Chapter 2

EVALUATION METHODS FOR LINEAR MODELS

2.0 NOTATION AND TERMINOLOGY.

Suppose we have a model under consideration which can be written in the form:

$$\mathbf{Y = X\beta + e}$$

where

Y is an (n x 1) vector of observations

X is an (n x κ) matrix of known predictors

β is a (κ x 1) vector of parameters

ε is an (n x 1) vector of errors

The least squares estimator of β is the value b, given by:

$$\hat{\mathbf{b}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}\,.$$

The fitted values are obtained from:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}$$

We use the notation

$$\mathbf{TSS} = \sum_{i=1}^{n} \left(\mathbf{y}_i - \bar{\mathbf{y}}\right)^2 \textbf{ (Total Sum of Squares)}$$

$$\mathbf{ESS} = \sum_{i=1}^{n} \left(\hat{\mathbf{y}}_i - \bar{\mathbf{y}}\right)^2 \textbf{ (Explained Sum of Squares)}$$

$$\mathbf{RSS} = \sum_{i=1}^{n} \left(\hat{\mathbf{y}}_i - \mathbf{y}_i\right)^2 \textbf{ (Residual Sum of Squares)}$$

Some of the methods most commonly used in the literature for model selection are the following ones:

2.1 RESIDUAL MEAN SQUARE CRITERION

Let p be, the number of predictors used in a linear model ($p \leq \kappa$). An estimate of the RMS (Residual Mean Square) is given by:

$$MSE = \frac{RSS}{n - p}$$

for all the possible $2^{\kappa}-1$ models.

The criterion is based on the function(see Drapper N.R. and Smith(1981)):

$$S^2(p) = \frac{MSE(p)}{\binom{\kappa}{p}}$$

that is, the average of the MSE for all the models of dimension p. The criterion works well for large data sets.

After the estimation of MSE and $S^2(p)$, the procedure is the following:

- We plot the average $S^2(p)$ against p. (Figure 1)

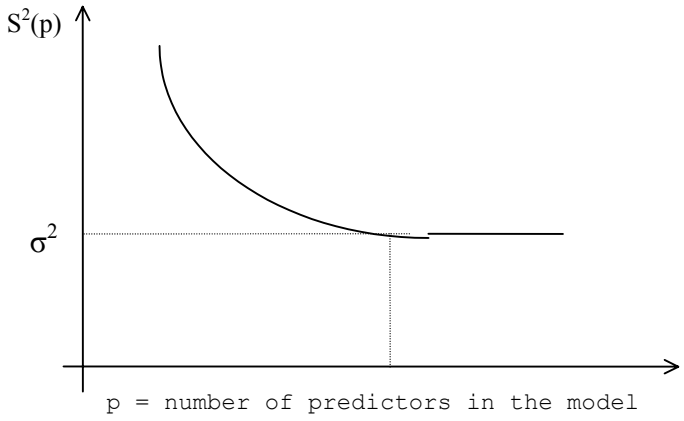

Figure 1: *Plot of the average $S^2(p)$ against p.*

- At the beginning, the curve is a descending one. As more and more predictor variables are added to an

already overfitted* equation, the residual mean square will tend to stabilize - the curve becomes a line - and approach the true value of σ^2 as the number of variables increases, provided that all important variables have been included. Thus, this procedure gives us an «asymptotic» estimate of σ^2 with which we can choose a model, or models, whose residual variance estimate is close to RMS(κ) - the RMS of the model that contains all the predictors- and which contain the fewest predictor variables to achieve that. The elbow of the curve indicates the appropriate number of predictors included in the model.

2.2 THE COEFFICIENT OF DETERMINATION

The most common measure of the goodness of fit of a linear model is the coefficient R^2. This measure, known as the coefficient of determination, is defined by (e.g Drapper N.R. and Smith(1981)):

$$R^2 = \frac{\text{Sum of Squares due to regression}}{\text{Total Sum of Squares}} \Rightarrow$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \ .$$

That is, R^2 measures the «proportion of total variation about the mean $\overline{Y}$ explained by the regression». The coefficient of determination lies always between 0 and 1 and the fit of a model is satisfactory if R^2 is close to unity. A value of R^2 is usually regarded as large, if it is greater than 0.7 or 0.8. An advantage of this measure is that it can be calculated very easily.

* The fitting of regression equations that involve more predictor variables than are necessary to obtain a satisfactory fit to data is called overfitting.

According to this measure, the best model is the one with the largest R^2 coefficient. However, R^2 can always be increased after the inclusion of any predictor. Therefore, the model :

$$Y = X\beta + \gamma w + e$$

has always a higher R^2 than the model

$$Y = X\beta + e$$

whatever the extra predictor variable w is. This occurs because the R^2 coefficient is an ascending function of the inclusive predictors in the model. If n is the total number of observations and p is the number of the inclusive predictors ($p \leq \kappa$), it can be proved (see, e.g., Panas(1993)), that :

$$E(R^2) = \frac{p - 1}{n - p}$$

which means that the mean of the distribution of R^2 increases as p increases. A solution to this problem is to estimate the appropriate number of dimensions, p, graphically. The procedure is described in the next steps:

- We estimate R^2 for all the possible $2^{\kappa}-1$ regressions.
- We plot the maximum value of R^2 coefficient which corresponds to every dimension p, towards p. (Figure 2)

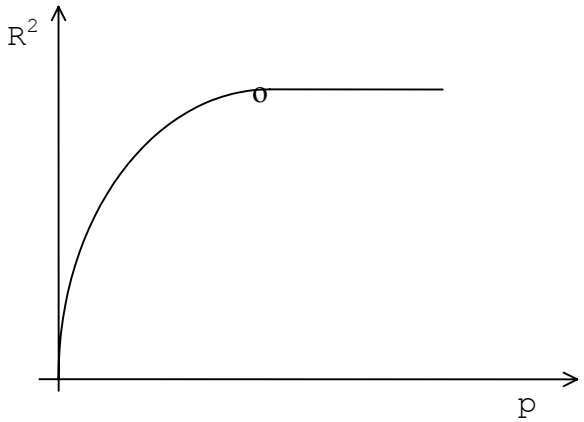

Figure 2: *Plot of the maximum value of R^2 coefficient which corresponds to every dimension p, against p.*

- At the beginning, the curve is an ascending one. As p increases the curve becomes almost a line. The value of p that corresponds to the elbow of the curve is the appropriate number of predictors that have to be included in the model.

2.3 ADJUSTED COEFFICIENT OF DETERMINATION

An alternative method of R^2 is a modification of R^2 which takes into account not only the fit of the model but the number of predictors as well while, at the same time, it penalizes the inclusion of extra variables. The modified R^2, known as the adjusted R^2, is obtained through the formula(e.g Drapper N.R. and Smith(1981)):

$$R^2_{adj} = 1 - \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} \cdot \left[\frac{n-1}{n-p}\right]$$

$$= 1 - \frac{n-1}{n-p}\left(1 - R^2\right)$$

R^2_{adj} may decrease, despite the inclusion of a new variable in a model. This happens when the improvement of the fit of the model that is achieved after the inclusion of the new variable is not large enough. R^2_{adj} does not lie in [0,1]. It can take negative values as well. However we can choose between the potential models finding the one with the highest R^2_{adj}.

Let us now assume that, in a given model we omit an predictor, say x_j . It can be proved (Greene W.H 1993), that R^2_{adj} will become larger if the t-ratio for the coefficient β_j is smaller than 1. Otherwise, R^2_{adj} becomes smaller. This is not the case with R^2. The omission of a variable leads always to a model with smaller R^2.

R^2_{adj} is obviously an improvement of R^2. Sometimes, however this improvement is not adequate.

Some alternative criteria that also penalize the increase of the number of parameters and can be used for model selection are given in the sequel.

2.4 COEFFICIENT OF MULTIPLE CORRELATION

The square root of R^2 is called coefficient of multiple correlation and it is symbolized by R. Many statisticians use R instead of R^2. The apparent reason of using R is that $0<R^2<1$ and consequently $R>R^2$. Nevertheless it is difficult to explain R and especially the sign of R.

2.5 SELECTION OF REGRESSION COEFFICIENTS.

The most common method to adjust a model in a set of data is the 'least square method'. Given κ predictors there are $2^{\kappa}-1$ possible models among which, we have to chose the 'best' one.

After having adjusted the model to the data, we have to select these predictors that are significant for the model. That is, we have to test the hypothesis:

$$\begin{Bmatrix} H_0: \beta_i = 0 \\ H_1: \beta_i \neq 0 \end{Bmatrix} \quad (2.5.1)$$

If we have not enough evidence to reject H_0 then the contribution of the corresponding X_i in the model is not significant and we have to omit the term $\beta_i X_i$ from the model.

There are different methods for the choice of the regression coefficients. The most important are:

2.5.1 THE BACKWARD ELIMINATION PROCEDURE.

The backward elimination method is more economical than the 'all regressions' method in the sense that it tries to examine only the 'best' regressions containing a certain number of variables. The method starts by computing a regression equation containing all predictors and systematically it eliminates this variable that has not a significant contribution to the model in the sense that it has the smallest value of t-statistic. The basic steps of the procedure are the following:

- We estimate the regression equation that contains all the predictors.
- For every predictor X_i, we test the hypothesis (2.5.1) and we accept the null hypothesis that the i-predictor is not significant, if:

$$|t_i| = \frac{\hat{\beta}_i}{S_{\hat{\beta}}} < t_{n-p-1;\frac{\alpha}{2}} \quad (2.5.2)$$

  (where $\hat{\beta}_i$ is the least square estimator of the coefficient β_i of the variable X_i and $S_{\hat{\beta}_i}$ is the estimated standard deviation of the estimator β_i of the regression coefficient). Then, we eliminate that predictor that presents the smallest t-statistic.
- In the sequel, we estimate the regression equation with κ-1 predictors. We also test the hypothesis (2.5.1)and we eliminate that predictor with the smallest $|t_i|$ statistic that is also smaller than $t_{n-p;\alpha/2}$.
- We continue the procedure till all the $|t_i|$ statistics are greater than the critical value of the t-distribution.

We have to point out that the elimination of the i-predictor can equivalently be tested using the F instead of the t-distribution.

2.5.2 THE FORWARD SELECTION PROCEDURE.

The method of the forward selection procedure uses the opposite logic from the one used in the backward elimination method. In every step, a predictor variable is added in the model. The steps of the procedure are the following:

- We estimate all the correlations between all the predictors and the response Y. We choose that predictor that presents the highest correlation with the response. Let this predictor be, X_1.

- We estimate the model with predictor X_1. We also estimate the t-statistic to test the hypothesis $H_0:\beta_1=0$. If we accept H_0, that is if:

$$|t| = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}} < t_{n-2;\frac{\alpha}{2}} \quad (2.5.3)$$

  we have to stop the procedure. If not, we have to choose the second predictor.

- For the selection of the second predictor, we estimate the partial correlations of Y with all the predictors, keeping X_1 constant. Let X_2 be this predictor that presents the highest absolute value of the partial correlation coefficient. In the sequel, we estimate the model with predictors X_1, X_2 and we test the hypothesis $H_0:\beta_2=0$. If

$$|t| = \begin{cases} \leq t_{n-3;\alpha/2} & \text{we stop the procedure and we consider the model that contains only the predictor } X_1 \\ \geq t_{n-3;\alpha/2} & \text{we procced to the selection of a third predictor} \end{cases}$$

- When the last selected predictor presents a |t| statistic smaller than the critical value, the procedure is terminated.

The advantage of this method is that it is computationally easy. On the other hand this method has a number of disadvantages. The correlations between the response and the predictor variables take generally large values. Another disadvantage is that offers only the possibility of adding an independent variable and from the time that a variable is incorporated in the model it is retained even if the value of the statistical function t doesn't exceed the critical level in any stage of the method.

We have to point out that the incorporation of the i-predictor can equivalently be tested using F instead of t-distribution.

2.5.3 THE STEPWISE REGRESSION PROCEDURE.

This method looks like the forward selection procedure but can be considered as a combination of the previous two procedures.

Initially, it incorporates a predictor in the equation, following the same method as the forward procedure. The difference is that in every stage we totally reexamine the predictors that have already been incorporated in the model using the criterion of the backward procedure. So, it is possible to eliminate some predictors that previously had been considered significant. The procedure continues until we arrive at a subset of predictors for which none of the predictors presents statistic |t| smaller than the critical value of t-distribution.

2.6 RIDGE REGRESSION

The method of ridge regression was first suggested by **A.E. Hoerl (1962),** and provides a different way for model selection. The procedure is intended to overcome ''ill conditioned'' situations where correlations between the various predictors in the model cause the X´X matrix to be close to singular (determinant equal to zero), giving rise to unstable parameter estimates. The estimation of the regression coefficients β of a linear model is :

$$\hat{b}(\vartheta)=(X'X + \theta\ I_r)^{-1}\ X'Y$$

where θ is a positive number - usually lying in the range (0,1). The selection procedure eliminates these variables that give the smaller $\hat{b}_i(\vartheta)$ in absolute value, that is, these $\hat{b}_i(\vartheta)$ that have the least predictive efficiency. After this it eliminates these variables whose regression coefficients tend to zero when θ increases.

Ridge regression is useful and appropriate in circumstances where it is believed that the values of the regression parameters are unlikely to be ''large''. The choice of θ is essentially equivalent to an expression of how big one believes those b's to be. In circumstances where one cannot accept the idea of restrictions on the b's, ridge regression would be completely inappropriate. A disadvantage of the method is that it doesn't clarify when the procedure finishes. It should be noted that, in applications, ridge regression has not usually been used as a model selection procedure. We mention this use only as a possibility.

There is a large and growing literature on the many aspects and generalization of ridge regression. Some important comments are given in the assignment of Draper and Nostrand (1979).

2.7 MALLOWS C_P STATISTIC

An alternative statistic, which has gained popularity in recent years is the C_p statistic, initially suggested by **C.L. Mallows (1973).** This has the form:

$$C_p = \frac{RSS(p)}{s^2} + 2(p + 1) - n$$

where

RSS(p) is the residual sum of squares from a model containing p parameters

and

s^2 is the residual mean square of the model that contains all the predictors.

It can be shown that a model with p predictors is adequate if:

$$E(C_p)=p$$

It follows that a plot of C_p versus p will show up the 'adequate models' as points fairly close to the C_p=p line. The best model is chosen after inspecting the C_p plot. We would look for a regression with a low C_p value about equal to p.

It can be proved that the statistic C_p is related to R^2 and R^2_{adj} through the formulas:

$$C_p = \frac{(n - \kappa)\left(1 - R_p^2\right)}{\left(1 - R_\kappa^2\right)} + 2(p + 1) - n$$

where κ is the total number of the parameters that can be included in the regression, and

$$C_p = \frac{(n - p)\left(1 - R^2_{adj}(p)\right)}{\left(1 - R^2_{adj}(\kappa)\right)} + 2(p + 1) - n$$

where $R^2_{adj}(\kappa)$ is the adjusted coefficient of determination that is estimated when all the parameters that can be included in the regression are considered.

2.8 HOCKING'S S_P CRITERION.

A criterion quite similar to C_P is the S_P criterion given by **Hocking (1976).** The criterion is based on the minimization of the quantity:

$$S_P = \frac{RSS(p)}{(n - p + 1)(n - p - 1)}$$

where RSS(p) is the residual sum of squares of a model that includes p variables from the κ candidate variables.

2.9 CROSS VALIDATION - PRESS CRITERION

One of the most useful methods in model selection problems is the cross validation (CV) method. Similar to other model selection methods, the CV method selects a model by minimizing the overall expected squared prediction error. The idea is simply to split the data into two parts, using one part to derive a prediction rule and then judge the goodness of the prediction by matching its outputs with the rest of the data (hence the name cross validation). The first part contains n_c data points used for fitting a model (model construction), whereas the second part contains $n_v = n - n_c$ data points reserved for assessing the predictive ability of the model(model validation). There are $\binom{n}{n_v}$ different ways to split the data set. Cross validation, selects the model with the best average predictive ability calculated based on all(or some) different ways of data splitting.

Clearly, the computational complexity of this method increases as n_v increases. That is why the simplest cross-validation with n_v=1(leave one-out CV) has been the main focus of researchers' attention in the last years. Suppose that p is the number of predictions in a regression equation. The basic steps of the leave one-out CV are the following (Predictive Sum of Squares procedure (**Allen 1971):**

1. Delete the first set of observations on the response and predictor variables.
2. Fit all possible regression models to the remaining n-1 data points.
3. Use each fitted model to predict Y_1 by $\hat{Y}_{1p}$(say) and so obtain a predictive discrepancy $(Y_1 - \hat{Y}_{1p})$ for all the possible regression models.
4. Repeat steps 1,2 and 3 , but deleting the second observation to give $(Y_2 - \hat{Y}_{2p})$ values, the third observation to give $(Y_3 - \hat{Y}_{3p})$ values and so on, to n deletions.
5. For each subset regression model calculate the predictive discrepancy sum of squares :

$$\text{PRESS} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{ip}\right)^2$$

6. Choose the «best» subset regression. This will have a comparatively small predictive sum of squares but will not involve too many predictors.

The CV with n_v=1 is inconsistent in the sense that the probability of selecting the model with the best predictive ability does not converge to 1 as the total number of observations $n \rightarrow \infty$, and is too conservative in the sense that it tends to select an unnecessarily large model (overfitting). That is why, many investigators

propose leave-n_v-out CV, finding methods that reduce the amount of computation :

- **Jun Shao (1993),** proposes a leave-n_v-out cross validation where $n_v/n \rightarrow 1$ as $n \rightarrow \infty$. Selecting n_v in this way, rectifies the inconsistency of the leave-one-out CV.

- **Ping Zhang (1993)** also proposes a leave-n_v-out cross validation. He proves that the general -leave-n_v-out CV is asymptotically equivalent to the FPE crierion. (Final Prediction Error) and proposes two computationally more feasible methods, the r- fold CV (MCV_k*)and the repeated learning-testing method (RLT_k), which is essentially a bootstrap method.

2.10 BOOTSTRAP

Bootstrap is a data-resampling method, based on computers, which substitutes considerable amounts of computation in place of theoretical analysis. Generally, bootstrap replies to how accurate is an estimation $\hat{\theta}$ of an unknown parameter θ using many replications.(B. Efron and R. Tibshirani (1986)).

Regression Models

Let, the data set z for a linear regression model consists of n points z_1, z_2 , … , z_n , where each z_i is itself a pair, say

$$z_i = (x_i , y_i) \qquad i = 1,2,\ldots,n$$

The structure of the linear model is expressed as:

$$y_i = x_i \beta + e_i \qquad i = 1,2,\ldots,n$$

- **Bootstrapping Residuals (Efron 1979)**

⇒Given $z_i = (x_i , y_i)$ $i = 1,2,\ldots,n$ we obtain the least square estimates of the regression coefficients:

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

⇒We calculate the residuals

$$\hat{e}_i = y_i - x_i\hat{\beta} \quad \text{for } i = 1,2,\ldots,n$$

⇒We estimate the empirical distribution of the $\hat{e}_i$:

$\hat{F}$: probability 1/n on $\hat{e}_i$ for $i = 1,2,\ldots,n$

⇒We select a random sample of bootstrap error terms

$$\hat{F}(\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n) = e^*$$

⇒Bootstrap responses y_i^* are generated :

$$y_i^* = x_i \hat{\beta} + e_i^* \quad \text{for } i = 1,2,\ldots,n$$

It may seem strange that the x_i are the same for the bootstrap data as for the actual data. This happens because we are treating x_i as fixed quantities rather than random.

⇒The bootstrap L.S.E estimate $\hat{\beta}^*$:

$$\hat{\beta}^* = (X'X)^{-1}X'Y^*$$

is the minimizer of the residual squared error for the bootstrap data,

$$\sum_{i=1}^{n}\left(y_i^* - x_i\hat{\beta}^*\right)^2 = \min_b \sum_{i=1}^{n}\left(y_i^* - x_i b\right)^2$$

- **Bootstrapping Pairs (Efron 1982)**

⇒Given $z_i = (x_i, y_i)$ $i = 1,2,\ldots n$, let $\hat{F}$ be the distribution putting probability 1/n on each of the n points.

⇒The bootstrap data set now, is $z^* = (z_1^*, z_2^*, \ldots, z_n^*)$.

⇒We obtain the L.S.E:

$$\hat{\beta}^* = \left(X^{*\prime}X^*\right)^{-1}X^{*\prime}Y^*$$

$\Rightarrow$ Then, the bootstrap predictions are:

$$\hat{y}^* = \hat{\beta}^* X^*$$

$\Rightarrow$ The residual squared error is:

$$RSE = \sum_{i=1}^{n}\left(y_i^* - \hat{y}_i^*\right)^2$$

Model Comparison

Let X,D,E be the data matrix of all the variables and two design matrices -subsets of X- respectively. Then, the L.S.E for the two models are:

$$\hat{\beta}(D) = \left(D'D\right)^{-1}D'Y \quad \text{and} \quad \hat{\beta}(E) = \left(E'E\right)^{-1}E'Y .$$

The residual squared errors for the two models are:

$$RSE(D) = \sum_{i=1}^{n}\left(y_i - \hat{y}_i(D)\right)^2 \quad \text{and} \quad RSE(E) = \sum_{i=1}^{n}\left(y_i - \hat{y}_i(E)\right)^2$$

Using bootstrapping residuals, or bootstrapping pairs, we estimate the above residual squared errors. Generally, small values of the residual squared error indicate good prediction. The question how do the D and E variables compare as predictors of the response, can be phrased as a comparison between RSE(D) and RSE(D). A handy comparison statistic is

$$\hat{\theta} = \frac{1}{n}\left[RSE(E) - RSE(D)\right]$$

A positive value of $\hat{\vartheta}$ would indicate that the E variables are not as good as the D variables and that the D variables are better predictors. But we ca not decide if this is really true until we understand the statistical variability of $\hat{\vartheta}$. Confidence intervals using bootstrap methods is a good way to answer if the difference is statistically significant or if it is of practical importance.

Jun Shao (1996) proposes an alternative bootstrap procedure. He argues, that although the bootstrap estimates have good properties, the bootstrap selection procedure is inconsistent, in the sense that the probability of selecting the optimal subset of variables does not converge to 1, as $n \to \infty$. This inconsistency can be rectified by modifying the sampling method used in drawing bootstrap observations.

For bootstrapping pairs, it is found that instead of drawing n bootstrap observations, much less bootstrap observations should be sampled: the bootstrap selection procedure becomes consistent if we draw m bootstrap observations with $m \to \infty$ and $m/n \to 0$.

For bootstrapping residuals, he suggests multiplying the residuals by a factor $\sqrt{n/m}$, where m satisfies $m/n \to 0$ and $m \to \infty$.

2.11 LIKELIHOOD RATIO TEST (χ^2 - Test)

Let $\hat{\Theta}_S$, $\hat{\Theta}_R$ be the M.L.E for the saturated model (the model that contains all the possible predictors) and the reduced model(for which the predictors is a subset of the predictors of the saturated model), respectively. Let also $L(\hat{\Theta}_S)$ and $L(\hat{\Theta}_R)$ be the maximum likelihoods. If a model fits well the data then we would expect $L(\hat{\Theta}_R)$to be quite as large as $L(\hat{\Theta}_S)$. We define the function

$$\Omega = \frac{L\left(\hat{\Theta}_R\right)}{L\left(\hat{\Theta}_S\right)} \quad .$$

Using this ratio we can test the hypothesis:

H_0 : Reduced Model is better

H_1 : Saturated Model is better

Values of Ω close to 1 indicate that H_0 describes well the data, while values of Ω close to 0 indicate significant lack of fit of the reduced, compared with the saturated model. In order to test the hypothesis we have to know the distribution of Ω, when H_0 is true.

We define the quantity

$$\text{Deviance} = D = -2\log\Omega.$$

Asymptotically, when H_0 is true, the deviance follows χ^2_{n-p} where n is the number of observations and p the number of the predictors of the reduced model.

So,

$$D = -2\log\left(\frac{L\left(\hat{\Theta}_R\right)}{L\left(\hat{\Theta}_S\right)}\right) = 2\left[\log L\left(\hat{\Theta}_S\right) - \log L\left(\hat{\Theta}_R\right)\right]$$

Suppose that we have the following sequence of nested models:

$$M_0 \subset M_1 \subset M_S$$

where M_S is the saturated model and M_0 , M_1 , M_S have p, p+q and n predictors, respectively. We want to test the hypothesis :

$$H_0: M_0 \text{ model is beter}$$
$$H_1: M_1 \text{ model is better}$$

Furthermore, let D_0 be the deviance that is derived when comparing M_0 and M_S , D_1 the deviance (Deviance=-2logL(Θ)) that is derived when comparing M_1 and M_S , and $\hat{\Theta}_0$, $\hat{\Theta}_1$ and $\hat{\Theta}_S$ the M.L.E for the three models, respectively. We can test the previous hypothesis using the likelihood ratio statistic:

$$\Delta D = D_0 - D_1 =$$

$$= 2[\log L(\hat{\Theta}_S; y) - \log L(\hat{\Theta}_0; y)]$$

$$- 2[\log L(\hat{\Theta}_S; y) - \log L(\hat{\Theta}_1; y)] =$$

$$= 2\,[\log L(\hat{\Theta}_1; y) - \log L(\hat{\Theta}_0; y)] = 2\log\frac{L\left(\hat{\Theta}_1; y\right)}{L\left(\hat{\Theta}_0; y\right)}.$$

$D_0 \sim X^2_{n-p}$ and $D_1 \sim X^2_{n-p-q}$. Under the assumption of independence, $\Delta D \sim X^2_q$. That is, if we find a large value of ΔD, we reject the null hypothesis of M_0, i.e. we prefer model M_1 to M_0. (Likelihood Ratio test or X^2 - test)

2.12 AKAIKE INFORMATION CRITERION (AIC).

The use of the asymptotic chi-square method for model comparison is, limited to the case where one model is a restricted version of the other. The AIC statistic, introduced by **Akaike [1973, 1977],** can be used in cases, where the asymptotic chi-square test is not feasible. The AIC statistic associated with a model is defined by the formula:

$$AIC = -2 \ln L + 2 \text{ d.f}$$

where L is the maximum likelihood of the model and d.f. is the effective number of parameters in the model. In the case of linear regression models the problem simplifies to the minimization of the function

$$AIC = n\ln(\hat{\sigma}^2) + 2p$$

Note that adding twice the degrees of freedom compensates the error of the large value of the maximum likelihood as a consequence of increasing the number of parameters. If more parameters are used to describe the data, it is natural to get a larger likelihood, possibly without improving a true goodness of fit, and the AIC avoids this spurious improvement of fit by penalizing the use of additional parameters. The model which gives the minimum AIC value is considered the best fitting model.

2.13 BAYESIAN INFORMATION CRITERION (BIC)

A Bayesian version of the AIC is one that **Akaike (1978,1979)** called BIC. For the linear regression models it is defined by the function:

$$BIC = n \ln(\hat{\sigma}^2) + p \ln(n)$$

In this criterion the parameter p penalizes $\hat{\sigma}^2$ more than AIC since for n>7, ln(n)>2.

2.14 AMEMIYA PREDICTION CRITERION (PC)

Amemiya (1980) suggested a criterion similar to the one by Akaike that is based on the minimization of the quantity:

$$PC = \hat{\sigma}^2 \frac{n + p}{n - p}$$

2.15 HANNAN'S CRITERION (HC)

Hannan (1981) suggested a criterion that is based on the minimization of the quantity:

$$HC = n \ln(\hat{\sigma}^2) + p (2\ln(\ln(n))$$

2.16 THEIL'S RESIDUAL VARIANCE CRITERION (RVC)

The following criterion was suggested by **Theil (1961)** and is based on the minimization of the quantity:

$$RVC = \hat{\sigma}^2 \frac{n}{n - p}$$

2.17 PARZEN'S CRITERION FOR AUTOREGRESSIVE TRANSFER FUNCTIONS (CAT)

Parzen (1977) suggested the following criterion, called CAT:

$$CAT(p) = \begin{cases} -\left(1 + \frac{1}{n}\right) & p = 0 \\ \frac{1}{n}\sum_{j=1}^{n} \frac{1}{\hat{\sigma}_j^2} - \frac{1}{\hat{\sigma}_p^2} & p = 1,2,3 \end{cases}$$

where $\hat{\sigma}_j^2$ is the unbiased estimator of σ_n^2 when an AR(j) model (autoregressive series of order j) is fitted in the data and n is the number of observations. The order p is selected when CAT(p) takes the minimum value.

2.18 BAYESIAN MODEL CHOICE.

In Bayesian statistics a prior belief for a model is given by the prior probability and we are interested in a posterior belief of the model, that is the posterior probability.

Assume that we have a countable set M of competing models for a given set of data ε. Let model $m \in M$ have a vector θ_m of unknown parameters, the dimension of which may vary from model to model. The posterior probability of a model m given the data ε, is given by:

$$\pi(m \mid \varepsilon) = \frac{\pi(m) \cdot \int_{\theta_m} \pi(\varepsilon \mid m, \theta_m) \cdot \pi(\theta_m \mid m)\, d\theta_m}{\sum_{m \in M} \pi(m) \cdot \int_{\theta_m} \pi(\varepsilon \mid m, \theta_m) \cdot \pi(\theta_m \mid m)\, d\theta_m}$$

where $\pi(\varepsilon \mid m, \theta_m)$ is the likelihood given the model m and the parameter vector θ_m, $\pi(m)$ is the prior probability for model m, and $\pi(\theta_m|m)$ is the prior of the parameter vector θ_m given the model m. Inference about the model selection problem may be done using the Bayes Factor (BF) of model m_i against model m_j given by

$$BF = \frac{\pi(m_i \mid \varepsilon)}{\pi(m_j \mid \varepsilon)} \cdot \frac{\pi(m_j)}{\pi(m_i)} = \frac{\int_{\theta_{m_i}} \pi(\varepsilon \mid m_i, \theta_{m_i}) \cdot \pi(\theta_{m_i} \mid m_i)\, d\theta_{m_i}}{\int_{\theta_{m_j}} \pi(\varepsilon \mid m_j, \theta_{m_j}) \cdot \pi(\theta_{m_j} \mid m_j)\, d\theta_{m_j}}.$$

However, the Bayes Factor require evaluation of the integrals in the nominator and the denominator. These integrals are in general difficult to calculate.

Because of this difficulty numerical or asymptotic methods are necessary to obtain posterior summaries of interest. A sampling strategy of parameters θ, called MCMC (Markov Chain Monte Carlo) is a tool to overcome these difficulties.

Methods such as SSVS (Stochastic Search Variable Selection- George and McCullogh(1993)) and Reversible Jump (Peter Green(1995)) are also used for the Bayesian model choice.

2.19 MODEL SELECTION FOR TIME SERIES

The criteria that are mostly used in time series, trying an autoregressive-moving average model of order p and q with maximum likelihood L and N time points, are:

- AIC=-2lnL + 2(p+q+1)
- AICC=$-2\ln L + \frac{2N(p+q+1)}{(N-p-q-2)}$

  (Corrected Akaike Criterion)
- BIC (Bayesian Information Criterion)

If we want to test the adequacy of the model we use the **«Portmanteau Test». (Box and Jeckins (1970))**

Let $\hat{\alpha}_t$ be the residuals from a time series, and $r(\hat{\alpha}_t)$ be the autocorrelations of the $\hat{\alpha}_t$. Then it is known that:

$$r(\hat{\alpha}_t) \sim N(0,1/N)$$

and that

$$Q = N\sum_{s=1}^{K} r_s^2(\hat{\alpha}_t) \sim \chi^2_{K-p-q}$$

where K is the lag of the model.

So, testing the hypothesis

H_0 : The model is a satisfactory one

H_A : not H_0

we have enough evidence to reject H_0 for values of Q greater than χ^2_{K-p-q}.