Model building

Assoc. Prof. Susan Balaba Tumwebaze

Dr. Thomas L. Odong

Dr. Hellen Namawejje

Model building

- When you have many regressors to include in the model, a subset of the regressors need to be selected.

- Finding the appropriate subset of regressors for the model is called the **variable selection problem**.

- There methods used to select a subset of regressor to be included in the model. These methods are called stepwise-type procedures

Aim of model building

- To determine a model that is simple yet fits best

Regression Equation

- Consider a regression equation stated here;

Y(Growth) =rain fall + soil type + variety + sun light + fertilizer

$Y=\beta_0+\beta_1 x_1+ \beta_2 x_2 \ldots+\beta_6 X_6$

Variable Selection Procedures

- Stepwise Regression
- Forward Selection
- Backward Elimination

Iterative; one independent variable at a time is added or deleted based on the *F* statistic or P-value

---

- Best-Subsets Regression

Different subsets of the independent variables are evaluated

Stepwise selection methods

- There are three methods of model selection namely; *forward selection, backward selection and stepwise selection*

Forward selection

- Begins with the assumption that there are no

  Regressors in the model other than the intercept.

  i.e., y=beta_0

  Criterion, alpha=0.05

- The regressor are inserted in the model one at time.
  - Y=beta_0 + soil type, soil type must have $p<0.05$
- The first regressor selected will be one with the largest correlation with the response.
- Once a variable is in, it stays in.

Forward selection-continued

Step One: The first regressor is entered;

- The first regressor must have $p<0.05$

Step Two: The second regressor entered is one with the largest correlation with the response after adjusting for effect of the first one.

- The second regressor added must also have $p<0.05$
- The same applies to third, …..nth
- Y=beta_0+soil type($p<0.05$)+ fertility ($p<0.05$)+…..+variety ($P<0.05$)

Note: It is done by the Statistical packages automatically

Backward selection

- Backward selection begins with a model that includes all the k regressors.

- Y=beta_0+beta_1X_1+…..+beta_6x_6 (assuming you had 6 variables)

- Criteria alpha <0.05

- Partial F-value for each regressor is computed and compared with F-to-remove (alpha<0.05).

Backward selection-Continued

- If the partial F-value is less that the F-to-remove, the regressor is removed or p-value greater than the set significance level (alpha)
- Once the variable is out, it stays out
- Backward selection terminates when the none of the regressor in the model is not less than F-to –remove (alpha<0.05)
- Y=beta_0+beta_1x_1+beta_2x_2+beta_3x_3+beta_4x_4

Stepwise Regression

- Stepwise regression combines both forward and backward selection process.

- Has two F-statistics (i.e., F-to-enter and F-out)

i.e., alpha<0.05 (enter) and alpha<0.05 (remove)

- Y=beta_0+Rainfall

- Y=beta_0+Rainfall+ soil type

- You can remove Rainfall if adding soil type makes it alpha>0.05

Stepwise -Continued

- Variables included in the model are assessed in both ways.
- Variables removed at one step could be re-entered at a later step
- Rules are set for a variable to enter the model and to be removed from the model using p-values or partial F-statistics
- None of the methods results into the best regressors

Methods for evaluating subset regression models

Methods for evaluating subset regression models

What we consider:

- Choose one with largest R-squared adjusted
- MSE-Smallest
- AIC- Smallest is the best
- Mallow Cp= this must be equal to the number of parameters used in the model. If they are 4 variables, Cp =5, (the 4 variables betas+ 1 intercept=5). A good model has Cp=p=k+1
- Prediction Sum of Squares (PRESS)- the smaller the PRESS value, the better the model's predictive ability.

Methods_continued

Coefficient of multiple determination R-sq

- R-sq= (SSr/Syy)

- A plot of R-sq against p-terms or (p-1) regressor, shows R-sq increasing with increasing P.

- Adjusted R-sq does not increase as more regressors are added to the model, this is preferred to R-sq (see Montgomery and Peck, 1992).

- *When using R-sq or adj R-sq, choose a model with the highest values.*

Appropriate model

- After all the above step, choose the best model
- Test for regression assumptions/validation of assumptions under regression
- Analyze the model

Practical in R

Lets move to R