Building a Regression Model.

1) Think about the regression in advance.

a) What variables are needed? Is it available? Is it readily measurable? Are there redundancies?

b) How many observations are needed? More variables to be examined requires more cases.

Data should be edited (outliers).

2) Model considerations.

We frequently have more variables than we need. We must consider some technique to reduce the number of variables.

Considerations on variable reduction. We want to determine what would be a good subset of variables for the model, including polynomial terms and interactions.

One thing we cannot do is fit all the variables and leave off the non-significant ones. Since the behavior of fully adjusted variables is unpredictable, particularly when two are highly correlated, leaving off all non-significant variables may well leave off an important variable which is correlated with another variable (such that neither appears significant).

All possible regressions Procedure:

All possible (or reasonable) regressions are fitted. Some criteria is used to determine which few models would be worthy of more consideration.

first, you cannot fit a model which has more variables than observations (you shouldn't even be very close).

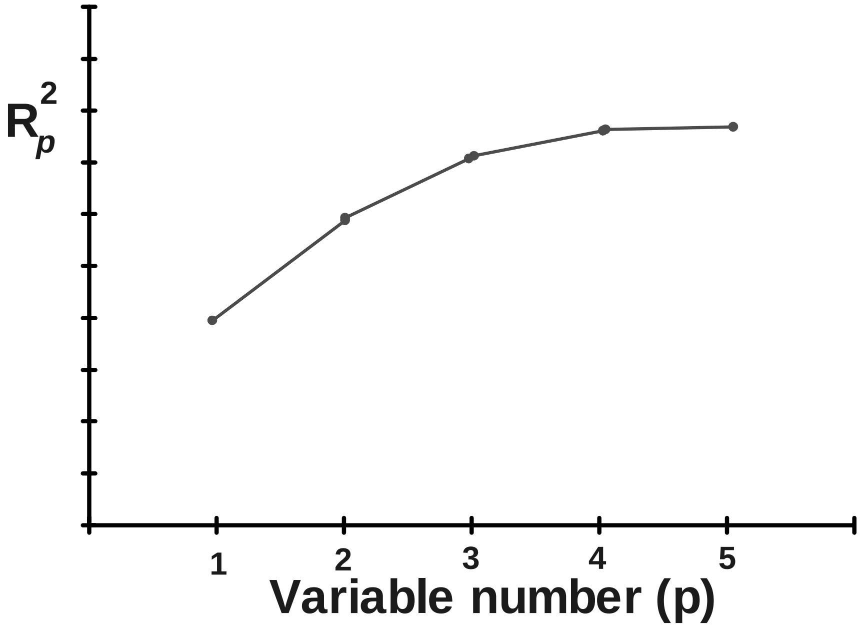
Then we evaluate all possible models,

Criterion:

a) R_p^2 criterion $= \frac{\text{SSReg}_p}{\text{SSTotal}} = 1 - \frac{\text{SSError}_p}{\text{SSTotal}}$

Note: This is simply the coeff of mult determination.

We do not search for a “maximum" because adding variables always makes R^2 go up. However, it should level off after all important variables are in. Therefore, plot values of R^2 on the number of observations.



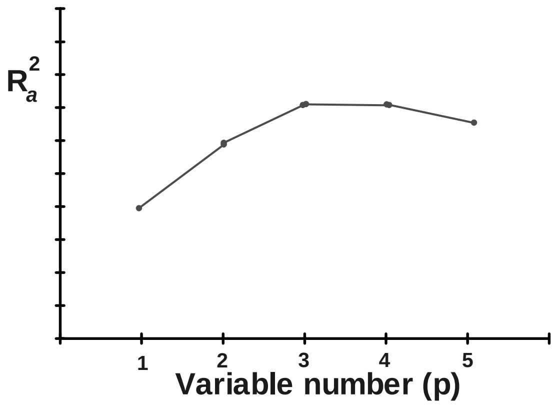
R_p^2 1 2 3 4 5 6 No. variables

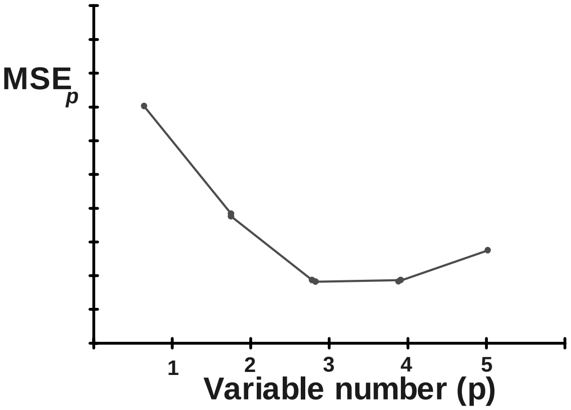
b) MSE_p or R_a^2 criterion $= 1 - \frac{\text{n}-1}{\text{n}-\text{p}} * \frac{\text{SSError}}{\text{SSTotal}} = 1 - \frac{\text{MSE}}{\frac{\text{SSTotal}}{\text{n}-1}}$

This R_a^2 is the adjusted coeff of mult determination. This value can go back down for increasing number so variables added. Since for a given problem the SSTotal/n-1 is constant, this value will go up or down as MSE goes up or down, hence the

MSE_p criterion : this can also be graphed, and examined for a minimum

We want to reduce the MSE_p value, and cause it to be a minimum. If values which take up 1 df do not account for at least 1 MSE unit is SS, then the MSE_p will go up. Of course, these variables have an F value of 1, and may not be significant.





c) C_p criterion

Consider the Total Mean Square Error : $E(\hat{Y}_i) - \mu_i$

where $E(\hat{Y}_i)$ is the predicted value for a particular regression, and μ_i is the true regression response.

If we consider the random error component, we get the usual, $\sigma^2_{\hat{Y}} + B^2$, where the B is a bias term

$$\sigma^2_{\hat{Y}} + (E(\hat{Y}_i) - \mu_i)^2,$$

where the failure of the regression line to fall on the true value is a bias

Then the Total Mean Square Error is given by

$$\sum_{i=1}^{n}\left[\sigma^2_{\hat{Y}} + (E(\hat{Y}_i) - \mu_i)^2\right] = \sum_{i=1}^{n}\sigma^2_{\hat{Y}} + \sum_{i=1}^{n}(E(\hat{Y}_i) - \mu_i)^2$$

If we have a good measure of the value of σ^2, then we can create a criteria by calculating

$$\frac{\text{TotalSSE}}{\text{TrueMSE}} = \frac{\Sigma\sigma^2_{\hat{Y}} + \Sigma(E(\hat{Y}_i) - \mu_i)^2}{\sigma^2}$$

If the full model is *carefully chosen*, the estimate of this the C_p statistic, and is given by

$$C_p = \frac{SSE_p}{MSE(X_1, X_2, ..., X_{p\text{-}1})} - (n - 2p)$$

where SSE_p comes from the fitted regression. This ratio should have a value approximately equal to p when there is no bias.

C_p

Graphic follows pattern similar to MSE_p above

Use of the C_p statistic requires more care in choosing the full set of variables, in order for MSE(Full) to be unbiased. Include all quadratic and cubic terms, all interactions, but exclude all extemporaneous variables.

d) PRESS$_p$ criterion (PRESS = Prediction SS)

This criterion is based on deleted residuals.

$$d_i = Y_i - Y_{i(i)} = \frac{e_i}{1-h_{ii}}$$

There are n deleted residuals in each regression, and

$$\mathrm{PRESS}_p = \sum_{i=1} d_i^2 = (Y_i - Y_{i(i)})^2$$

Since the value of d$_i$ is a measure of prediction error, regressions with small PRESS$_p$ values are good candidates.

This statistic can be calculated and graphed similar to the MSE$_p$ statistic

Stepwise Regression : we know that we should not delete batches of variable from regressions when they are not significant, because we do not know what the effect of adjusting for each other has been (except when order dependent TYPE I SS has been used, as in a pure polynomial).

Generally, when we wish to examine the effect of adding and deleting variables, this should be done one variable at a time. There is an algorithm to do this automatically called stepwise regression.

FORWARD stepwise regression:

Consider p − 1 candidate variables. We want to select the best set.

1) Fit all p − 1 SLR's, and evaluate the usual F statistic

$$F_k^* = \frac{MSR(X_k)}{MSE(X_k)}$$

Out of the p − 1 variables, the one with the best F statistic is the one which gives the greatest reduction in $MSR(X_k)$

2) The BEST variable has been selected. Now, is there a 2 variable model which is good?

Suppose that the first fit of SLR showed that X_{37} was best. Now fit all possible (p-2) regressions with X_{37} included, and find the best of these with the statistic,

$$F_k^* = \frac{MSR(X_k|X_{37})}{MSE(X_{37},X_k)} = \frac{b_k}{s_{b_k}}$$

We have assumed that X_{37} belongs in the model, and fit all 2 factor regressions to see if any other variable works well with X_{37}. Suppose that X_9 enters, and is the best variable in conjunction with X_{37}.

3) The BEST 2 factor model has been selected. Now, is there a 3 variable model which is good? We assume that X_{37} and X_9 belong in that model, these are retained.

Now fit all possible (p-3) regressions with X_{37} and X_9 included.

$$F_k^* = \frac{MSR(X_k|X_{37},X_9)}{MSE(X_{37},X_9,X_k)} = \frac{b_k}{s_{b_k}}$$

This process can be continued until no variables will enter into the model and be significant.

This process is called FORWARD stepwise regression. Sometimes, no single variable will enter and be significant, but some group of 2 or 3, between them, will lower the MSE sufficiently for all to be significant. FORWARD will not catch this.

BACKWARDS elimination.

1) Fit the full model (p-1 variables). Examine the F statistic (TYPE II, fully adjusted for all other variables in the model).

Find the least significant variable, and eliminate this variable only. THIS IS DONE ONE VARIABLE AT A TIME.

2) Fit the full model, minus the eliminated variable (p-2 variables). Again find the least significant and eliminate this variable.

3) Repeat until all variables in the model are significant.

Sometimes, a variable added to a model by FORWARD selection is not significant at some later point, after other variables are in the model.

STEPWISE selection (FORWARD selection with a backward glance)

This selection proceeds as does FORWARD, but after each addition all variables in the model are examined to insure that they remain significant. If not significant, they are eliminated.

Additional Considerations:

1) At what level do we enter a variable? When do we stop?

Sometimes we want to examine variables at a level lower than α=0.05, perhaps α=0.10 or α=0.15. Do we want only significant variables, or do we want to identify variables which may be important even if not significant? Also, these may become significant later as MSE is reduced.

SAS PROGRAMMING

```
PROC REG;
MODEL Y = X1 X2 X3 X4 X5 X6 / SELECTION=FORWARD;
```

Selection options: FORWARD, BACKWARD, STEPWISE, NONE (full model)
These do the selection techniques discussed. Additional options working with these are;

SLENTRY = alpha value, level of entry for stepwise and forward

SLSTAY = alpha value, level of remaining for backward and stepwise

INCLUDE = n, automatically puts the first n variables of the list into the regression (do not remove)

Other selection techniques

MAXR and MINR : instead of examining the F statistic, these approaches do stepwise regression by evaluating the increase in R^2 for adding and swapping variables to get the “best one, ... two, ..., three" factor models.

RSQUARE, ADJRSQ, CP : These are like all possible regressions. These will simply calculate the R_p^2, R_a^2 and C_p statistic for all models to find the best ones.

Additional options

START = n

STOP = n (will stop all methods)

BEST = n

eg. with these options we could determine the best 5 models at each level, starting with 3 factors, ending with 6 factors

Model Validation

Problem : When numerous variables are examined, especially when the number of observations is relatively small, it is possible that a model is developed which predicts very well for the data base, but does not work well for the real world (sometimes called prediction bias).

How to avoid this?

Validation data

The best way is to develop the model, then collect new data and validate the model.

Alternatively, we can set aside part of the database as a validation set.

How large a part to set aside? You generally want 6 to 10 times the number of observations as variables to be examined. This should be enough to develop a reliable model. The remaining variables can be set aside for validation.

If not enough observations are present, try a 50:50 split.

Validation procedure

Various possibilities;

1) Recalculate the model on the validation data and compare the parms

2) Use the model to get $\hat{Y}$ values for the validation data and fit the observed on the predicted to compare. The MSE and the R^2 should be similar

Example: a model to predict salinity on the coast from discharges nearby. Original model examined combinations of 6 tributaries (with lags) on 19 variables, and achieved R^2 of 77%. Model was 2 factors (lagged and unlagged sums), and was planned for use.

Only fit 45% of the variation in 65 new observations.

New examination revealed simpler model fit 58% of the variation (season and sum tribs lagged 1 mo). Stepwise on many variables & combinations came up with a 69% model which needs validation.