h_{ii} = h'_ih_i = (h_{i1}\dots h_{in}) \begin{pmatrix} h_{i1} \\ \vdots \\ h_{in} \end{pmatrix} = \sum_{j=1}^n h_{ij}^2$   
 $h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \Rightarrow 1 = h_{ii} + \sum_{j \neq i} \frac{h_{ij}^2}{h_{ii}} \ge 0$ 

Proof. Part 2: 
$$h_{ii} = h_{ii}^2 + h_{ij}^2 + \sum_{r \neq i,j} h_{ir}^2 \implies h_{ii} - h_{ir}^2 \le h_{ij}^2 \quad h_{ii} - h_{ir}^2 \quad \text{max when } h_{ii} = \frac{1}{2} \implies h_{ij}^2 \le \frac{1}{4} \implies h_{ij} \in [-.5, .5]$$

*Proof.* Read off  $i^{th}$  diagonal of H from  $X_c(X'_cX_c)^{-1}X'_c$ 

### 6.5 Types of Residuals

1. Regular  $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ 

2. Internally standardized residuals  $r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$ 

3. Externally standardized residuals  $r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$  where (i) represents the estimate with the  $i^{th}$  entry deleted

4. Deleted/Jackknife Residuals  $\frac{Y_{(i)} - \hat{Y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + X'_i(X'_{(i)}X_{(i)})^{-1}X_i}}$  where  $X_{(i)}$  is the design matrix without the  $i^{th}$  row and  $X_i$  is the  $i^{th}$  row of the design matrix

 $\hat{Y}_{(i)}$  : obtained by estimating  $\widehat{oldsymbol{eta}}$  based on 1,..., i-1, i+1, ..., n data points

$$Var(Y_i - \hat{Y}_i) = Var(Y_i) + Var(\hat{Y}_{(i)})$$
$$= \sigma^2 + X'_i Var(\hat{\beta}_{(i)}) X_i$$
$$= \sigma^2 (1 + X'_i (X'_{(i)} X_{(i)})^{-1} X_i)$$

$$\hat{\epsilon}_{(i)} = Y_i - X_i \widehat{\beta}_{(i)}$$

Theorem 11.

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - \frac{\widehat{\epsilon}_i}{1 - X_{ii}} (X'X)^{-1} X_i$$

 $\widehat{oldsymbol{eta}}_{(i)}$  obtained by excluding the  $i^{th}$  observation

*Proof.* 1. Show that  $X'X = X'_{(i)}X_{(i)} + X_iX'_i$ .

- 2. Show that  $X'Y = X'_{(i)}Y_{(i)} + X_iY_i$ .
- 3. Show that  $(X'X)^{-1}X'_{(i)}Y_{(i)} = \widehat{\beta} (X'X)^{-1}X_iY_i$

4. 
$$\widehat{\boldsymbol{\beta}}_{(i)} = ((X'X)^{-1} + \frac{(X'X)^{-1}X_iX_i'(X'X)^{-1}}{1 - h_{ii}})(X'_{(i)}Y_{(i)})$$

Hint: if B is nonsingular, square matrix, and c is vector,  $[B - CC']^{-1} = B^{-1} + \frac{B^{-1}CC'B^{-1}}{1 - C'B^{-1}C}$ 5. Put 3 and 4 together to obtain the result

$$\hat{\epsilon}_{(i)} = \frac{\epsilon_i}{1 - h_{ii}}$$

Proof.

$$\begin{aligned} \hat{\epsilon}_{(i)} &= Y_i - X_i \widehat{\beta}_{(i)} = Y_i - X_i' [\widehat{\beta} - \frac{\hat{\epsilon}_i}{1 - h_{ii}} (X'X)^{-1} X_i] \\ &= Y_i - X_i \widehat{\beta} + \frac{X_i' \hat{\epsilon}_i (X'X)^{-1} X_i}{1 - h_{ii}} \\ &= Y_i - \hat{Y}_i + \frac{\hat{\epsilon}_i}{1 - h_{ii}} X_i' (X'X)^{-1} X_i \\ &= \hat{\epsilon}_i + \frac{\epsilon_i h_{ii}}{1 - h_{ii}} \end{aligned}$$

Show that 
$$\frac{\hat{\epsilon}_i}{\hat{\epsilon}_{(i)}\sqrt{1-h_{ii}}} = \frac{\hat{\epsilon}_{(i)}}{\sqrt{Var(\hat{\epsilon}_{(i)})}}$$
 where  $\sqrt{Var(\hat{\epsilon}_{(i)})} = \hat{\sigma}_{(i)}\sqrt{1+X_i'(X_{(i)}'X_{(i)})^{-1}X_i}$ 

$$Var(\hat{\epsilon}_{(i)}) = Var(\frac{\hat{\epsilon}}{1 - h_{ii}}) = \frac{1}{(1 - h_{ii})^2} Var(\hat{\epsilon}_i) = \frac{\sigma^2 (1 - h_{ii})}{(1 - h_{ii})^2} = \frac{\sigma^2}{1 - h_{ii}} \text{ which is estimated by } \frac{\sigma_{(i)}^2}{1 - h_{ii}}$$
$$\frac{\hat{\epsilon}_{(i)}}{\sqrt{Var(\hat{\epsilon}_{(i)})}} = \frac{\hat{\epsilon}_i}{1 - h_{ii}} \frac{\sqrt{1 - h_{ii}}}{\hat{\sigma}_{(i)}} = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}$$

#### 6.5.1 Hypothesis Testing with Residuals

 $H_0:\ i^{th}$  data point (externally studentized residual) not an outlier  $H_A:\ {\rm not}\ {\rm so}$ 

$$\frac{\hat{\epsilon}_i}{\sigma_{(i)}^2 \sqrt{1-h_{ii}}} \sim t_{n-p-1,\alpha/(2n)} = t_{(n-1)-p}$$

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}^2 \qquad \frac{(n-p)\hat{\sigma}_{(i)}^2}{\sigma^2} \sim \chi_{n-p-1}^2$$
On

High leverage point  $> \frac{2p}{n}$ 

### 6.5.2 Residuals in R

- Regular residuals: residuals(mod1)
- Internally standardized residuals: rstandard(mod1)

- Externally studentized residuals: rstudent(mod1)
- Leverage values: hatvalues(mod1)
- Cook's Distance: cooks.distance(mod1)

# 6.6 Cook's Distance

$$D_i := \frac{(\widehat{\beta} - \widehat{\beta}_{(i)})' X' X (\widehat{\beta} - \widehat{\beta}_{(i)})}{p \widehat{\sigma}^2} = \frac{\sum_{j=1}^n (\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{p \sigma^2}$$
  
If  $r_i := \frac{\widehat{\epsilon}_i}{\widehat{\sigma} \sqrt{1 - h_{ii}}}$ , then Cook's Distance  $= \frac{r_i^2}{p} (\frac{h_{ii}}{1 - h_{ii}})$ .  $D_i$  is an abnormal value if  $D_i > \frac{4}{n - p}$ 

# 6.7 Box Cox Transformation

Y - original response

$$Y(\lambda) = \begin{cases} \frac{Y^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0\\ \log(y) & \text{if } \lambda = 0 \end{cases}$$

then the resulting Y is normally distributed.  $Y(\lambda) \sim N(X\beta, \sigma^2 I)$ .  $\lambda$  is unknown. Found through likelihood and maximization.

$$f(Y_i, X_i, i = 1, \dots, n | \lambda, \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} exp\{\frac{-1}{2\sigma^2}(Y_i(\lambda) - \beta_0 - \beta_1 X_i)\}$$

Maximize  $\lambda, \beta_0, \beta_1, \sigma^2$ 

# Bibliography

- C. A. Bache, J. W. Serum, W. D. Youngs, and D. J. Lisk. Polychlorinated Biphenyl residues: Accumulation in Cayuga Lake trout with age. *Science*, 117:1192–1193, 1972.
- George E. P. Box and George C. Tiao. Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading, MA, 1973.
- George E. P. Box and Paul W. Tidwell. Transformations of the independent variables. *Technometrics*, 4:531–550, 1962.
- J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart. *Linpack Users' Guide*. SIAM, Philadelphia, 1979.
- G. A. F. Seber. Linear Regression Analysis. Wiley, New York, 1977.
- G. W. Stewart. Introduction to Matrix Computations. Academic Press, New York, 1973.