Linear Statistical Modeling Methods with SAS

Model Selection and Weighted Least Squares

Xuemao Zhang
East Stroudsburg University

March 25, 2024

Outline

- Strategies for Model Building
- Criteria for Model Selection
  - R^2 and adjusted R^2
  - Mallows' C_p Criterion
  - AIC_p and BIC_p Criteria
  - $PRESS_p$ Criterion
- Automatic Search Procedures for Model Selection
  - "Best" Subsets Algorithms
  - Stepwise Regression Methods
- Weighted Least Squares

Strategies for Model Building

For the following MLR model, model building means specifying the form of the model: the response, predictors and regressors. The dilemman evolves from an uncertainty of what terms (regressors) to be included in the model.

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1 X_1(Z_1, Z_2, \ldots, Z_p) \\ &\quad + \beta_2 X_2(Z_1, Z_2, \ldots, Z_p) \\ &\quad + \ldots + \beta_k X_k(Z_1, Z_2, \ldots, Z_p). \end{aligned}$$

Strategies for Model Building:

- Which predictor variables and regressors should be included.
- Structure for their use:
  - **(a)** first order model
  - **(b)** second order model
  - **(c)** different function forms (transformations)

Model Building:

(1) Try different models
  - **(a)** limit choices to those that make sense to the experts
  - **(b)** eliminate more complicated structures that are not confirmed by the data

(2) Compare the different models and select the best one

(3) Model Validation (to be discussed in machine learning)

Criteria for Model Selection: R^2 and adjusted R^2

The use of R^2, coefficient of multiple determination, is to identify several "good" subsets of X variables-in other words, subsets for which R^2 is high.

The R^2 criterion is equivalent to using the error sum of squares SSE since

$$R^2 = 1 - \frac{SSE}{SST}.$$

Since R^2 does not take account of the number of parameters in the regression model, the adjusted coefficient of multiple determination R_a^2 has been suggested as an alternative criterion:

$$R_a^2 = 1 - \frac{\text{SSE}/(n-1-k)}{SS_{total}/(n-1)} = 1 - \left(\frac{n-1}{n-1-k}\right)\frac{SSE}{SST}.$$

Criteria for Model Selection: Mallows' C_p Criterion

This criterion is concerned with the total mean squared error of the n fitted values for each subset regression model. The mean squared error (MSE) of $\widehat{Y}_i|_{X=x_0=(1,x_{1i},\ldots,x_{ki})}$ is defined as

$$MSE(\widehat{Y}_i) = E\left[\widehat{Y}_i - \mu_i\right]^2,$$

where μ_i is the **true mean response** when the levels of the predictor variables X are those for the ith case, $i = 1, \ldots, n$.

It can be shown that

$$MSE(\widehat{Y}_i) = \left[E(\widehat{Y}_i) - \mu_i\right]^2 + Var(\widehat{Y}_i),$$

where $Var(\widehat{Y}_i)$ is the variance of the fitted value $\widehat{Y}_i$,
$E(\widehat{Y}_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$ based on a proposed model

$$E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

Note. If the above model is the **true model**, then

$$E(\widehat{Y}_i) = \mu_i, i = 1, 2, \ldots, n.$$

Criteria for Model Selection: Mallows' C_p Criterion

Now the total mean squared error for all n fitted values is given by

$$\sum_{i=1}^{n} MSE(\widehat{Y}_i) = \sum_{i=1}^{n} \left[E(\widehat{Y}_i) - \mu_i \right]^2 + \sum_{i=1}^{n} Var(\widehat{Y}_i).$$

The criterion measure, denoted by Γ_p, is simply the total mean squared error divided by σ^2:

$$\Gamma_p = \frac{1}{\sigma^2} \left[\sum_{i=1}^{n} \left[E(\widehat{Y}_i) - \mu_i \right]^2 + \sum_{i=1}^{n} Var(\widehat{Y}_i) \right].$$

It can then shown that an estimator of Γ_p is C_p:

$$C_p = \frac{SSE_p}{MSE(X_1, \ldots, X_k)} - (n - 2p)$$

where $p = 1 + k$, where SSE_p is the error sum of squares for the fitted subset regression model with $p = k + 1$ parameters (k X variables).

When there is no bias in the regression model with k X variables so that $E(\widehat{Y}_i) = \mu_i$,

$$E(C_p) \approx p.$$

Criteria for Model Selection: Mallows' C_p Criterion

Remark.

- One favors the candidate model with the smallest C_p value.
- Norm: $C_p = p$ suggests that the model contains no estimated bias.
- In practical situations, some candidate models yield $C_p < p$.
- A C_p much larger than p occurs with a heavily biased model.

Criteria for Model Selection: AIC_p and BIC_p Criteria

- Two popular alternatives that also provide penalties for adding predictors are Akaike's information criterion (AIC_p) and Schwarz' Bayesian criterion BIC_p (also called SBC_p).
  - p is the number of predictors in the model.
- We search for models that have **small values** of AIC_p, or BIC_p, where these criteria are given by:

$$AIC_p = n \ln SSE_p - n \ln n + 2p$$
$$SBC_p = n \ln SSE_p - n \ln n + (\ln n)p$$

- Models with **small** SSE_p will do well by these criteria, as long as the penalties $2p$ for AIC_p and $(\ln n)p$ for BIC_p are not too large.

Criteria for Model Selection: $PRESS_p$ Criterion

Residuals may have two problems when used to evaluate model fit:

- They have different standard errors depending on the value of the regressor.
- They are not independent.

The ordinary residuals are not uncorrelated in general. The **PRESS residuals** remedy this problem.

Consider the model

$$E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

PRESS residuals are computed as follows:

- First, observation (X_i, Y_i) is omitted from the data and the least squares line fit to the remaining data, giving the parameter estimates $\hat{\beta}_0^{(i)}, \hat{\beta}_1^{(i)}, \ldots, \hat{\beta}_k^{(i)}$.
- Next, the **deleted fitted value**, $\hat{Y}_{(i)} = \hat{\beta}_0^{(i)} + \hat{\beta}_1^{(i)} X_{1i} + \cdots \hat{\beta}_k^{(i)} X_{ki}$ is computed, $i = 1, \ldots, n$.
- Then, the **deleted residual** or **PRESS residuals** $e_{(i)} = Y_i - \hat{Y}_{(i)}$ **is computed,** $i = 1, \ldots, n$**.**

Criteria for Model Selection: $PRESS_p$ Criterion

The $PRESS_p$ Criterion is the sum of the squared PRESS residuals over all n cases

$$PRESS_p = \sum_{i=1}^{n} (Y_i - \hat{Y}_{(i)})^2.$$

Remark. Models with small $PRESS_p$ values are considered good candidate models.

$PRESS_p$ values can be calculated without requiring n separate regression runs, each time deleting one of the n cases. Namely,

$$PRESS_p = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

Example

Surgical Unit Example

A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 108 patients was available for analysis. From each patient record, the following information was extracted from the preoperation evaluation:
X_1 - blood clotting score
X_2 - prognostic index
X_3 - enzyme function test score
X_4 - liver function test score
X_5 - age, in years
X_6 - indicator variable for gender (0 = male, 1 = female)
X_7 and X_8 - indicator vmiables for history of alcohol use:

Alcohol Use	X_7	X_8
None	0	0
Moderate	1	0
Severe	0	1

These constitute the pool of potential explanatory or predictor variables for a predictive regression model. The response variable is survival time, which was ascertained in a followup study.

Example

```
data ch9tab01;
input x1 x2 x3 x4 y@@;
logy = log(y);
label x1 = 'blood-clotting'
x2 = 'prognostic'
x3 = 'enzyme'
x4 = 'liver function'
y = 'survival';
cards;
6.7  62   81  2.59  200 5.1  59   66  1.70  101
7.4  57   83  2.16  204 6.5  73   41  2.01  101
7.8  65  115  4.30  509 5.8  38   72  1.42   80
5.7  46   63  1.91   80 3.7  68   81  2.57  127
6.0  67   93  2.50  202 3.7  76   94  2.40  203
6.3  84   83  4.13  329 6.7  51   43  1.86   65
5.8  96  114  3.95  830 5.8  83   88  3.95  330
7.7  62   67  3.40  168 7.4  74   68  2.40  217
6.0  85   28  2.98   87 3.7  51   41  1.55   34
7.3  68   74  3.56  215 5.6  57   87  3.02  172
5.2  52   76  2.85  109 3.4  83   53  1.12  136
6.7  26   68  2.10   70 5.8  67   86  3.40  220
6.3  59  100  2.95  276 5.8  61   73  3.50  144
5.2  52   86  2.45  181 1.2  76   90  5.59  574
5.2  54   56  2.71   72 5.8  76   59  2.58  178
3.2  64   65  0.74   71 8.7  45   23  2.52   58
5.0  59   73  3.50  116 5.8  72   93  3.30  295
5.4  58   70  2.64  115 5.3  51   99  2.60  184
2.6  74   86  2.05  118 4.3   8  119  2.85  120
4.8  61   76  2.45  151 5.4  52   88  1.81  148
5.2  49   72  1.84   95 3.6  28   99  1.30   75
8.8  86   88  6.40  483 6.5  56   77  2.85  153
3.4  77   93  1.48  191 6.5  40   84  3.00  123
4.5  73  106  3.05  311 4.8  86  101  4.10  398
5.1  67   77  2.86  158 3.9  82  103  4.55  310
6.6  77   46  1.95  124 6.4  85   40  1.21  125
6.4  59   85  2.33  198 8.8  78   72  3.20  313
;
run;
```

Example

- Scatter-plot matrix

```
proc sgscatter data=ch9tab01;
title "Scatterplot Matrix for Surgical Data";
matrix logy x1 x2 x3 x4;
run;
```

- Correlation matrix

```
proc corr data=ch9tab01;
var logy x1 x2 x3 x4;
run;
```

Example

- Let's use the full model

```
proc reg data=ch9tab01;
model logy = x1-x4/ selection=rsquare adjrsq cp mse sse aic bic;
run;  /*PRESSp is not a selection criterion in SAS*/
```

```
proc reg data=ch9tab01 OUTEST=RegOut PRESS;
model logy = x1 x2 x3 x4;
run;
```

```
proc print data=RegOut;
run;
```

"Best" Subsets Algorithms

With k regressors or predictors $X_1, \ldots, X_k$, the all subsets approach examines

1. All one-variable models

$$E(Y) = \beta_0 + \beta_1 X_1$$

$$\vdots$$

$$E(Y) = \beta_0 + \beta_1 X_k$$

2. All two-variable models

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\vdots$$

$$E(Y) = \beta_0 + \beta_1 X_{k-1} + \beta_2 X_k$$

$$\vdots$$

The number of possible models, $2^k - 1$, grows rapidly with the number of regressors. A variety of automatic computer-search procedures using the model selection criteria discussed in the last section have been developed.

Stepwise Regression Methods

In those occasional cases when the pool of potential X variables contains 30 to 40 or even more variables, use of a "best" subsets algorithm may not be feasible.

We consider one type of Stepwise method: **Forward Stepwise**, which consists of starting with no X variable in the model and adding the significant ones one at at time, until we are satisfied with the last model.

Each time when we try to add X_i to the regression model, we test

$$H_0 : \beta_i = 0 \text{ versus } H_A : \beta_i \neq 0.$$

at a **sle**(significance level for entry).

Furthermore, when there is already one or more X variable in the model, we test if all the X variables can stay at a **sls**(significance level for staying in the model) after a new X variable is added to the regression model. In general, we must specify sle and sls such that

$$sle \leq sls.$$

Stepwise Regression Methods

Suppose there are k regressors, $X_1, \ldots, X_k$.

Step 1. The **stepwise regression** routine first fits a simple linear regression model. Add one X variable, say X_i, in the regression model. Fit the regression model

$$E(Y) = \beta_0 + \beta_1 X_i$$

and test

$$H_0 : \beta_1 = 0 \text{ versus } H_A : \beta_1 \neq 0.$$

for $i = 1, \ldots, k$. The X_i variable with the largest F-test statistic (or smallest p-value),

$$F^* = \left[\frac{\widehat{\beta}_i}{s.e.(\widehat{\beta}_i)} \right]^2$$

is the candidate for first addition. If the p-value for the $F - test$ is less than the **sle**(significance level for entry), the X variable is added; Otherwise, the program terminates with no X variable considered sufficiently helpful to enter the regression model

Stepwise Regression Methods

Step 2. Assume X_1 is the variable entered at step 1. The stepwise regression routine now fits all regression models with two X variables. Add another X variable, say X_j, in the regression model. Fit the regression model

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_j$$

and test

$$H_0 : \beta_2 = 0 \text{ versus } H_A : \beta_2 \neq 0.$$

Choose the X_j with largest F test statistic (smallest p-value)

$$F^* = \left[\frac{\widehat{\beta}_j}{s.e.(\widehat{\beta}_j)} \right]^2.$$

If the p-value for the $F - test$ is less than the **sle**(significance level for entry), the X_j variable is added; Otherwise, the program terminates and the regression model include X_1 only.

Stepwise Regression Methods

Step 3. After the second step, suppose the model obtained is

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

Now the stepwise regression routine examines whether any of the other X variables already in the model should be dropped. We test

$$H_0 : \beta_1 = 0 \text{ versus } H_A : \beta_1 \neq 0$$

since $H_0 : \beta_2 = 0$ versus $H_A : \beta_2 \neq 0$. was tested in step 2. If the p-value is larger than the **sls**(significance level for staying in the model), X_1 should be dropped; otherwise, both X_1 and X_2 stay.

Step 4. Suppose both X_1 and X_2 are now in the model. The stepwise regression routine now examines which X variable is the next candidate for addition, then examines whether any of the variables already in the model should now be dropped, and so on until no further X variables can either be added or deleted, at which point the search terminates.

Note that the stepwise regression algorithm allows an X variable, brought into the model at an earlier stage, to be dropped subsequently if it is no longer helpful in conjunction with variables added at later stages.

Stepwise Regression Methods

Other Stepwise Procedures:

- **Forward Selection.** The forward selection search procedure is a simplified version of forward stepwise regression, omitting the test whether a variable once entered into the model should be dropped.
- **Backward Elimination.** The backward elimination search procedure is the opposite of forward selection. The backward elimination procedure consists of starting with all possible X variables in the model and eliminating the most non-significant ones one at at time, until we are satisfied with the remaining model.

Example

- Stepwise Regression Method

```
proc reg data = ch9tab01;
model logy = x1-x4/ selection = stepwise slentry= 0.01 slstay= 0.05;
run;
```

Unequal Error Variances

- The model assumption of `constant variance` may be invalid

If **transformation** is not helpful in reducing or eliminating unequal variances of the error terms, an alternative is weighted least squares, a procedure based on a generalization of multiple regression model.

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Unequal Error Variances

The generalized multiple regression model can then be expressed as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

where ε has a multivariate distribution with mean $\mathbf{0}$ and variance-covariance matrix

$$Var(\varepsilon) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}.$$

To obtain unbiased estimators of $\beta_0, \beta_1, \ldots, \beta_k$ with minimum variance, we must take into account that the different Y observations for the n cases no longer have the same reliability. Observations with small variances provide more reliable information about the regression function than those with large variances.

Weighted Least Squares

The weighted least squares estimates of β, denoted by $\widehat{\beta}_w = (\widehat{\beta}_{0w}, \widehat{\beta}_{1w}, \ldots, \widehat{\beta}_{kw})$ is obtained by minimizing

$$SSE_w = \sum_{i=1}^{n} w_i (y_i - \widehat{y}_i)^2,$$

where $w_i = 1/\sigma_i^2$. In matrix notation, the weighted least squares method minimizes

$$SSE_w = (\mathbf{y} - \mathbf{X}\widehat{\beta}_{\mathbf{w}})' \mathbf{W} (\mathbf{y} - \mathbf{X}\widehat{\beta}_{\mathbf{w}}),$$

where

$$\mathbf{W} = (Var(\varepsilon))^{-1} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_n^2 \end{bmatrix}.$$

Weighted Least Squares

Therefore, the weighted least squares normal equations are

$$\mathbf{X}^{'}\mathbf{W}\mathbf{X}\widehat{\beta}_{w} = \mathbf{X}^{'}\mathbf{W}\mathbf{Y}.$$

And the least squares estimators are:

$$\boxed{\widehat{\beta}_{w} = (\mathbf{X}^{'}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{W}\mathbf{Y}.}$$

The variance-covariance matrix of the weighted least squares estimated regression coefficients $\widehat{\beta}_{w}$ is

$$Var(\widehat{\beta}_{w}) = (\mathbf{X}^{'}\mathbf{W}\mathbf{X})^{-1}.$$

Weighted Least Squares - Error Variances Known up to Proportionality Constant

If the relative weights w_i are a constant multiple of the unknown true weights $1/\sigma_i^2$,

$$w_i = c\left(\frac{1}{\sigma_i^2}\right),$$

where c is the proportionality constant.

It can be shown that the weighted least squares estimators are unaffected. Furthermore,

$$Var(\widehat{\beta}_w) = c(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}.$$

This matrix is unknown because the proportionality constant c is not known. It can be estimated as

$$S^2(\widehat{\beta}_w) = MSE_w(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1},$$

where MSE_w is based on the weighted squared residuals:

$$MSE_w = \frac{\sum w_i(y_i - \widehat{y}_i)^2}{n-1-k} = \frac{\sum w_i e_i^2}{n-1-k}.$$

That is, MSE_w is an estimator of the proportionality constant c.

Weighted Least Squares - Error Variances Unknown

Since ε is a mean $\mathbf{0}$ random vector,

$$\sigma_i^2 = E(\varepsilon_i^2) - (E\{\varepsilon_i\})^2 = E(\varepsilon_i^2).$$

Hence, the squared residual e_i^2 is an estimator of σ_i^2. We can therefore estimate the variance function describing the relation of σ_i^2 to relevant predictor variables by first fitting the regression model using unweighted least squares and then regressing the squared residuals e_i^2 against the appropriate predictor variables.

Inference procedures when weights are estimated.

The variance-covariance matrix of the estimated regression coefficients is estimated by

$$S^2(\widehat{\beta}_w) = MSE_w(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1},$$

where MSE_w is based on the weighted squared residuals:

$$MSE_w = \frac{\sum w_i(y_i - \widehat{y}_i)^2}{n-1-k} = \frac{\sum w_i e_i^2}{n-1-k}.$$

Example

```
data ch11tab01;
input x y@@;
label x='age'
y='Dbp';
cards;
27 73 21 66 22 63 24 75
25 71 23 70 20 65 20 70
29 79 24 72 25 68 28 67
26 79 38 91 32 76 33 69
31 66 34 73 37 78 38 87
33 76 35 79 30 73 31 80
37 68 39 75 46 89 49 101
40 70 42 72 43 80 46 83
43 75 44 71 46 80 47 96
45 92 49 80 48 70 40 90
42 85 55 76 54 71 57 99
52 86 53 79 56 92 52 85
50 71 59 90 50 91 52 100
58 80 57 109
;
run;
```

Example

```
proc reg data=ch11tab01;
model y = x;
output out=temp r=resid;
plot y*x r.*x;
run;
```

- The nonconstant error variance can be seen from the residual plot.

Example

- Now let's see the absolute residuals plot.

```
data temp;
set temp;
absr = abs(resid);
run;
```

```
proc gplot data = temp;
plot absr*x;
run;
```

- Or sgplot

```
proc sgplot data = temp;
scatter x=x y=absr;
run;
```

Example

- It seems that there is a linear relation between the error standard deviation and X. We therefore regressed the absolute residuals against X and record the predicted values which are the estimated expected standard deviation.

```
proc reg data = temp;
model absr = x;
output out = temp1 p = s;
run;
```

Example

- The fitted model is
$$\hat{s} = -1.54948 + 0.19817x.$$
- The weights are then obtained by using $w_i = \frac{1}{(\hat{s}_i)^2}$

```
data temp1;
set temp1;
w = 1/(s**2);
run;
```

```
proc print data = temp1 (obs = 10);
run;
```

Example

- Last, we fit the regression model using weighted least squares.

```
proc reg data = temp1;
weight w;
model y = x / clb;
run;
```

- It is interesting to note that the standard deviation is somewhat smaller than the standard deviation of the estimate obtained by ordinary least squares method.

License