Presented at the 2016 ICEAA Professional Development & Training Workshop - www.iceaaonline.com/atlanta2016

Develop PRESS for Nonlinear Equations

Dr. Shu-Ping Hu

29 March 2016
Tecolote Research PRT-204

Outline

- Objectives
- Definitions
- Error Terms and Models
- Calculate PRESS for Linear Models
  - Hat Matrix and Leverage Value
  - Single Regression Method
- Use of PRESS and Predicted R^2
- Algorithm
- Examples for Validation
- Conclusions

Objectives

- Explain the importance of using the predicted residual sum of squares (PRESS) and predicted R^2 statistics in regression analysis
- Recommend using PRESS and Predicted R^2 for cross-validation when deriving nonlinear CERs
- Develop an algorithm to generate PRESS and Predicted R^2 by a **single regression** for nonlinear equations
  - Avoid the downside of running nonlinear regression **multiple** times
  - Validate the single regression approach for three error terms: Additive, MUPE, and Log-Error

Definition – PRESS

- **Leave-one-out statistic**: PRESS is defined as the sum of the squares of all residuals such that the predicted value is calculated for the omitted observation in each refitted regression model:

$$\textbf{PRESS} = \sum_{i=1}^{n} w_i \left(y_i - \hat{y}_{i,-i}\right)^2$$

- Each residual in PRESS is termed

$$\textbf{PRESS Residual}_i = y_i - \hat{y}_{i,-i}$$

n = sample size
w_i = weighting factor for the i^{th} data point
y_i = i^{th} observation
$\hat{y}_{i,-i}$ = i^{th} predicted value from the equation fitted without the i^{th} observation

The smaller the PRESS, the better the model's predictive capability

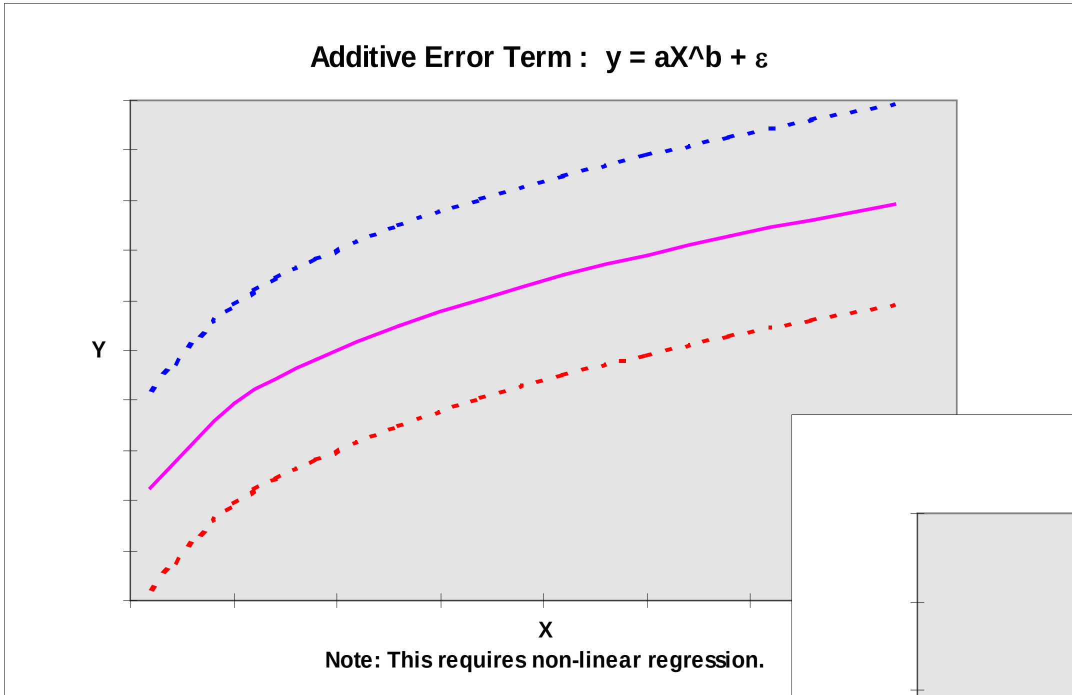
Definition – Predicted R^2

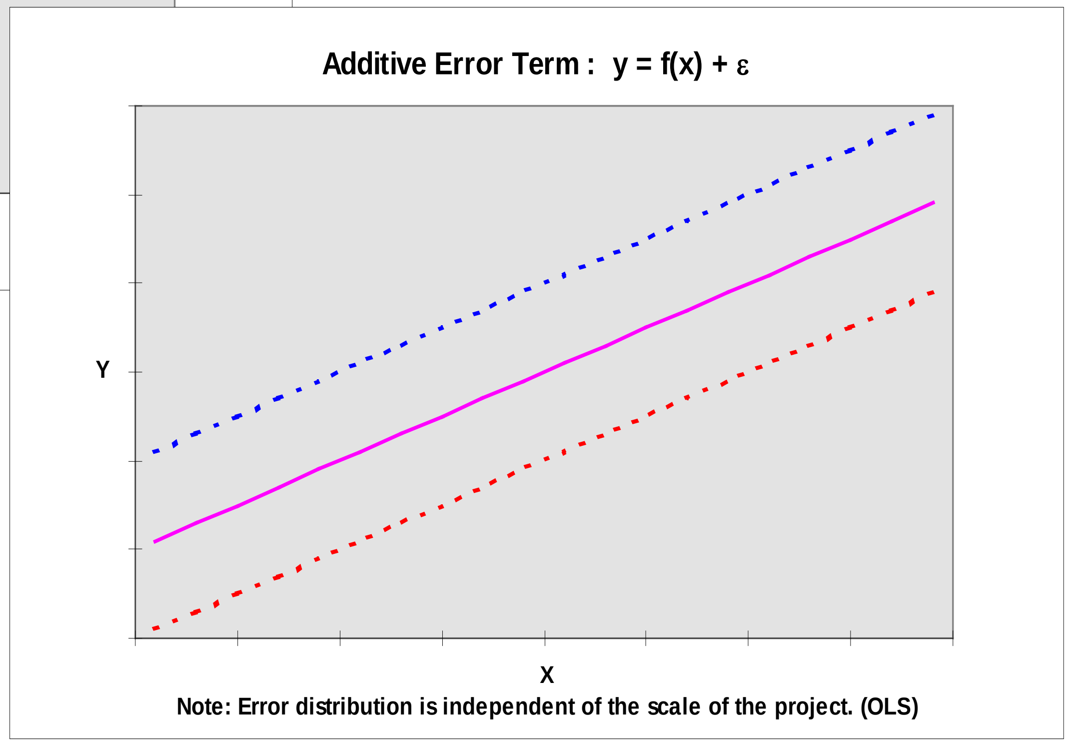
- Predicted R^2 = 1 – PRESS/SST; SST = $\Sigma_i \, w_i(y_i - \overline{y}_w)^2$
  - SST is the total sum of squares for the dependent variable
  - $\overline{y}_w$ is the weighted average of the dependent variable
- PRESS is an absolute measure; Predicted R^2 puts PRESS in perspective
- Predicted R^2 is a more useful measure than Adjusted R^2 for assessing the model's predictive power as it is calculated using observations **not** included in the curve fitting process

PRESS and Predicted R^2 are commonly used to determine
(1) how well the model predicts new observations
(2) if there are excessive independent variables

Error Terms and Models (1/3)
Additive Error Term: y = f(x) + ε

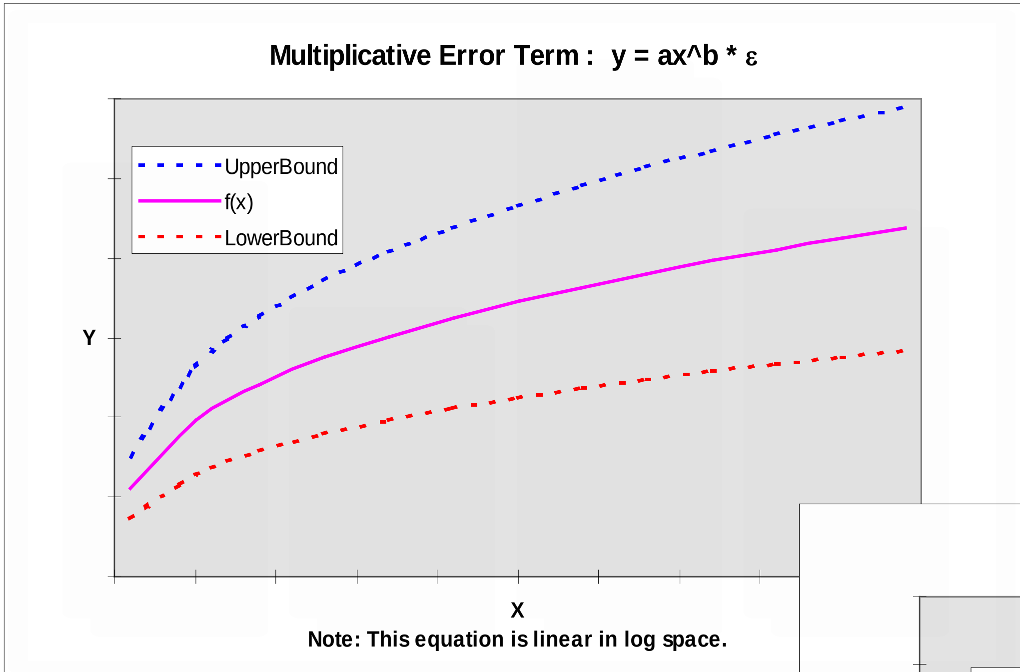
For additive-error models: curve fitting is done in unit space





Cost variation is independent of the scale of the project

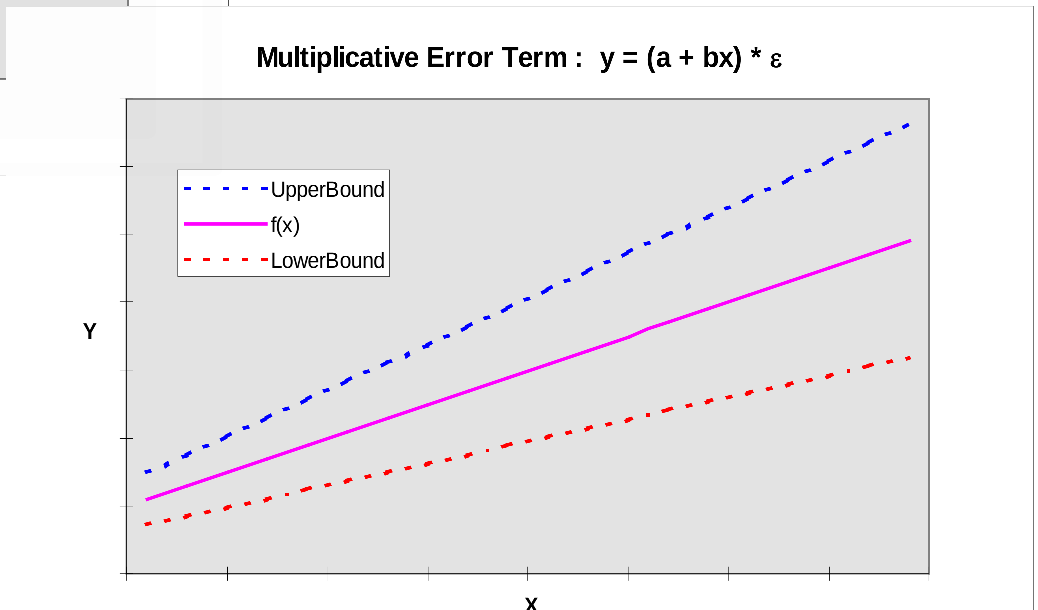
Error Terms and Models (2/3)
Multiplicative Error Term: y = f(x)*ε



Note: This equation is linear in log space.

Multiplicative error assumption is appropriate when

- Errors in the dependent variable are believed to be proportional to the magnitude of the function (the value of the variable)
- Dependent variable ranges over more than one order of magnitude



Note: This requires non-linear regression.

Cost variance is proportional to the scale of the project

Error Terms and Models (3/3)

Multiplicative Error Model: $y = f(x)*\varepsilon$

- Log-Error: $\varepsilon \sim LN(0, \sigma^2)$ ⇨ Least squares in log space
  - Error = Log (y) - Log f(x)
  - Minimize the sum of squared errors; process is done in log space

If f(x) is linear in log space, it is termed a log-linear or LOLS (log space **ordinary least squares**) CER

- MUPE: $E(\varepsilon) = 1$, $V(\varepsilon) = \sigma^2$ ⇨ Least squares in weighted space
  - Error = **(y-f(x))/f(x)**
  - Minimize the sum of squared (%) errors iteratively (i.e., minimize $\Sigma_i \{(y_i-f(x_i))/f_{k-1}(x_i)\}^2$, k is the iteration number)
  - MUPE, an iterative, **w**eighted **l**east **s**quares (WLS) has zero sample bias

variance of error term

Note: $E((y-f(x))/f(x)) = 0$

$V((y-f(x))/f(x)) = \sigma^2$

- ZMPE: $E(\varepsilon) = 1$, $V(\varepsilon) = \sigma^2$ ⇨ Least squares in weighted space
  - Error = **(y-f(x))/f(x)**
  - Minimize the sum of squared (percentage) errors with a constraint:

    $$\Sigma_i(y_i - f(x_i))/f(x_i) = 0$$

  - ZMPE is a constrained minimization process
  - Average sample bias is eliminated by the constraint

We will focus on additive-error, MUPE, and log-error equations in this paper

Calculate PRESS for Linear Models (1/2)
Hat Matrix and Leverage Value

- Hat Matrix: $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}$ $(\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ if $\boldsymbol{V} = \boldsymbol{I})$
- Leverage value (denoted by h_{ii}) is the test statistic for an extreme value of the predictors:

$$h_{ii} = w_i \mathbf{x}_i (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1} \mathbf{x}_i' \quad \text{for } i = 1, \ldots, n \ \S$$

- h_{ii} is the i^{th} diagonal element of H
  - $\sum_{i=1}^{n} \boldsymbol{h_{ii}} = \boldsymbol{Trace}(\boldsymbol{H}) = \boldsymbol{p}$ $(0 < h_{ii} < 1)$
  - If h_{ii} > 2p/n or 3p/n, data point i is flagged as a leverage point, which has an unusual value of the predictors (note: average of h_{ii}'s is p/n)
  - The hat matrix is very useful for outlier analysis (e.g., standardized residual)
  - The predicted value ($\hat{Y}$) can be expressed using H:

$$\hat{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}Y = \boldsymbol{H}Y$$

$$\hat{y}_i = h_{i1}y_1 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n \quad for\ i = 1, \ldots, n$$

$\boldsymbol{X}$ = the design matrix
$\mathbf{V}$ = the variance/covariance matrix of the errors
$\mathbf{W}$ = the weighting matrix = $\boldsymbol{V}^{-1}$
$\mathbf{x}_i$ = the i^{th} driver vector; $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$ if intercept is included. (k + 1 = p)
p = the total number of estimated parameters
n = the sample size
$\hat{\boldsymbol{\beta}}$ = the vector of estimated parameters
§ : this equation assumes data points uncorrelated

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$
$if\ \ p = 2, \mathbf{V} = \mathbf{I}$

The hat matrix only depends on the independent variables

Calculate PRESS for Linear Models (2/2) Single Regression Method

- Use a single regression to calculate the PRESS residual and PRESS:

$$\boldsymbol{PRESS\ Residual(i) = y_i - \hat{y}_{i,-i} = (y_i - \hat{y}_i)/(1 - h_{ii})} \quad \text{for i = 1,..., n}$$

$$PRESS = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i,-i})^2 = \sum_{i=1}^{n} w_i \frac{(y_i - \hat{y}_i)^2}{(1-h_{ii})^2}$$

  - $\hat{y}_{i,-i}$ = i^th predicted value from the equation fitted without data point i
  - $\hat{y}_i$ = i^th predicted value from the equation fitted using **all** data points
  - See Montgomery and Peck (1992) for a proof of the formula
  - Note: the proof is based upon a useful identity found in matrix inverse operation (see Morrison, 1976 or Bartlett, 1951):

$$(\mathbf{Z} + c\mathbf{bb'})^{-1} = \mathbf{Z}^{-1} - \frac{c}{1 + c\mathbf{b'Z}^{-1}\mathbf{b}} \mathbf{Z}^{-1}\mathbf{bb'Z}^{-1}$$

Z = p-by-p nonsingular matrix
b = p-by-1 vector
c = scalar

- This single regression method takes only one iteration instead of refitting multiple regression equations to compute PRESS for linear models

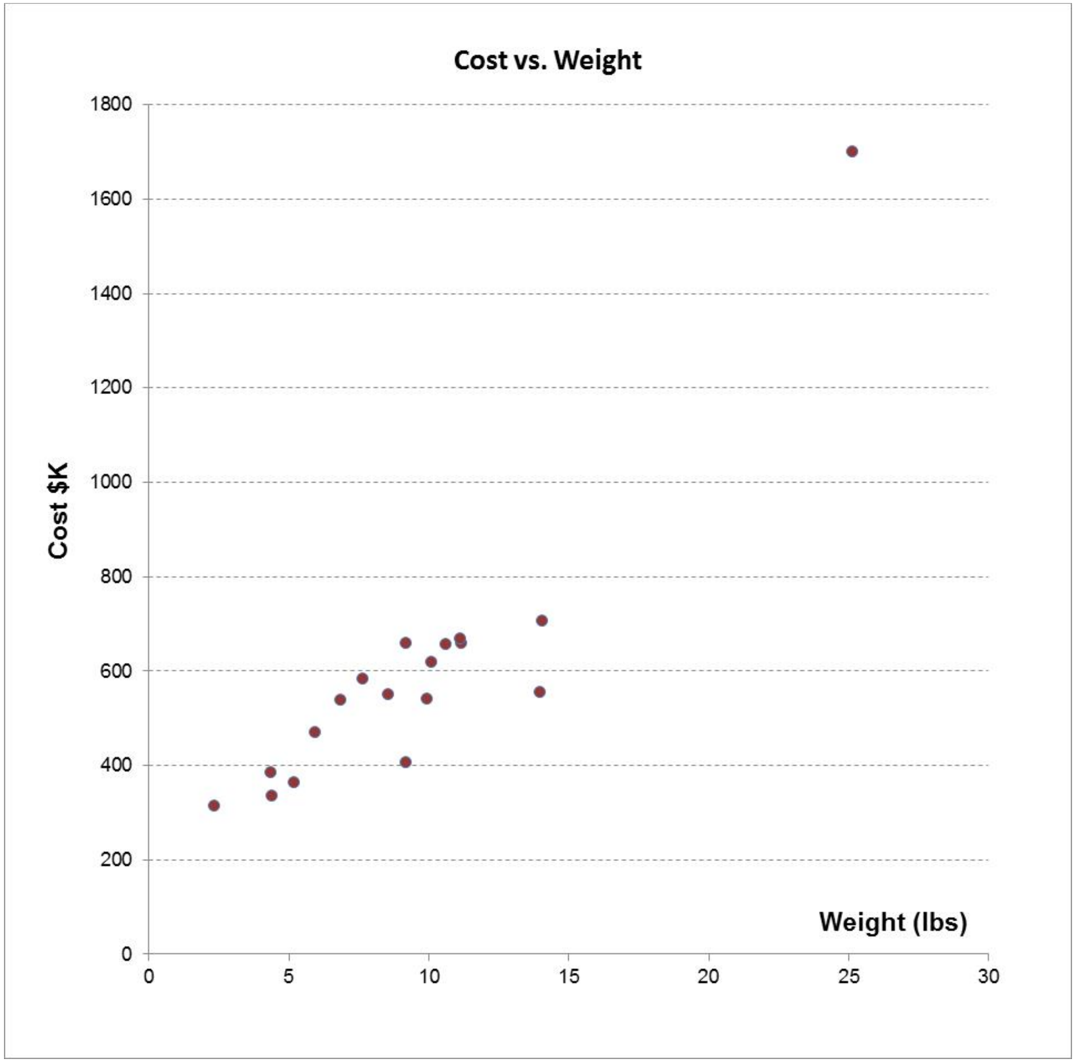
Use of PRESS and Predicted R^2 (1/2)

Example: Cost = a + b*Weight

- A hypothetical data set to predict the cost of a black box using

Data Point	Cost $K	Weight in lbs	WF
Obs 1	538.84	6.83	1.00
Obs 2	363.77	5.16	1.00
Obs 3	405.83	9.17	1.00
Obs 4	549.91	8.53	1.00
Obs 5	619.03	10.06	1.00
Obs 6	660.29	11.15	0.85
Obs 7	470.80	5.92	1.00
Obs 8	668.95	11.10	1.00
Obs 9	385.21	4.33	1.00
Obs 10	583.21	7.62	1.00
Obs 11	337.02	4.37	1.00
Obs 12	555.82	13.94	0.75
Obs 13	542.05	9.92	1.00
Obs 14	707.17	14.02	1.00
Obs 15	660.15	9.18	1.00
Obs 16	315.45	2.32	1.00
Obs 17	656.81	10.59	0.62
Obs 18	1701.28	25.12	1.00

WF is the weighting factor for each observation; the factors are assigned arbitrarily for demonstration purposes



Use of PRESS and Predicted R^2 (2/2)

Example: Cost = 78 + 55.5*(Weight) + ε

- Outlier analysis table:

Obs #	Cost	Predicted Y Value	Residual	Std. Residual	Leverage	Cook's Distance
1	538.84	457.1922	81.647777	0.719363	0.072174	0.020127
2	363.77	364.4906	-0.720644	-0.006438	0.097687	0.000002
3	405.83	587.0855	-181.255453	-1.584996	0.058111	0.077497
4	549.91	551.5591	-1.649100	-0.014431	0.059443	0.000007
5	619.03	636.4893	-17.459289	-0.152779	0.059406	0.000737
6	660.29	696.9951	-36.705110	-0.295601	0.056084	0.002596
7	470.80	406.6782	64.121811	0.568733	0.084478	0.014923
8	668.95	694.2196	-25.269613	-0.221850	0.065559	0.001727
9	385.21	318.4174	66.792596	0.602606	0.115163	0.023631
10	583.21	501.0451	82.164934	0.720981	0.064596	0.017948
11	337.02	320.6378	16.382199	0.147725	0.114248	0.001407
12	555.82	851.8678	-296.047807	-2.269557	0.080871	0.226605
13	542.05	628.7179	-86.667899	-0.758213	0.058960	0.018010
14	707.17	856.3086	-149.138601	-1.341295	0.109559	0.110678
15	660.15	587.6406	72.509448	0.634060	0.058105	0.012401
16	315.45	206.8425	108.607550	1.012128	0.170678	0.105413
17	656.81	665.9096	-9.099550	-0.062009	0.038389	0.000077
18	1701.28	1472.4688	228.811204	3.220733	0.636488	9.081357

Goodness-of-Fit statistics table:

Std Error (SE)	R2	R2 Adjusted
117.8320	85.72%	84.82%
Pearson's Corr Coef	**PRESS**	**R2 Predicted**
0.9258	599,480.8	61.46%

$$PRESS = \sum_{i=1}^{n} w_i \frac{(\text{Residual}_i)^2}{(1-\text{Leverage}_i)^2} = \sum_{i=1}^{n} w_i \frac{(y_i - \hat{y}_i)^2}{(1-h_{ii})^2} = 599{,}480.8$$

Predicted R^2 = 1 – PRESS/SST
= 1 – 599,480.8/1,555,385
= 61.46%

- The predicted R^2 (61%) is much smaller than the adjusted R^2 (85%) for this CER, which indicates that the model does not predict new observations as well as it fits existing data
- A log-linear model using this data set also delivers the same message

Algorithm
Partial Derivative Matrix (1/2)

- Given a nonlinear model, $f(\mathrm{x},\theta)$, with an additive error term $\boldsymbol{\epsilon}$:

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \epsilon_i = f_\mathrm{i} + \epsilon_i \quad (for\ i = 1, \ldots, n)$$

- Taylor series expansion of the nonlinear function f at a given point $\boldsymbol{\theta}^{\mathbf{0}}$

$$f(\mathbf{x}_\mathrm{i}, \boldsymbol{\theta}) \cong f(\mathbf{x}_\mathrm{i}, \boldsymbol{\theta}^0) + \left.\frac{\partial f_i}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} (\boldsymbol{\theta} - \boldsymbol{\theta}^0) = f_i^{\,o} + \left.\frac{\partial f_i}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} (\boldsymbol{\theta} - \boldsymbol{\theta}^0)$$

y_i = the observed dep variable value of the i^{th} data point (i = 1,…,n); n is the sample size

$f_i = f(\mathbf{x_i}, \boldsymbol{\theta})$ = the value of the hypothesized equation for the i^{th} data point

ϵ_i = the error term for the i^{th} data point with a mean of 0 and a variance σ_i^2

$\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ik})$, a set of k predictor variables for the i^{th} data point

$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_\mathrm{p})'$, a vector of unknown parameters; $\boldsymbol{\theta}^{\mathbf{0}} = (\theta_1^0, \theta_2^0, \ldots, \theta_\mathrm{p}^0)'$, a given point of $\boldsymbol{\theta}$

$(\partial f_i / \partial \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\mathbf{0}}}$ is the partial derivatives of $f(\mathbf{x}_i, \boldsymbol{\theta})$ with respect to (w.r.t.) $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta}^{\mathbf{0}}$

- The model can be further simplified using matrix notations:

$$\boldsymbol{y} \cong \boldsymbol{f}^{\mathbf{0}} + \boldsymbol{Z}_{\mathbf{0}}(\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathbf{0}}) + \boldsymbol{\epsilon} \quad \Rightarrow \boldsymbol{y} - \boldsymbol{f}^{\mathbf{0}} \cong \boldsymbol{Z}_{\mathbf{0}}(\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathbf{0}}) + \boldsymbol{\epsilon}$$

  - $\boldsymbol{f}^{\mathbf{0}} = (f_1^0, f_2^0, \ldots, f_\mathrm{n}^0)'$, an n x 1 vector of the hypothesized CER, evaluated at $\boldsymbol{\theta}^{\mathbf{0}}$
  - $\mathbf{Z}_\mathrm{o}$ is an n x p matrix of the partial derivatives of $\boldsymbol{f}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}^0$

Algorithm
Partial Derivative Matrix (2/2)

- **Given:** $y \cong f^0 + Z_0(\theta - \theta^0) + \epsilon \Rightarrow y - f^0 \cong Z_0(\theta - \theta^0) + \epsilon$

$$\mathbf{Z}_0 = \frac{\partial f}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = \begin{pmatrix} \frac{\partial f_1}{\partial \theta_1} & \frac{\partial f_1}{\partial \theta_2} & \cdots & \frac{\partial f_1}{\partial \theta_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_i}{\partial \theta_1} & \frac{\partial f_i}{\partial \theta_2} & \cdots & \frac{\partial f_i}{\partial \theta_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_n}{\partial \theta_1} & \frac{\partial f_n}{\partial \theta_2} & \cdots & \frac{\partial f_n}{\partial \theta_p} \end{pmatrix}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = \begin{pmatrix} \frac{\partial f_1}{\partial \boldsymbol{\theta}} \\ \cdots \\ \frac{\partial f_i}{\partial \boldsymbol{\theta}} \\ \cdots \\ \frac{\partial f_n}{\partial \boldsymbol{\theta}} \end{pmatrix}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = \begin{pmatrix} \mathbf{z}_1 \\ \cdots \\ \mathbf{z}_i \\ \cdots \\ \mathbf{z}_n \end{pmatrix}$$

Note: $\mathbf{z_i}$ is the $\mathrm{i^{th}}$ row vector of $\mathbf{Z_o}$

$$\mathbf{z}_i = \frac{\partial f_i}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^o} = \left(\frac{\partial f_i}{\partial \theta_1} \quad \frac{\partial f_i}{\partial \theta_2} \quad \cdots \quad \frac{\partial f_i}{\partial \theta_p}\right)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^o}$$

  - $\mathbf{Z}_o$ (an n-by-p partial derivative matrix) is used as a proxy for OLS's design matrix $\mathbf{X}$

- Nonlinear (NL) PRESS residual for data point *i* can be estimated by

$$(y_i - \hat{y}_{i,-i}) \cong \frac{(y_i - \hat{y}_i)}{(1 - w_i \mathbf{z}_i (\mathbf{Z}_0^{'} \mathbf{W} \mathbf{Z}_0)^{-1} \mathbf{z}_i')} = \frac{(y_i - \hat{y}_i)}{(1 - H_i)} \quad (i = 1, \ldots, n)$$

$$H_i = w_i \mathbf{z}_i (\mathbf{Z}_0^{'} \mathbf{W} \mathbf{Z}_0)^{-1} \mathbf{z}_i' \quad \textit{for NL}$$
$$h_{ii} = w_i \mathbf{x}_i (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{x}_i' \quad \textit{for WLS}$$

  - Just as h_{ii} for OLS, H_i is used to estimate the NL leverage value for the $\mathrm{i^{th}}$ data point

- PRESS for a nonlinear model is then approximated by

$$\textit{Nonlinear PRESS} = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i,-i})^2 \cong \sum_{i=1}^{n} w_i \frac{(y_i - \hat{y}_i)^2}{(1 - H_i)^2}$$

Note: w_i is weighting factor for the $\mathrm{i^{th}}$ data point

Use partial derivative matrix to approximate the "design matrix" for nonlinear CER

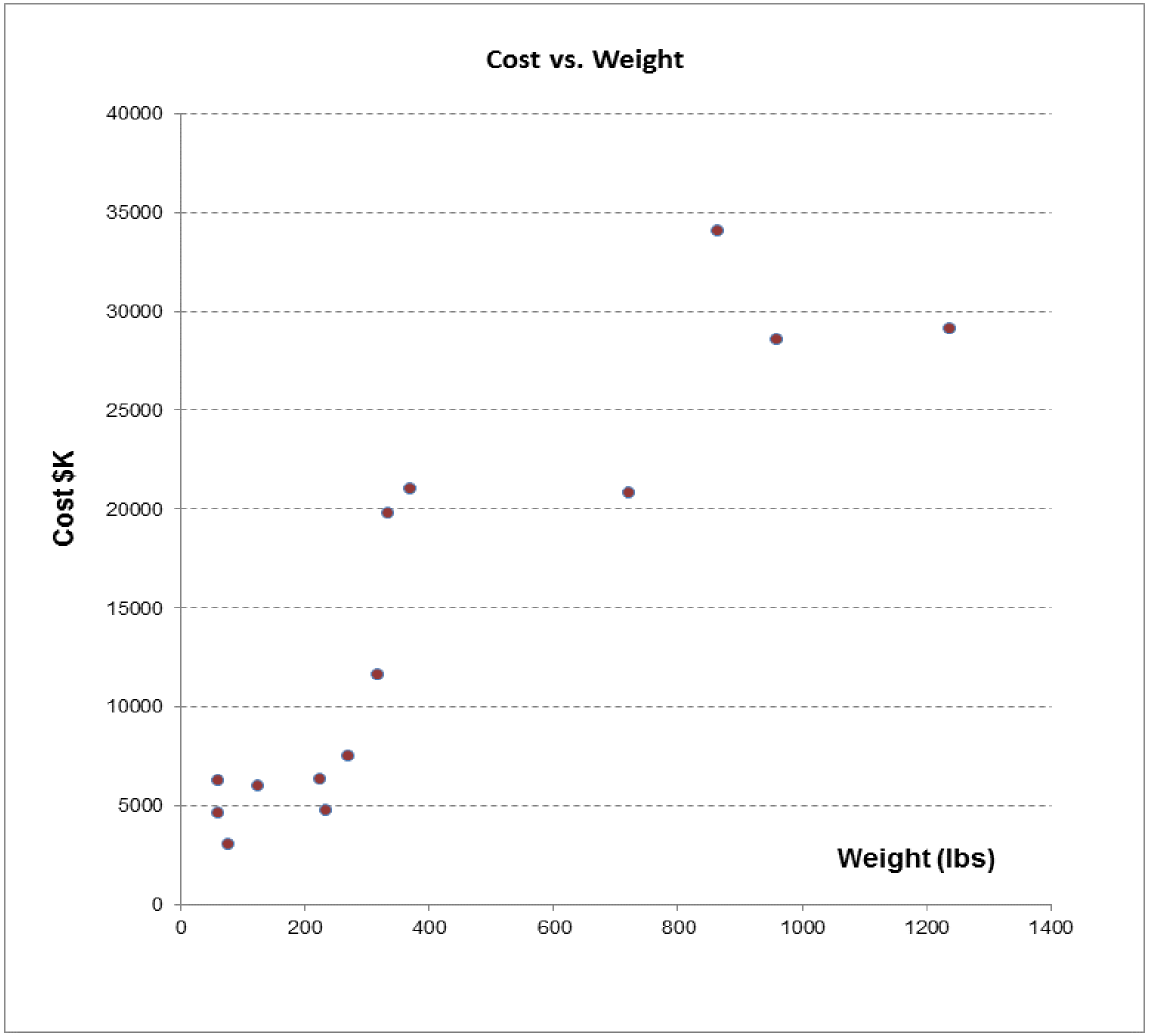
Examples for Validation (1/5)

Data Set

- A hypothetical data set of generic satellite electronic units:

Data Point	Cost $K	Weight in lbs	WF
Obs 1	3106.64	77.05	1.00
Obs 2	29166.32	1236.77	0.62
Obs 3	4820.48	232.14	1.00
Obs 4	34111.22	863.36	1.00
Obs 5	6387.04	224.40	1.00
Obs 6	20871.60	720.44	0.75
Obs 7	28621.92	959.33	1.00
Obs 8	19796.80	332.50	1.00
Obs 9	7526.40	269.42	0.50
Obs 10	6002.24	123.84	1.00
Obs 11	11668.48	316.15	1.00
Obs 12	6329.12	59.77	1.00
Obs 13	4683.20	59.17	1.00
Obs 14	21068.72	369.12	1.00

WF is the weighting factor for each observation; the factors are assigned arbitrarily for demonstration purposes



Examples for Validation (2/5)

Additive Error: $y = 225.6x(\text{Weight}^{0.7089}) + \varepsilon$ Adj. R^2 = 82.4%

- Comparison of PRESS Residuals between Single Regression and Direct Approach:

Data Point	WF (w)	Residual ($y_i - \hat{y}_i$)	H_i	$y_i - \hat{y}_{i,-i}$ Single Regr	$y_i - \hat{y}_{i,-i}$ LOO Def	% Difference
Obs 1	1.00	-1,800.10	0.08381	-1,964.77	-1,956.66	0.41%
Obs 2	0.62	-5,936.22	0.38303	-9,621.64	-9,688.84	-0.69%
Obs 3	1.00	-5,902.66	0.12094	-6,714.77	-6,660.69	0.81%
Obs 4	1.00	6,903.81	0.22278	8,882.69	8,868.27	0.16%
Obs 5	1.00	-4,081.41	0.12077	-4,642.04	-4,606.37	0.77%
Obs 6	0.75	-3,060.16	0.11428	-3,455.01	-3,454.69	0.01%
Obs 7	1.00	-696.20	0.29367	-985.66	-979.13	0.67%
Obs 8	1.00	5,963.23	0.11769	6,758.65	6,742.03	0.25%
Obs 9	0.50	-4,390.71	0.06033	-4,672.61	-4,656.59	0.34%
Obs 10	1.00	-866.56	0.10543	-968.69	-962.24	0.67%
Obs 11	1.00	-1,679.36	0.11862	-1,905.38	-1,893.06	0.65%
Obs 12	1.00	2,230.74	0.07178	2,403.23	2,385.89	0.73%
Obs 13	1.00	614.02	0.07131	661.17	657.43	0.57%
Obs 14	1.00	6,171.70	0.11555	6,977.99	6,968.17	0.14%
PRESS				332,790,557	331,656,343	0.34%
Pred R^2				75.81%	75.89%	0.1%

Single Regression Approx: $y_i - \hat{y}_{i,-i} \cong (y_i - \hat{y}_i)/(1 - H_i)$

$y_1 - \hat{y}_{1,-1} \cong (y_1 - \hat{y}_1)/(1 - H_1)$
$= (-1{,}800.1)/(1-0.08381) = -1{,}964.77$

$$PRESS = \sum_{i=1}^{n} w_i \left(y_i - \hat{y}_{i,-i}\right)^2$$

Single Regression Approximation:
Predicted R^2 = 1 – PRESS/SST
= 1 – 332,790,557/1,375,549,914
= 75.81%

LOO Definition:
Predicted R^2 = 1 – PRESS/SST
= 1 – 331,656,343/1,375,549,914
= 75.89%

- The last column of the table illustrates the percent differences of the PRESS residuals calculated by a single regression and their definition. The percent differences are all within 1%.
- The percent difference between the two PRESS statistics is 0.34%, which is well within 1%. The respective Predicted R^2 measures are 75.81% and 75.89%. They are about the same.

Examples for Validation (3/5)

MUPE Error Term

- MUPE's PRESS statistic is the sum of all percentage errors:

$$MUPE\text{'s } PRESS = \sum_{i=1}^{n} w_i (y_i - \hat{y}_{i,-i})^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_{i,-i})^2}{(\hat{y}_{i,-i})^2} = \sum_{i=1}^{n} (\%\ Error_{(i)})^2$$

- For the MUPE CER, the weighting factor (w_i) for each data point is the square of the reciprocal of its predicted value
- The predicted values (circled in the equation) are derived from the leave-one-out model to match the definition
- For the single regression method, the predicted value, $\hat{y}_{i,-i}$, is calculated as the actual cost minus the "approximated" PRESS residual: $\hat{y}_{i,-i} \cong Actual - PRESS\ Residual = y_i - (y_i - \hat{y}_{i,-i})$
  - where $(\boldsymbol{y_i} - \boldsymbol{\hat{y}_{i,-i}})$ is computed using a single regression approximation
  - This mod is made to match the leave-one-out definition

Examples for Validation (4/5)

MUPE Error: $y = 241.06x(Weight^{0.69115}) \times \varepsilon$ Adj. R^2 = 68.7%

- Comparison of MUPE's PRESS Residuals & % Errors between Single Regression and Direct Approach (Def):

Data Point	Residual $(y_i - \hat{y}_i)$	H_i	$y_i - \hat{y}_{i,-i}$ Single Regr	$y_i - \hat{y}_{i,-i}$ LOO Def	% Error Single Regr	% Error LOO Def
Obs 1	-1,748.13	0.19334	-2,167.11	-2,139.71	41%	41%
Obs 2	-3,898.10	0.24316	-5,150.50	-5,049.08	15%	15%
Obs 3	-5,583.84	0.07352	-6,026.96	-6,026.13	56%	56%
Obs 4	8,319.85	0.17099	10,035.90	9,947.85	-42%	-41%
Obs 5	-3,776.26	0.07447	-4,080.09	-4,079.19	39%	39%
Obs 6	-1,887.45	0.14205	-2,199.96	-2,185.35	10%	9%
Obs 7	881.53	0.19013	1,088.48	1,075.85	-4%	-4%
Obs 8	6,459.69	0.07424	6,977.74	6,975.68	-54%	-54%
Obs 9	-4,005.93	0.07145	-4,314.18	-4,314.19	36%	36%
Obs 10	-736.93	0.11922	-836.68	-833.13	12%	12%
Obs 11	-1,211.84	0.07296	-1,307.22	-1,307.06	10%	10%
Obs 12	2,255.86	0.24700	2,995.83	2,956.66	-90%	-88%
Obs 13	638.25	0.24933	850.24	835.49	-22%	-22%
Obs 14	6,732.88	0.07812	7,303.42	7,298.36	-53%	-53%
PRESS					242.9%	237.9%
Pred R^2					56.4%	57.3%

Single Regression Approx: $y_i - \hat{y}_{i,-i} \cong (y_i - \hat{y}_i)/(1 - H_i)$

- $y_1 - \hat{y}_{1,-1} \cong (y_1 - \hat{y}_1)/(1 - H_1) = (-1{,}748.13)/(1 - 0.19334) = -2{,}167.11$
- $\hat{y}_{1,-1} = y_1 - (y_1 - \hat{y}_{1,-1}) = 3106.64 - (-2167.11) = 5274.41$
- $\%\text{Error} = 2167.11/5274.41 = 41\%$

$$MUPE's\ PRESS = \sum_{i=1}^{n} \left(\%\ Error_{(i)}\right)^2$$

Single Regression Approximation:
Predicted R^2 = 1 – PRESS/SST
= 1 – 2.429/5.56962 = **56.4%**

LOO Definition:
Predicted R^2 = 1 – PRESS/SST
= 1 – 2.379/5.56962 = **57.3%**

- The individual PRESS percentage error approximated by the single regression method closely follows its definition except for observation 12, which is off by 2% (blue columns). The individual PRESS residuals derived by the single regression are also very close to those calculated directly (orange columns); the differences are all within 2%.
- The percentage difference between the two MUPE's PRESS statistics is 2.1%, which is still reasonably good. The respective Predicted R^2 measures are calculated to be 56.4% and 57.3%. The difference is less than 1%.

Examples for Validation (5/5)

Log Error: $y = 200.1x(\text{Weight}^{0.7167}) \times \varepsilon$ Adj. R^2 = 76.7%

- Comparison of PRESS Residuals between Single Regression and Direct Approach:

Data Point	WF (w)	Residual $\ln(y_i) - \ln(\hat{y}_i)$	H_i	$\ln(y_i) - \ln(\hat{y}_{i,-i})$ Single Regr	$\ln(y_i) - \ln(\hat{y}_{i,-i})$ LOO Def	% Diff
Obs 1	1.00	-0.3711	0.19835	-0.4629	-0.4629	0%
Obs 2	0.62	-0.1210	0.17465	-0.1466	-0.1466	0%
Obs 3	1.00	-0.7222	0.07859	-0.7838	-0.7838	0%
Obs 4	1.00	0.2932	0.19893	0.3660	0.3660	0%
Obs 5	1.00	-0.4165	0.07927	-0.4523	-0.4523	0%
Obs 6	0.75	-0.0683	0.12403	-0.0780	-0.0780	0%
Obs 7	1.00	0.0422	0.22098	0.0542	0.0542	0%
Obs 8	1.00	0.4330	0.08311	0.4722	0.4722	0%
Obs 9	0.50	-0.3834	0.03893	-0.3989	-0.3989	0%
Obs 10	1.00	-0.0526	0.12212	-0.0599	-0.0599	0%
Obs 11	1.00	-0.0595	0.08118	-0.0648	-0.0648	0%
Obs 12	1.00	0.5225	0.25448	0.7009	0.7009	0%
Obs 13	1.00	0.2286	0.25693	0.3076	0.3076	0%
Obs 14	1.00	0.4203	0.08843	0.4611	0.4611	0%
PRESS				2.2969	2.2969	0%
Pred R^2				70.93%	70.93%	0%

Single Regression Approximation:

$$ln(y_i) - ln(\hat{y}_{i,-i}) \cong (ln(y_i) - ln(\hat{y}_i))/(1 - H_i)$$

$$ln(y_1) - ln(\hat{y}_{1,-1}) \cong (ln(y_1) - ln(\hat{y}_1))/(1 - H_1) = (-0.3711)/(1-0.19835) = -0.4629$$

$$PRESS = \sum_{i=1}^{n} w_i\left(\ln(y_i) - \ln(\hat{y}_{i,-i})\right)^2$$

Single Regression Approx:
Predicted R^2 = 1 – PRESS/SST
= 1 – 2.2969/7.9017 = 70.93%

LOO Definition:
Predicted R^2 = 1 – PRESS/SST
= 1 – 2.2969/7.9017 = 70.93%

- This case shows there is no difference between the two PRESS statistics as well as their respective predicted R^2 measures. This is because the power-form equation becomes linear in log space and the single regression method (based upon the partial derivative matrix Z_o) precisely estimates the leverage values for the log-linear CER.

Conclusions

- PRESS and Predicted R^2 are useful statistics for cross-validation
- Use PRESS and Predicted R^2 for all regression equations
- Apply a single regression algorithm to approximate PRESS and Predicted R^2 for nonlinear equations, including MUPE
  - The algorithm has been validated for three error terms using a sample data
  - Recommend this simple, effective approach to generate PRESS for nonlinear CERs
- Future study: Develop PRESS and Predicted R^2 for ZMPE CERs

References

- Allen, D. M., “The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction,” Technometrics, 16, (1974), 125–127.
- Bartlett, M. S., “An inverse adjustment arising in discriminant analysis,” Annals of Mathematical Statistics, Volume 22 (1951), pp. 107-111.
- Belsley, D. A.; E. Kuh; R. E. Welsch, “Regression Diagnostics,” John Wiley & Sons, Inc., 1981.
- Book, S. A. and N. Y. Lao, “Minimum -Percentage-Error Regression under Zero-Bias Constraints,” Proceedings of the 4th Annual U.S. Army Conference on Applied Statistics, 21-23 Oct 1998, U.S. Army Research Laboratory, Report No. ARL-SR-84, November 1999, pages 47-56.
- Cook, R. D. and S. Weisberg, “Residuals and Influence in Regression,” Chapman and Hall, 1982.
- Draper, N. R. and H. Smith, “Applied Regression Analysis (2nd edition),” New York: John Wiley & Sons, Inc., 1981.
- Hu, S., “The Minimum-Unbiased-Percentage-Error (MUPE) Method in CER Development,” 3rd Joint Annual ISPA/SCEA International Conference, Vienna, VA, 12-15 June 2001.
- Hu, S. and A. R. Sjovold, "Multiplicative Error Regression Techniques," 62nd MORS Symposium, Colorado Springs, Colorado, 7-9 June 1994.
- Hu, S. and A. Smith, “Why ZMPE When You Can MUPE,” 6th Joint Annual ISPA/SCEA International Conference, New Orleans, LA, 12-15 June 2007.
- Montgomery D. C. and E. A. Peck, “Introduction to Linear Regression Analysis (2nd edition),” Wiley -Interscience, 1992.
- Morrison, D. F., “Multivariate Statistical Methods, 2nd ed.”, New York: McGraw-Hill Book Company, 1976.
- Morrison, D. F., “Applied Linear Statistical Methods”, New Jersey: Prentice-Hall, Inc. 1983.
- Seber, G. A. F. and C. J. Wild, “Nonlinear Regression,” New York: John Wiley & Sons, 1989, pages 37, 46, 86-88.
- Wedderburn, R. W. M., “Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method,” Biometrika, Vol. 61, Issue 3 (December 1974), pages 439-447.
- Weisberg, S., “Applied Linear Regression, 2nd Edition,” New York: John Wiley & Sons, 1985, pages 87-88.