Contents

1 Theoretical Aspects of Fitting Models **2**
1.1 The Law of Parsimony . . . 2
1.2 Model building . . . 2
1.3 Overfitting . . . 2
1.4 Variable-Selection Procedures . . . 2
1.5 Validation and Testing . . . 2
1.6 The Coefficient of Determination . . . 3
1.7 The Adjusted Coefficient of Determination . . . 3
1.8 Akaike Information Criterion . . . 4
1.8.1 Schwarz's Bayesian Information Criterion . . . 4
1.9 AIC and BIC in Two-Step Cluster Analysis . . . 4
1.10 Multi-collinearity . . . 6
1.11 Variance Inflation Factor (VIF) . . . 6
1.12 Tolerance . . . 6

1 Theoretical Aspects of Fitting Models

1.1 The Law of Parsimony

Ockham's razor, sometimes known as the law of parsimony, is simply a maxim that states that simple explanations are usually better than complicated ones. **Ockham's razor** was originally proposed by a monk named William of Ockham. (He did not call it "Ockham's razor" or even "my razor." This is a name that has been given to it over time.)

Another version of this principle is the Law of parsimony . This says that if you are choosing between two theories, choose the one with the fewest assumptions. Assumptions here means claims of fact that have no evidence. A theory that doesn't have many assumptions, and is very simple, is called a parsimonious theory.

In the context of statistics, the law of parsimony can be interpreted as an adequate model which requires the fewest independent variables is the preferred model.

1.2 Model building

The traditional approach to statistical model building is to find the most parsimonious model that still explains the data. The more variables included in a model (overfitting), the more likely it becomes mathematically unstable, the greater the estimated standard errors become, and the more dependent the model becomes on the observed data. Choosing the most adequate and minimal number of explanatory variables helps to find out the main sources of influence on the response variable, and increases the predictive ability of the model. As a rule of thumb, there should be more than 10 observations for each variable in the model.

1.3 Overfitting

Overfitting occurs when a statistical model does not adequately describe of the underlying relationship between variables in a regression model. When overfitting happens, the model predicts the fitted data very well, but predicts future observations poorly.

Overfitting generally occurs when the model is excessively complex, such as having too many parameters (i.e. predictor variables) relative to the number of observations. A model which has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data.

1.4 Variable-Selection Procedures

In regression analysis, variable-selection procedures are aimed at selecting a reduced set of the independent variables - the ones providing the best fit to the model, in keeping with the Law of Parsimony.

1.5 Validation and Testing

When you have sufficient data, you can subdivide your data into three parts called the training, validation, and test data. Rather than estimating parameter values from the entire data set, the data set is broken into three distinct parts. During the ***variable selection*** process, models are fit on the training data, and the prediction error for the models so obtained is found by

using the validation data. Validation is the process of using part of a data set to estimate model parameters, and using the other part to assess the predictive ability of the model. Validation can be used to assess whether or not overfitting has occurred.

This prediction error on the validation data can be used to decide when to terminate the selection process or to decide what effects to include as the variable selection process proceeds. Finally, once a selected model has been obtained, the test set can be used to assess how the selected model generalizes on data that played no role in selecting the model.

1 The training set is the part that estimates model parameters.

2 The validation set is the part that assesses or validates the predictive ability of the model.

3 The test set is a final, independent assessment of the models predictive ability.

A validation set is a portion of a data set used to assess the performance of prediction or classification models that have been fit on a separate portion of the same data set (the training set). Typically both the training and validation set are randomly selected, and the validation set is used as a more objective measure of the performance of various models that have been fit to the the training data (and whose performance with the training set is therefore not likely to be a good guide to their performance with data that they were not fit to).

It is difficult to give a general rule on how many observations you should assign to each role. One important textbook recommended that a typical split might be 50% for training and 25% each for validation and testing.

1.6 The Coefficient of Determination

The coefficient of determination R^2 is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. It is the proportion of variability in a data set that is accounted for by the statistical model. It provides a measure of how well future outcomes are likely to be predicted by the model.

R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1.0 indicates that the regression line perfectly fits the data.

In the case of simple linear regression, the coefficient of determination is equivalent to the squared value of the Pearson correlation coefficient. (Consider this to be co-incidental, rather than a definition)

1.7 The Adjusted Coefficient of Determination

Adjusted R^2 (often written as and pronounced "R bar squared") is a modification of R^2 that adjusts for the number of predictor terms in a model. Adjusted R^2 is used to compensate for the addition of variables to the model. As more independent variables are added to the regression model, unadjusted R^2 will generally increase but there will never be a decrease. This will occur even when the additional variables do little to help explain the dependent variable. To compensate for this, adjusted R^2 is corrected for the number of independent variables in the model, increases only if the new term improves the model more than would be expected by chance. If too many predictor variables are being used, this will be reflected in a reduced

adjusted R^2. The adjusted R^2 can be negative, and will always be less than or equal to R^2. The result is an adjusted R^2 than can go up or down depending on whether the addition of another variable adds or does not add to the explanatory power of the model. Adjusted R^2 will always be lower than unadjusted.

Adjusted R square is generally considered to be a more accurate goodness-of-fit measure than R square. It has become standard practice to report the adjusted R^2, especially when there are multiple models presented with varying numbers of independent variables.

1.8 Akaike Information Criterion

Akaike's information criterionis a measure of the goodness of fit of an estimated statistical model. The AIC was developed by Hirotsugu Akaike under the name of "an information criterion" in 1971. The AIC is a ***model selection*** tool i.e. a method of comparing two or more candidate regression models. The AIC methodology attempts to find the model that best explains the data with a minimum of parameters. (i.e. in keeping with the law of parsimony)

The AIC is calculated using the "likelihood function" and the number of parameters (Likelihood function : not on course). The likelihood value is generally given in code output, as a complement to the AIC. Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest AIC being the best. (Although, a difference in AIC values of less than two is considered negligible).

The Akaike information criterion is a measure of the relative goodness of fit of a statistical model. It was developed by Hirotsugu Akaike, under the name of "an information criterion" (AIC), and was first published by Akaike in 1974.

$$\mathrm{AIC} = 2p - 2\ln(L)$$

- p is the number of free model parameters.
- L is the value of the Likelihood function for the model in question.
- For AIC to be optimal, n must be large compared to p.

1.8.1 Schwarz's Bayesian Information Criterion

An alternative to the AIC is the Schwarz BIC, which additionally takes into account the sample size n.

$$\mathrm{BIC} = p\ln n - 2\ln(L)$$

1.9 AIC and BIC in Two-Step Cluster Analysis

(Removed from Last Week's Class due to Version Update)

Two-Step Cluster Analysis guides the decision of how many clusters to retain from the data by calculating measures-of-fit such as ***Akaikes Information Criterion (AIC)*** or ***Bayes Information Criterion (BIC)***.

These are relative measures of goodness-of-fit and are used to compare different solutions with different numbers of segments.("Relative" means that these criteria are not scaled on a range of, for example, 0 to 1 but can generally take any value.)

Important: Compared to an alternative solution with a different number of segments, smaller values in AIC or BIC indicate an increased fit.

SPSS computes solutions for different segment numbers (up to the maximum number of segments specified before) and chooses the appropriate solution by looking for the smallest value in the chosen criterion. However, which criterion should we choose?

- AIC is well-known for overestimating the correct number of segments
- BIC has a slight tendency to underestimate this number.

Thus, it is worthwhile comparing the clustering outcomes of both criteria and selecting a smaller number of segments than actually indicated by AIC. Nevertheless, when running two separate analyses, one based on AIC and the other based on BIC, SPSS usually renders the same results.

Once you make some choices or do nothing and go with the defaults, the clusters are formed. At this point, you can consider whether the number of clusters is "good". If automated cluster selection is used, SPSS prints a table of statistics for different numbers of clusters, an excerpt of which is shown in the figure below. You are interested in finding the number of clusters at which the Schwarz BIC becomes small , but also the change in BIC between adjacent number of clusters is small.

The decision of how much benefit accrued by another cluster is very subjective. In addition to the BIC, a high ratio of distance of measures is desirable. In the figure below, the number of clusters with this highest ratio is three.

Autoclustering statistics

		Schwarz's Bayesian Criterion (BIC)	BIC Change[1]	Ratio of BIC Changes[2]	Ratio of Distance Measures[3]
Number of Clusters	1	6827.387			
	2	5646.855	-1180.532	1.000	1.741
	3	5000.782	-646.073	.547	1.790
	4	4672.859	-327.923	.278	1.047
	5	4362.908	-309.951	.263	1.066
	6	4076.832	-286.076	.242	1.193
	7	3849.057	-227.775	.193	1.130
	8	3656.025	-193.032	.164	1.079
	9	3482.667	-173.358	.147	1.162
	10	3343.916	-138.751	.118	1.240
	11	3246.541	-97.376	.082	1.128
	12	3168.733	-77.808	.066	1.093
	13	3103.950	-64.783	.055	1.022
	14	3042.116	-61.835	.052	1.152
	15	2998.319	-43.796	.037	1.059

1. The changes are from the previous number of clusters in the table.
2. The ratios of changes are relative to the change for the two cluster solution.
3. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Figure 1: Schwarz Bayesian Information Criterion

1.10 Multi-collinearity

When choosing a predictor variable you should select one that might be correlated with the criterion variable, but that is not strongly correlated with the other predictor variables. However, correlations amongst the predictor variables are not unusual. The term multi-collinearity is used to describe the situation when a high correlation is detected between two or more predictor variables. Such high correlations cause problems when trying to draw inferences about the relative contribution of each predictor variable to the success of the model.

1.11 Variance Inflation Factor (VIF)

The Variance Inflation Factor (VIF) measures the impact of multi-collinearity among the variables in a regression model.

There is no formal VIF valu-e for determining presence of multi-collinearity. Values of VIF that exceed 10 are often regarded as indicating multicollinearity, but in weaker models values above 2.5 may be a cause for concern. In many statistics programs, the results are shown both as an individual R^2 value (distinct from the overall R^2 of the model) and a Variance Inflation Factor (VIF). When those R^2 and VIF values are high for any of the variables in your model, multi-collinearity is probably an issue.

1.12 Tolerance

Tolerance is simply the reciprocal of VIF, and is computed as

$$\text{Tolerance} = \frac{1}{VIF}$$

Whereas large values of VIF were unwanted and undesirable, since tolerance is the reciprocal of VIF, larger than not values of tolerance are indicative of a lesser problem with collinearity. In other words, we want large tolerances.