Contents

1 Law of Parsimony

Parsimonious: The simplest plausible model with the fewest possible number of variables.

2 Partial Correlation in Linear Models

Partial correlation is a measure of the strength and direction of a linear relationship between two continuous variables whilst controlling for the effect of one or more other continuous variables (also known as 'covariates' or 'control' variables). Although partial correlation does not make the distinction between independent and dependent variables, the two variables are often considered in such a manner (i.e., you have one continuous dependent variable and one continuous independent variable, as well as one or more continuous control variables).

3 Variable Selection Procedures

Often we require the optimal set of predictor variables to adequately describe the data, without overfitting. As such we will use variable selection procedures. In statistical methods, the order in which the predictor variables are entered into (or taken out of) the model is determined according to the strength of their correlation with the response variable. Actually there are several versions of this method, called forward selection, backward selection and stepwise selection.

The main procedures are as follows:

- Forward Selection
- Backward Elimination
- Stepwise Selection

The essential concept is the estimation of the relationship between a predictor variable and a response variable after controlling for the effects of other predictors in the equation. One such estimate is the semi-partial correlation coefficient.

3.1 Model Selection

Model selection, also known as variable selection or feature selection, is the task of selecting a statistical model from a set of potential models, given the data.

3.2 Stepwise Regression

- Stepwise regression combines forward selection and backward elimination. At each step, the best remaining variable is added, provided it passes the significant at 5% criterion, then all variables currently in the regression are checked to see if any can be removed, using the greater than 10% significance criterion.

- It is not guaranteed to find the best subset of independents but it will find a subset close to the best.The process continues until no more variables are added or removed. This is the one we shall use.

3.3 Variable Selection Procedures in SPSS

- **Enter**: This is the forced entry option. SPSS will enter at one time all specified variables regardless of significance levels.
- **Forward**: This method will enter variables one at a time, based on the significance value to enter.
- **Backward**: This enters all independent variables at one time and then removes variables one at a time based on a preset significance value to remove.
- **Stepwise**: This combines both forward and backward procedures. Since inter correlations are complex, the variance due to certain variables will change when new variables are entered into the equation. This is the most frequently used of the regression methods.
- **Remove**: This is the forced removal option. It requires an initial regression analysis usingthe Enter procedure. In the next block (Block 1 of 1) you may specify one or morevariables to remove. SPSS will then remove the specified variables and run the analysis again.

There are different ways that the relative contribution of each predictor variable can be assessed. In the "simultaneous" method (which SPSS calls the **Enter** method), the researcher specifies the set of predictor variables that make up the model. The success of this model in predicting the criterion variable is then assessed.

In contrast, ***hierarchical*** methods enter the variables into the model in a specified order. The order specified should reflect some theoretical consideration or previous findings. If you have no reason to believe that one variable is likely to be more important than another you should not use this method. As each variable is entered into the model its contribution is assessed. If adding the variable does not significantly increase the predictive power of the model then the variable is dropped.

In ***statistical*** methods, the order in which the predictor variables are entered into (or taken out of) the model is determined according to the strength of their correlation with the criterion variable. Actually there are several versions of this method, called forward selection, backward selection and stepwise selection.

4 Variable Selection Procedures (SPSS Options)

- **Enter:** This is the forced entry option. SPSS will enter at one time all specified variables regardless of significance levels.
- **Forward:** This method will enter variables one at a time, based on the significance value to enter.
- **Backward:** This enters all independent variables at one time and then removes variables one at a time based on a preset significance value to remove.

- **Stepwise:** This combines both forward and backward procedures. Since inter correlations are complex, the variance due to certain variables will change when new variables are entered into the equation. This is the most frequently used of the regression methods.
- **Remove:** This is the forced removal option. It requires an initial regression analysis usingthe Enter procedure. In the next block (Block 1 of 1) you may specify one or morevariables to remove. SPSS will then remove the specified variables and run the analysis again.

There are different ways that the relative contribution of each predictor variable can be assessed. In the simultaneous method (which SPSS calls the **Enter** method), the researcher specifies the set of predictor variables that make up the model. The success of this model in predicting the criterion variable is then assessed.
In contrast, hierarchical methods enter the variables into the model in a specified order. The order specified should reflect some theoretical consideration or previous findings. If you have no reason to believe that one variable is likely to be more important than another you should not use this method. As each variable is entered into the model its contribution is assessed. If adding the variable does not significantly increase the predictive power of the model then the variable is dropped.

4.1 Forward Selection with SPSS

In Forward selection, SPSS enters the variables into the model one at a time in an order determined by the strength of their correlation with the criterion variable. The effect of adding each is assessed as it is entered, and variables that do not significantly add to the success of the model are excluded.
Step 1
Firstly, the predictor variable with the largest squared correlation with the dependent variable Y is entered into the model. Since this is the first step of the selection procedure, entering the predictor with the largest squared correlation is equivalent to entering the predictor with the **largest squared semi-partial** correlation.
It may seem trivial to bring up the idea of semi-partial correlation at step 1 of the procedure, but we do so because at subsequent steps, the criterion for entrance into the regression equation will be the squared semi-partial correlation (or equivalently, the amount of variance contributed by the new predictor over and above variables already entered into the equation).
Step 2
The unselected predictor variable with the largest squared semi-partial correlation with the dependent variable (hence referred to as Y) is selected. That is, the predictor with the largest correlation with Y after being adjusted for the first predictor, is entered if it meets entrance criteria in terms of preset statistical significance for entry, what SPSS refers to as `PIN` (probability of entry) criteria. It is important to note that even once this new predictor is entered at step 2, the predictor entered at step 1 remains in the equation, even if it's new semi-partial correlation with Y is now less than what it was at step 1.
This is the nature of the forward selection procedure, it does not re-evaluate already-entered predictors into the model after adding new variables. That is, it only add predictors to the model.

Step 3
The next unselected predictor with the largest squared semi-partial correlation with Y is then selected. That is, the predictor with the largest correlation with Y after being adjusted for both of the first two predictors is entered.
Selection for entrance of this variable is conditional upon its relationship with the previously entered variables at step 1 and step 2.

Hence, for a variable to be entered at step 3, SPSS asks the question, *"Which among available variables currently not entered into the regression equation contribute most to variance explained in Y given that variables entered at steps 1 and 2 remain in the model?"* Translated into statistical language, what this question boils down to is selecting the variable that has the largest statistically significant squared semi-partial correlation with Y.

Subsequent Steps
We do not detail subsequent steps for the reason that they mimic the preceding steps. It is worth noting that we didn't even really need to detail steps 2 and 3, and could have just stated the "rule" of forward regression by referring to the first step alone.

Summary

- We can state the general rule of forward regression as follows:

  > Forward regression, at each step of the selection procedure from step 1 through subsequent steps, chooses the predictor variable with the greatest squared semi-partial correlation with the response variable for entry into the regression equation. The given predictor will be entered if it satisfies entrance criteria (significance level, PIN) specified in advance by the researcher.

- The above is the simplest way to describe the procedural routine of how forward regression operates. What is perhaps most noteworthy about the above rule is what is not included just as much as what is included in the statement. Notice that nowhere in the rule is there any mention of removal of predictors at any step of the selection process.

- In this procedure, once a predictor is selected into the model, it cannot be removed. Other predictors may be added at future steps, but predictors already in the model remain in the model. As we will see, this is in contrast to SPSS's stepwise regression, in which we can specify criteria for both adding and removing predictors at each step.

4.2 Backward Selection with SPSS

In Backward selection, SPSS enters all the predictor variables into the model. The weakest predictor variable is then removed and the regression re-calculated. If this significantly weakens the model then the predictor variable is re-entered otherwise it is deleted. This procedure is then repeated until only useful predictor variables remain in the model.

4.3 Stepwise Selection with SPSS

Stepwise is the most sophisticated of these statistical methods. Each variable is entered in sequence and its value assessed. If adding the variable contributes to the model then it is retained, but all other variables in the model are then re-tested to see if they are still contributing

to the success of the model. If they no longer contribute significantly they are removed. Thus, this method should ensure that you end up with the smallest possible set of predictor variables included in your model.

4.4 Stepping Method Criteria

Stepwise methods include or remove one independent variable at each step, based (by default) on the probability of F (p-value). The limits for the criteria controlling variable inclusion or removal can be specified by defining probabilities for **F-to-enter/F-to-remove** (or otherwise **values of F-to-enter/F-to-remove**, not recommended without a very thorough understanding of the F-distribution).

The following three stepwise methods are available.

- Stepwise Based on the p-value of F (probability of F), SPSS starts by entering the variable with the smallest p-value; at the next step again the variable (from the list of variables not yet in the equation) with the smallest p-value for F and so on.

  Variables already in the equation are removed if their p-value becomes larger than the default limit due to the inclusion of another variable. The method terminates when no more variables are eligible for inclusion or removal.

  This methods is based on both probability-to-enter (PIN) and probability to remove (POUT).

- Backward Elimination: First all variables are entered into the equation and then sequentially removed. For each step SPSS provides statistics, namely R^2. At each step, the largest probability of F is removed (if the value is larger than POUT.

- Forward Forward selection: at each step the variable not yet in the equation with the smallest probability of F is entered. as long as the value is smaller than PIN. The procedure stops when there are no variables that meet the entry criterion.

4.5 The Remove Option (SPSS Option)

In addition to the Enter, Stepwise, Forward and Backward methods, SPSS also offers the Remove method in which variables are removed from the model in a block - the use of this method will not be described here.

- If you have no theoretical model in mind, and/or you have relatively low numbers of cases, then it is probably safest to use Enter, the simultaneous method. Statistical procedures should be used with caution and only when you have a large number of cases.

- This is because minor variations in the data due to sampling errors can have a large effect on the order in which variables are entered and therefore the likelihood of them being retained. However, one advantage of the Stepwise method is that it should always result in the most parsimonious model.

- This could be important if you wanted to know the minimum number of variables you would need to measure to predict the dependent variable. If for this, or some other reason, you decide to select a statistical method, then you should really attempt to validate your results with a second independent set of data.

- This can be done either by conducting a second study, or by randomly splitting your data set into two halves . Only results that are common to both analyses should be reported.

4.6 F Values

- At each step, SPSS performs the following calculations: for each variable currently in the model, it computes the t-statistic for its estimated coefficient, squares it, and reports this as its **F-to-remove** statistic; for each variable not in the model, it computes the t-statistic that its coefficient would have if it were the next variable added, squares it, and reports this as its **F-to-enter** statistic.

- At the next step, the program automatically enters the variable with the highest F-to-enter statistic, or removes the variable with the lowest F-to-remove statistic, in accordance with certain specified values.

- (Important: F = t-squared)

5 Variable Selection Procedures

6 Stepwise Regression using R

SPSS can be very opaque in determining how particularly statistical routines are carried out. Conversely the statistical programming language R is usually quite clear, once a familiarity with the language has been developed.

For variable selection procedures, R used the AIC criterion. When comparing multiple candidate models, the candidate model with the lowest AIC value is the best model. We will use R output to revise variable selection procedures. Recall that we used the ***mtcars*** data set. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (197374 models). For this data, we tried to determine the optimal set of independent variables to predict the dependent variables ***mpg*** (miles per gallon).

cyl Number of cylinders

disp Displacement (cu.in.)

hp Gross horsepower

drat Rear axle ratio

wt Weight (lb/1000)

qsec 1/4 mile time

vs V/S

am Transmission (0 = automatic, 1 = manual)

gear Number of forward gears

carb Number of carburetors

6.1 Backward Elimination

The initial model contains all of the independent variables. Candidate models, whereby each of the independent variables are individually removed from the model are fitted. The AIC value for each reduced model is computed. The unreduced model is also used for comparison. The AIC values are tabulated to determine which removal results in the lowest AIC value. In this first case, the removal of cyl would reduced the AIC value from 70.898 (see bottom row) to 68.915. Thus the independent variable cyl is removed from the set of IVs.

```
Start:  AIC=70.9
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

Df Sum of Sq    RSS    AIC
```

```
- cyl   1    0.0799 147.57 68.915
- vs    1    0.1601 147.66 68.932
- carb  1    0.4067 147.90 68.986
- gear  1    1.3531 148.85 69.190
- drat  1    1.6270 149.12 69.249
- disp  1    3.9167 151.41 69.736
- hp    1    6.8399 154.33 70.348
- qsec  1    8.8641 156.36 70.765
<none>              147.49 70.898
- am    1   10.5467 158.04 71.108
- wt    1   27.0144 174.51 74.280
```

In the second phase, the process is repeated. This time removing vs results in an AIC value of 66.973. It is then removed from the set of IVs. For this phase, the unreduced model is the model fitted by all independent variables except cyl, which was removed in the previous phase.

```
Step:  AIC=68.92
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

Df Sum of Sq    RSS    AIC
- vs    1    0.2685 147.84 66.973
- carb  1    0.5201 148.09 67.028
- gear  1    1.8211 149.40 67.308
- drat  1    1.9826 149.56 67.342
- disp  1    3.9009 151.47 67.750
- hp    1    7.3632 154.94 68.473
<none>              147.57 68.915
- qsec  1   10.0933 157.67 69.032
- am    1   11.8359 159.41 69.384
- wt    1   27.0280 174.60 72.297
```

this process continues until the removal of an IV will not results in an improvement in AIC. This is indicated by having the $< none >$ (i.e unreduced model) having the lowest AIC value. At the end of the output is the optimal model, according to the backward elimination procedure, using the IVs : am , qsec and wt.

```
Step:  AIC=61.31
mpg ~ wt + qsec + am
```

```
Df Sum of Sq    RSS    AIC
<none>              169.29 61.307
- am    1    26.178 195.46 63.908
- qsec  1   109.034 278.32 75.217
- wt    1   183.347 352.63 82.790

Call:
lm(formula = mpg ~ wt + qsec + am)

Coefficients:
(Intercept)           wt         qsec           am
9.618       -3.917        1.226        2.936
```

6.2 Stepwise Regression

Stepwise Regression differs from Backward Elimination, in that it allows IVs to be re-introduced. Hence the + signs from the second phase onwards.

```
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

Df Sum of Sq    RSS    AIC
- cyl   1    0.0799 147.57 68.915
- vs    1    0.1601 147.66 68.932
- carb  1    0.4067 147.90 68.986
- gear  1    1.3531 148.85 69.190
- drat  1    1.6270 149.12 69.249
- disp  1    3.9167 151.41 69.736
- hp    1    6.8399 154.33 70.348
- qsec  1    8.8641 156.36 70.765
<none>              147.49 70.898
- am    1   10.5467 158.04 71.108
- wt    1   27.0144 174.51 74.280

Step:  AIC=68.92
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

Df Sum of Sq    RSS    AIC
- vs    1    0.2685 147.84 66.973
- carb  1    0.5201 148.09 67.028
- gear  1    1.8211 149.40 67.308
- drat  1    1.9826 149.56 67.342
- disp  1    3.9009 151.47 67.750
```

```
- hp    1    7.3632 154.94 68.473
<none>              147.57 68.915
- qsec  1   10.0933 157.67 69.032
- am    1   11.8359 159.41 69.384
+ cyl   1    0.0799 147.49 70.898
- wt    1   27.0280 174.60 72.297
```

Again, the procedure finishes when it is found that the unchanged model has the lowest of all possible AIC values.

```
Step:  AIC=61.31
mpg ~ wt + qsec + am

Df Sum of Sq    RSS    AIC
<none>              169.29 61.307
+ hp    1     9.219 160.07 61.515
+ carb  1     8.036 161.25 61.751
+ disp  1     3.276 166.01 62.682
+ cyl   1     1.501 167.78 63.022
+ drat  1     1.400 167.89 63.042
+ gear  1     0.123 169.16 63.284
+ vs    1     0.000 169.29 63.307
- am    1    26.178 195.46 63.908
- qsec  1   109.034 278.32 75.217
- wt    1   183.347 352.63 82.790

Call:
lm(formula = mpg ~ wt + qsec + am)

Coefficients:
(Intercept)           wt         qsec           am
9.618       -3.917        1.226        2.936
```

7 Stepwise Logistic Selection

Stepwise logistic regression involves the stepwise (or one-by-one) selection of variables, providing a fast and effective method to screen a large number of variables, and to fit multiple logistic regression equations simultaneously.

In stepwise selection, an attempt is made to remove any insignificant variables from the model before adding a significant variable to the model.

Stepwise binary logistic regression is very similar to stepwise multiple regression in terms of its advantages and disadvantages. Stepwise logistic regression is designed to find the ***most parsimonious*** set of predictors that are most effective in predicting the dependent variable.

7.1 Procedure for Stepwise Selection

- Variables are added to the logistic regression equation one at a time, using the statistical criterion of reducing the ***-2 Log Likelihood error*** for the included variables. (Recall: The lower the -2LL value, the better the fit of the model).
- After each variable is entered, each of the included variables are tested to see if the model would be better off the variable were excluded. This does not happen often.
- The process of adding more variables stops when all of the available variables have been included or when it is not possible to make a statistically significant reduction in -2 Log Likelihood using any of the variables not yet included.
- Categorical variables are added to the logistic regression as a group. It is possible, and often likely, that not all of the individual dummy-coded variables will have a statistically significant individual relationship with the dependent variable.

7.2 SPSS Implementation

SPSS provides a table of variables included in the analysis and a table of variables excluded from the analysis. It is possible that none of the variables will be included. It is possible that all of the variables will be included.

The order of entry of the variables can be used as a measure of relative importance.

Once a variable is included, its interpretation in stepwise logistic regression is the same as it would be using other methods for including variables.

7.3 Advantages and Disadvantages

- Stepwise logistic regression can be used when the goal is to produce a predictive model that is parsimonious and accurate because it excludes variables that do not contribute to explaining differences in the dependent variable.
- Stepwise logistic regression is less useful for testing hypotheses about statistical relationships. Its usage is recommended only for exploratory purposes, rather that as a formal procedure.
- Stepwise logistic regression can be useful in finding relationships that have not been tested before. Its findings invite one to speculate on why an unusual relationship makes sense.
- It is not legitimate to do a stepwise logistic regression and present the results as though one were testing a hypothesis that included the variables found to be significant in the stepwise logistic regression.
- Using statistical criteria to determine relationships is vulnerable to over-fitting the data set used to develop the model at the expense of generalisability.

> Menard (1995: 54) writes, "there appears to be general agreement that the use of computer-controlled stepwise procedures to select variables is inappropriate for theory testing because it capitalizes on random variations in the data and produces results that tend to be idosyncratic and difficult to replicate in any sample other than the sample in which they were originally obtained."

7.4 Forward Selection

You can estimate models using block entry of variables or any of the following stepwise methods: forward conditional, forward LR, forward Wald, backward conditional, backward LR, or backward Wald.

Forward selection is the usual option for a stepwise regression, starting with the constant-only model and adding variables one at a time. The forward stepwise logistic regression method utilizes the likelihood ratio test which tests the change in 2LL between steps to determine automatically which variables to add or drop from the model.

Method selection allows you to specify how independent variables are entered into the analysis. Using different methods, you can construct a variety of regression models from the same set of variables.

1 Enter. A procedure for variable selection in which all variables in a block are entered in a single step.

2 Forward Selection (Conditional). Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on conditional parameter estimates.

3 Forward Selection (Likelihood Ratio). Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on the maximum partial likelihood estimates. (LR stands for Likelihood Ratio and is considered the criterion least prone to error.)

4 Forward Selection (Wald). Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of the Wald statistic.

5 Backward Elimination (Conditional). Backward stepwise selection. Removal testing is based on the probability of the likelihood-ratio statistic based on conditional parameter estimates.

6 Backward Elimination (Likelihood Ratio). Backward stepwise selection. Removal testing is based on the probability of the likelihood-ratio statistic based on the maximum partial likelihood estimates.

7 Backward Elimination (Wald). Backward stepwise selection. Removal testing is based on the probability of the Wald statistic.

7.5 Cross Validation of Stepwise Regression

When stepwise logistic regression is used, some form of validation analysis is a necessity. We will use 75/25% cross-validation.

To do cross validation, we randomly split the data set into a 75% training sample and a 25% validation sample. We will use the training sample to develop the model, and we test its effectiveness on the validation sample to test the applicability of the model to cases not used to develop it.

In order to be successful, the follow two questions must be answers affirmatively: Did the stepwise logistic regression of the training sample produce the same subset of predictors produced by the regression model of the full data set?

If yes, compare the classification accuracy rate for the 25% validation sample to the classification accuracy rate for the 75% training sample. If the **shrinkage** (accuracy for the 75% training sample - accuracy for the 25% validation sample) is 2% (0.02) or less, we conclude that validation was successful.

Note: shrinkage may be a negative value, indicating that the accuracy rate for the validation sample is larger than the accuracy rate for the training sample. Negative shrinkage (increase in accuracy) is evidence of a successful validation analysis.

If the validation is successful, we base our interpretation on the model that included all cases.