# CHAPTER 12: SOME VARIABLE SELECTION PROCEDURES

Researchers sometimes do not have a specific model that they wish to test. In other words, given the data that has been collected on the system the researcher would like to find the best model. There are a number of methods called ***variable selection procedures*** that can be used to accomplish this task. We will examine some of the methods available on the SPSS system.

Consider the example in Section 11.2 where we recorded seriousness, age, TV news, and victimization (note the data is now for 20 people). Suppose the psychologist wished to consider the following predictor variables of seriousness ($Y$).

| Predictors | SPSS variables |
|---|---|
| age | age |
| age squared | agesq |
| TV news | tvnews |
| Victimization | victim |
| interaction of age and TV news | compute atv=age*tvnews |
| interaction of age and Victimization | compute avic=age*victim |
| interaction of age, TV news and Victimization | compute atvic=age*tvnews*victim |

The question the researcher wishes to answer is: Of all of the above predictors, which set gives the best model? There are a number of ways this could be answered.

## 12.1 FORWARD SELECTION METHOD

This method enters the predictors one at a time until a model is found, such that entering any more variables would not improve the model significantly. This is done using $r$ or $R^2$ in the simplest model as the criterion for strongest predictor of $y$. The computer selects the $x$ variable that is most highly correlated with $y$ first. After removing the effect of this predictor from $y$, the next strongest predictor, given the first is selected, is added to the model giving a two variable linear model, and so on. Notice once variables are entered into the model using the forward procedures they cannot be dropped from the model.

SPSS first selects the strongest predictor (the predictor with the smallest $p$-value in the $T$ column or *pin* value), then the next strongest is found given the first. The procedure can be run using syntax (given below) or using the windows method by clicking the arrow where the enter command is located and selecting forward.

**Commands (syntax)**

```
regression variables = serious age agesq tvnews victim atv avic atvic/
        criteria = pin(.05)/
        dependent = serious/
        forward
```

For the above example, the *p*-value for entry is .05. The forward print out below gives the best model as

$$E(x|y) = 13.49 + .09(ATV)$$

The interaction of age and tvnews was the highest correlated with seriousness ($y$) and entered first. Since none of the variables appear in the *Variables Entered/Removed* section, with the *p*-value is less than .05, the procedure stops at this point.

**Variables Entered/Removed[a]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | Interaction of age and TV news | . | Forward (Criterion: Probability-of-F-to-enter <= .050) |

a. Dependent Variable: Seriousness

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .898[a] | .806 | .795 | 7.302 |

a. Predictors: (Constant), Interaction of age and TV news

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 3982.721 | 1 | 3982.721 | 74.689 | .000[a] |
| | Residual | 959.829 | 18 | 53.324 | | |
| | Total | 4942.550 | 19 | | | |

a. Predictors: (Constant), Interaction of age and TV news

b. Dependent Variable: Serious

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | 13.493 | 3.187 | | 4.233 | .000 |
| | Interaction of age and TV news | 9.946E-02 | .012 | .898 | 8.642 | .000 |

a. Dependent Variable: Serious

## 12.2 BACKWARD ELIMINATION METHOD

The *Backward* method enters all predictors into a model. It then removes any variables that are not making a significant contribution to the prediction equation using the *t* test. One can also designate the level of significance desired to remove such variables. The method continues to drop variables until no other variables can be removed, and those that remain contribute significantly to the model.

SPSS in our example would first fit the model

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_1 x_3 + \beta_6 x_1 x_2 + \beta_7 x_1 x_2 x_3$$

Next, the weakest predictor (i.e., has the largest *p*-value) would be dropped. Given this model, the next weakest predictor is dropped, and so on. The criterion for dropping is the *p*-value in the *t* column to remove or *pout*. In the commands below, *p*-values of *t* greater than 0.1 are dropped.

**Commands**

```
regression variables = serious age agesq tvnews victim atv avic atvic/
        criteria = pout(0.1)/
        dependent = serious/
        backward
```

The best model using this procedure is

$$E(y|x) = -32.74 + 6.48(TVNEWS) + 32.51(VICTIM) + .82(AGE) - .88(AVIC)$$

The first step is to fit all variables into the model. Note that the *p*-value of the variable that is highest is dropped from the model in the next step. If we examine the *Variables Entered/Removed* section, *t* column, this is the variable ATVIC.

**Variables Entered/Removed[b]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | Interaction of age, TV news and victimization, Age squared, Amount of TV news watched (hrs/wk), Previous victim of crime, Age, Interaction of age and victimization, Interaction of age and TV news[a] | . | Enter |
| 2 | . | Interaction of age, TV news and victimization | Backward (criterion: Probability of F-to-remove >= .100). |
| 3 | . | Interaction of age and TV news | Backward (criterion: Probability of F-to-remove >= .100). |
| 4 | . | Age squared | Backward (criterion: Probability of F-to-remove >= .100). |

a. All requested variables entered.

b. Dependent Variable: Seriousness

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .946[a] | .895 | .833 | 6.583 |
| 2 | .946[b] | .895 | .846 | 6.328 |
| 3 | .946[c] | .895 | .857 | 6.101 |
| 4 | .945[d] | .894 | .865 | 5.919 |

a. Predictors: (Constant), Interaction of age, TV news and victimization, Age squared, Amount of TV news watched (hrs/wk), Previous victim of crime, Age, Interaction of age and victimization, Interaction of age and TV news

b. Predictors: (Constant), Age squared, Amount of TV news watched (hrs/wk), Previous victim of crime, Age, Interaction of age and victimization, Interaction of age and TV news

c. Predictors: (Constant), Age squared, Amount of TV news watched (hrs/wk), Previous victim of crime, Age, Interaction of age and victimization

d. Predictors: (Constant), Amount of TV news watched (hrs/wk), Previous victim of crime, Age, Interaction of age and victimization

**ANOVA[e]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 4422.464 | 7 | 631.781 | 14.577 | .000[a] |
| | Residual | 520.086 | 12 | 43.341 | | |
| | Total | 4942.550 | 19 | | | |
| 2 | Regression | 4422.008 | 6 | 737.001 | 18.406 | .000[b] |
| | Residual | 520.542 | 13 | 40.042 | | |
| | Total | 4942.550 | 19 | | | |
| 3 | Regression | 4421.521 | 5 | 884.304 | 23.761 | .000[c] |
| | Residual | 521.029 | 14 | 37.216 | | |
| | Total | 4942.550 | 19 | | | |
| 4 | Regression | 4417.002 | 4 | 1104.251 | 31.517 | .000[d] |
| | Residual | 525.548 | 15 | 35.037 | | |
| | Total | 4942.550 | 19 | | | |

a. Predictors: (Constant), Interaction of age, TV news and victimization, Age squared, Amount of TV news watched (hrs/wk), Previous victim of crime, Age, Interaction of age and victimization, Interaction of age and TV news

b. Predictors: (Constant), Age squared, Amount of TV news watched (hrs/wk), Previous victim of crime, Age, Interaction of age and victimization, Interaction of age and TV news

c. Predictors: (Constant), Age squared, Amount of TV news watched (hrs/wk), Previous victim of crime, Age, Interaction of age and victimization

d. Predictors: (Constant), Amount of TV news watched (hrs/wk), Previous victim of crime, Age, Interaction of age and victimization

e. Dependent Variable: Seriousness

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | -36.657 | 68.621 | | -.534 | .603 |
| | Age | .561 | 1.381 | .481 | .406 | .692 |
| | Age squared | 9.361E-03 | .045 | .788 | .207 | .840 |
| | Amount of TV news watched (hrs/wk) | 9.900 | 20.813 | .770 | .476 | .643 |
| | Previous victim of crime | 32.402 | 28.496 | 1.025 | 1.137 | .278 |
| | Interaction of age and TV news | -9.01E-02 | .615 | -.813 | -.146 | .886 |
| | Interaction of age and victimization | -.993 | 1.914 | -1.266 | -.519 | .613 |
| | Interaction of age, TV news and victimization | 1.956E-02 | .191 | .173 | .103 | .920 |
| 2 | (Constant) | -32.401 | 52.533 | | -.617 | .548 |
| | Age | .479 | 1.084 | .411 | .442 | .666 |
| | Age squared | 6.948E-03 | .037 | .585 | .187 | .854 |
| | Amount of TV news watched (hrs/wk) | 8.813 | 17.220 | .686 | .512 | .617 |
| | Previous victim of crime | 29.775 | 12.017 | .942 | 2.478 | .028 |
| | Interaction of age and TV news | -5.19E-02 | .470 | -.468 | -.110 | .914 |
| | Interaction of age and victimization | -.800 | .328 | -1.019 | -2.440 | .030 |
| 3 | (Constant) | -27.038 | 19.118 | | -1.414 | .179 |
| | Age | .474 | 1.044 | .406 | .454 | .657 |
| | Age squared | 2.972E-03 | .009 | .250 | .348 | .733 |
| | Amount of TV news watched (hrs/wk) | 6.943 | 2.845 | .540 | 2.440 | .029 |
| | Previous victim of crime | 29.723 | 11.576 | .941 | 2.568 | .022 |
| | Interaction of age and victimization | -.801 | .316 | -1.021 | -2.536 | .024 |
| 4 | (Constant) | -32.746 | 9.565 | | -3.424 | .004 |
| | Age | .829 | .222 | .711 | 3.736 | .002 |
| | Amount of TV news watched (hrs/wk) | 6.483 | 2.447 | .505 | 2.650 | .018 |
| | Previous victim of crime | 32.520 | 8.092 | 1.029 | 4.019 | .001 |
| | Interaction of age and victimization | -.883 | .205 | -1.125 | -4.316 | .001 |

a. Dependent Variable: Seriousness

## 12.3 STEPWISE SELECTION

The *Stepwise* method is similar to the *Forward* method; however, at each stage all variables entered are examined to see whether they are necessary after entering the last variable. In some cases, a variable entered earlier may be dropped.

For example, $x_1$ and $x_2$ are in the model; however, when $x_3$ is added, the *t* statistic associated with $x_2$ is no longer significant, and is therefore dropped.

**Commands**

```
regression variables = serious age agesq tvnews victim atv avic atvic/
        criteria=pin(.05) pout(0.1)/
        dependent = serious/
        stepwise
```

The model selected using this procedure is

$$E(y|x) = 13.49 + .09(ATV)$$

The stepwise output in this situation gives the same model as the forward procedure.

**Variables Entered/Removed[a]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | Interaction of age and TV news | . | Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100). |

a. Dependent Variable: Seriousness

**Excluded Variables[b]**

| Model | | Beta In | t | Sig. | Partial Correlation | Collinearity Statistics: Tolerance |
|---|---|---|---|---|---|---|
| 1 | Age | -1.245[a] | -1.966 | .066 | -.430 | 2.323E-02 |
| | Age squared | -.147[a] | -.242 | .812 | -.059 | 3.065E-02 |
| | Amount of TV news watched (hrs/wk) | .089[a] | .334 | .743 | .081 | .160 |
| | Previous victim of crime | -.003[a] | -.026 | .980 | -.006 | .992 |
| | Interaction of age and victimization | -.097[a] | -.917 | .372 | -.217 | .974 |
| | Interaction of age, TV news and victimization | -.132[a] | -1.243 | .231 | -.289 | .932 |

a. Predictors in the Model: (Constant), Interaction of age and TV news

b. Dependent Variable: Seriousness

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | 13.493 | 3.187 | | 4.233 | .000 |
| | Interaction of age and TV news | 9.946E-02 | .012 | .898 | 8.642 | .000 |

a. Dependent Variable: Seriousness

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 3982.721 | 1 | 3982.721 | 74.689 | .000[a] |
| | Residual | 959.829 | 18 | 53.324 | | |
| | Total | 4942.550 | 19 | | | |

a. Predictors: (Constant), Interaction of age and TV news

b. Dependent Variable: Seriousness

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .898[a] | .806 | .795 | 7.302 |

a. Predictors: (Constant), Interaction of age and TV news

Note that the different procedures may give the psychologist different models.

## 12.4 MULTICOLLINEARITY

Suppose we have a single dependent variable *y* and *k* independent variables. Further, let $y = f(x_1, x_2, ..., x_k)$ denote the functional relationship between $(x_1, x_2, ..., x_k)$ and *y*. Note that we are in a different situation than the one dependent variable and one independent or predictor variable case. When the predictor variables or *x*'s, $(x_1, x_2, ..., x_k)$ are highly correlated, we say *multicollinearity* exists. When the correlation among the variables is very high (say, .9 or more) problems may arise with the model estimates, their interpretation, and the significance level of tests. Other consequences of multicollinearity are large standard errors, which give rise to wide confidence intervals and non-significant or incorrect *t* statistics.

We assume hereafter that the relationship between the dependent variable $y$ and the independent variables $x_1,\ldots, x_k$ (perhaps re-expressions of the original independent variables) is of the form, ignoring for the moment the possibility of variation,

$$\begin{aligned} y &= f(x_1,\ldots,x_k) \\ &= \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \\ &= X'_{1xk}\,\beta_{kx1} \\ &= X'\beta \end{aligned}$$

In the statistical context, i.e. when a particular value for $(x_1,\ldots,x_k)'$ specifies a frequency distribution for $y$, we ***assume*** the average value of $y$ is given by

$$\mathbf{E[}y|x_1,\ldots,x_k\mathbf{] = \beta_1 x_1 + \ldots + \beta_k x_k}$$

and that changes in $(x_1,\ldots,x_k)$ affect, at most, the ***means*** of the frequency distributions. Read $E[y|x_1,\ldots,x_k]$ as the average value of $y$ given $(x_1,\ldots,x_k)$. If we put $e = y - \beta_1 x_1 - \ldots - \beta_k x_k$ then the frequency distribution of $e$ is constant as $(x_1,\ldots,x_k)$ changes.

Thus, we can write our model as

$$y = \beta_1 x_1 + \ldots + \beta_k x_k + e$$

and $e$ is referred to as the ***error term***.

If $f$ is the frequency of $e$, then for a particular value of $(x_1,\ldots,x_k)$ the frequency function of $y$ is given by $f(e - \beta_1 x_1 - \ldots - \beta_k x_k)$. We will ***assume*** hereafter that $f$ can be taken to be a density function, and the variance of the frequency distribution for $e$ exists and is equal to $\sigma^2$.

***In a psychological investigation our primary purpose will be to make inferences about the true value of the coefficients $\beta_1, \beta_2, \ldots, \beta_k$.***

To do this, we will be required to make a number of observations at different values of $(x_1,\ldots,x_k)'$.

Let $y_i$ denote the observation taken at

$$X'_{(i)} = (x_{i1},\ldots,x_{ik})$$

and let $e_i$ denote the error. Then, for $n$ observations we have in matrix notation

$$y_{nx1} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_1 x_{11} + & \ldots & + \beta_k x_{1k} + & e_1 \\ \beta_1 x_{21} + & \ldots & + \beta_k x_{2k} + & e_2 \\ . & & . & . \\ . & & . & . \\ . & & . & . \\ \beta_1 x_{n1} + & \ldots & + \beta_k x_{nk} + & e_n \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1k} \\ x_{21} & \ldots & x_{2k} \\ . & & . \\ . & & . \\ . & & . \\ x_{n1} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

$$= X_{nxk}\, \beta_{kx1} + e_{nx1}$$

$$= \boxed{\boldsymbol{X\beta + e}}$$

where ***X*** is called the ***design matrix***.

We assume that the form of the frequency distribution for *e* is normal: i.e.

$$f(e) = (s\pi\sigma^2)^{-\frac{1}{2}} e^{\left(-\frac{1}{2\sigma^2}e^2\right)}$$

**or**

$e \sim N(0, \sigma^2)$

The statistical model we have constructed here is

$$\boxed{(R, (2\pi\sigma^2)^{-\frac{1}{2}} e^{\left(-\frac{1}{2\sigma^2}(y_i - \beta_1 x_1 - \ldots - \beta_k x_{,})^2\right)} \beta_k \in R, \sigma \in R^+)}$$

called the ***linear model with normal error***.

For the normal linear model, the least squares estimator of β is given by

$$\boxed{\mathbf{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}}$$

The vector of residuals is

$$\boxed{\boldsymbol{Y} - \boldsymbol{X}\mathbf{b} = \boldsymbol{e}}$$

When the predictor variables $(x_1, x_2, ..., x_k)$ are correlated, we say ***multicollinearity*** exists. Where correlations are high among the $x$ variables, the computer has difficulty in calculating (i.e. rounding error, etc.) the matrix $(X'X)^{-1}$, which is necessary for many estimates (i.e. $b$ 's, standard errors, etc.). The psychologist might find, for example, the *F*-test of $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$ significant, with the *t*-test non-significant. The problem here is that the variables share information concerning $y$ .

One possible solution to this problem is to drop one or more of the correlated variables from the model. This could be accomplished on the basis of the correlations of the *x*'s. This may work; however, many times the problem is with a linear combination of variables and a bivariate correlation will not indicate a problem (high correlation).

**Problems when Correlations are High**

1. Rounding errors in the calculation of the B estimates, standard errors (this is a consequence of the calculation of $(X'X)^{-1}$ in the model with multicollinearity)

2. Results misleading – for example for the model

   $$E(x|y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

   Both *t* statistics for the betas 2 and 3 are non-significant; however, the *F*-test for the overall model ANOVA model is significant (indicate the *x* variables share information).

3. Can effect direction or sign of coefficients $\beta$ .

   - Expect significance in the problem under study to be in a particular direction, but contrary to expectation, one b estimate is positive (.8) and one is negative (-.7).

**Detection of Multicollinearity**

1. Calculate the correlation (*r*) between pairs of variables and examine large values of *r*.

2. Opposite signs than what expected for $b$ estimates.

3. Significant *F* test in the overall ANOVA test of the model but non-significant *t* tests of the individual parameters.

4. Calculate ***variance inflation factors*** for $b$ 's

The *t* test is non-significant since the standard errors ( $s_b^2$ ) are inflated as a consequence of multicollinearity.

We see that the standard error can be written as

$$s_b^2 = s^2(\frac{1}{(1-R_i^2)})$$

where $s^2$ estimates $O^{-2}$ and $R_i^2$ is *R* squared for the model that regresses $x_i$ on the remaining $x_1, x_2, ..., x_{i-1}, x_{i+1}, ...x_k$ . In other words, instead of *y*, you are now trying to predict an *x* from the other *x*'s in the model.

Note: $\boxed{\frac{1}{(1 - R_i^2)}}$ is called the ***variance inflation factor* (*VIF*)** for parameter $B_i$

$VIF_i$ is large when $R_i^2$ is large, since $x_1$ is strongly correlated to the other independent variables.

**In Practice: Problems Exists**

**If** *VIF* **is greater than 10** or $R_i^2$ is greater than .9 is a common guide.

Some software calculates ***tolerance.***

$$\boxed{Tol_i = \frac{1}{VIF_i} = 1 - R_i^2}$$

It is the reciprocal of *VIF* . Use a *Tol* value less than .1 as a cut-off.

In SPSS, click collinearity diagnostics for these statistics. An example of the SPSS output is given below. Note the VIF and TOL columns and their very high and low values, respectively, indicating problems with multicollinearity.

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | -36.657 | 68.621 | | -.534 | .603 | | |
| | Age | .561 | 1.381 | .481 | .406 | .692 | .006 | 159.967 |
| | Amount of TV news watched (hrs/wk) | 9.900 | 20.813 | .770 | .476 | .643 | .003 | 299.238 |
| | Previous victim of crime | 32.402 | 28.496 | 1.025 | 1.137 | .278 | .011 | 92.740 |
| | Age squared | 9.361E-03 | .045 | .788 | .207 | .840 | .001 | 1654.744 |
| | Interaction of age and TV news | -9.01E-02 | .615 | -.813 | -.146 | .886 | .000 | 3513.813 |
| | Interaction of age and victimization | -.993 | 1.914 | -1.266 | -.519 | .613 | .001 | 678.945 |
| | Interaction of age, TV news and victimization | 1.956E-02 | .191 | .173 | .103 | .920 | .003 | 322.604 |

a. Dependent Variable: Serious

## Possible Solutions to Multicollinearity

1. Drop correlated variables (one or more) using stepwise variable selection (a type of linear model variable selection procedure available on most statistical software packages) as an aid in variable selection.

2. Avoid multicollinearity by designing an experiment where the levels of $x$ are uncorrelated; however, given the time + cost for experiments, many researchers continue to use observational techniques for data gathering.

3. Check your data for common errors, such as including the same variable twice, including an index variable in the model that is composed of the variables already in the model, and errors in the coding of polynomials or dummy variables.

4. Always make use of previous research and your knowledge of the system under study. For example, this information should be helpful in determining what variables to include, how to create indices, guides on sample sizes, etc.

**Further readings**


- **Draper**, N. R. and H. **Smith**, 1981: Applied Regression Analysis. John Wiley and Sons, New York, 709 pp.
- **Gebotys R.,** 2008: PS600 Introduction to Linear Models Notes, //www.wlu.ca/~wwwpsych/gebotys

**12.5 Exercises**

1. Using the data from Problem 3, Exercise 9.7, find the best model. State your reasons clearly.