Multiple Regression
Model Selection in MLR

Model Selection

- **Having a large number of independent variables/predictors, decide which of them to include in the model.**

- **This is the problem of model selection, and it is a difficult one.**

- **Good model selection rests on this basic principle known as Occam's razor:**
  "The best scientific model is the simplest model that explains the observed data."

- **In terms of linear models, Occam's razor implies the Principle of Parsimony:**
  "A model should contain the smallest number of variables necessary to fit the data."
- **Models that include only the variables needed to fit the data are called parsimonious models. Much of the practical work of multiple regression involves the development of parsimonious models.**

- **Adding a new variable to a model can substantially change the coefficients of the variables already in the model.**

Model Selection

- **GOAL: Find a model that explains as much variability in Y as possible with the smallest number of predictors.**

- **NOTE: Models with fewer variables support the principle of parsimony , are easier to interpret and cheaper to use.**

- **STEP1: Consider only those predictors that should have some effect on the response- work with the expert in the application area.**

- **STEP 2: Use a predictor/model selection procedure.**

  - **"Stepwise procedure" use partial F-tests to decide about a predictor,**
  - **Use an overall measure of model quality. For example R^2 or MSE.**

Stepwise Procedures

- **Forward selection;**
- **Backward elimination;**
- **Stepwise regression.**

Characteristics of stepwise procedures

- **All are based on partial F-tests,**
- **All are automated in stat software,**
- **Reasonable for large number of possible predictors,**
- **Do not consider all possible models,**
- **Use user specified criteria for adding/deleting a predictor from the model.**

Forward selection

- **START: Intercept only model**
- **Add variables to the model one at a time. Once "in" a variable stays in the model.**

- **STEP 1. Consider all models with 1 predictor**
  - **Compute partial F-tests (or t-tests) for model coefficients.**
  - **The variable with highest significant F-statistic (or t-stat) stays in the model, say xi.**

- **STEP 2. Consider all models with xi and one additional predictor**
  - **Compute partial F-tests (or t-tests) for model coefficients.**
  - **The variable with highest significant F-statistic (or t-stat) stays in the model, say xj.**

- **ETC.**

- **STOP when can not add any variables. Resulting model called "best"**

Criteria for adding a predictor "highest significant F/t statistic"

- **SIGNIFICANT: p-value less than significance level or F stat larger than specified value Fo.**
- **Fo chosen large for many tests, so Fo must be large for many values of df, often Fo=4 because $F_{1, v, 0.05} \approx 4$ for many values of v.**
- **HIGHEST: consider only significant predictors, add the one with the largest F statistic.**
- **REMEMBER: For partial F-tests about β_i's: $F = t^2$. So, you can also use values of the t-stat to make the decisions. Usually F stats are used in all packages.**

Backward elimination

- **Similar idea to forward selection.**
- **START: A model with ALL predictors.**
  - **Remove variables from the model one at a time.**
  - **Once remove a variable from the model, it stays out of the model.**
- **STEP 1: Consider all models with one variable removed.**
  - **Compute partial F-tests (or t-tests) for model coefficients.**
  - **Eliminate the variable with the smallest F-statistic (or t-stat) out of those with F stat smaller than a specified value F*.**
- **STEP 2. Consider all models with second variable removed. Predictor removed in step 1 stays out.**
  - **Compute partial F-tests (or t-tests) for model coefficients.**
  - **Eliminate the variable with the smallest F-statistic (or t-stat) out of those with F stat smaller than a specified value F*.**
- **ETC.**
- **STOP when can not remove any variables. Resulting model called "best"**

Criteria for eliminating a predictor

- **“Smallest F/t stat among those smaller than Fo”**

- **“Smaller than Fo”: Fo chosen as a critical value for significance test or a specified value that is “small” for many tests. Again, often Fo=4 because $F_{1,\,v,\,0.05} \approx 4$ for many values of v.**

- **SMALLEST: consider only predictors with F< Fo. Eliminate one with smallest F among them.**

Stepwise regression

- **Combines forward selection and backward elimination.**
- **START: with a model**
- **FS step: Add a predictor**
- **BE step: Examine variables previously in the model for possible elimination.**
- **ETC.**
- **STOP: when nothing to add.**
- **In stepwise regression we may replace one variable with several predictors or vice verse'a.**

Overall measures of model quality

- **Adjusted R^2: R_a^2**
- **PRESS statistic**
- **Mallows Cp statistic**
- **S or S^2 (Mean squared error)**

- **These measures of fit can be used to compare all 2k models with k available predictors.**

- **They can be used to compare models with different number of predictors**

- **Drawback: They are practical if the number of predictors k is not very large.**

Adjusted R^2: R_a^2

- R_a^2 is R^2 adjusted for the number of explanatory variables

- Problem with R^2 : add a predictor to a MLR model, then R^2 increases and SSE decreases no matter how little explanatory power that predictor has!

- Solution: Look at R^2 = 1 – SSE/SSy (SSy=SST/n-1) and replace SSE by MSE=SSE/(n-p) and SSy by Sy^2, get adjusted R^2 :

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SSy/(n-1)} = 1 - \frac{MSE}{S_Y^2} = 1 - \frac{n-1}{n-p}\frac{SSE}{S_Y^2}$$

If p increases, then n-p decreases, so (n-1)/(n-p) increases

Independent of the model

Adjusted R^2

- $R_a^2 = 1 - \frac{n-1}{n-p}\frac{SSE}{S_Y^2}$
- **If we add a predictor, then (n-1)/(n-p) increases, and SSE decreases or SSE/SSy decreases.**

- **If predictor has little explanatory power, then**
  - **SSE/SSy decreases “a little”**
  - **Decrease in SSE/SSy can be offset by increase in (n-1)/(n-p) ending in**
  - **Little change in R_a^2**

Thus, R_a^2 is useful for comparing models with different numbers of predictors.

R_a^2 and MSE

- **Maximizing R_a^2 is equivalent to minimizing $MSE=S^2$.**
- **MINITAB has a "best subsets" option for choosing "the best" model**
- **Best subsets algorithm uses R_a^2 as a measure of quality, thus it searches for a model with the largest R_a^2 (smallest MSE).**
- **The "best subsets" process works for at most 20 predictors.**

PRESS Statistic

- **PRESS- Prediction residuals Sum of Squares**
- **Prediction residual i: $e_{(i)} = Y_i - \widehat{Y}_i$, where $\widehat{Y}_i$ is the regression estimate of Y_i (ith fitted value) based on a regression equation computed WITHOUT the ith observation.**
- **$PRESS = \sum_{i=1}^{n} e_{(i)}^2$ is an estimate of the error of new predictions**
- **Minimizing PRESS means having a regression equation which produced the smallest error when making new predictions.**
- **In MINITAB: to get prediction errors $e_{(i)}$ we use leverage statistics hi:**

$$e_{(i)} = \frac{e_i}{1-h_i},$$

where h_i =ith leverage statistic that can be reported as optional output (use Storage button).

Mallows Cp statistic

- **Mallows Cp statistic is designed to achieve a good compromise between**
  - **Explaining as much variability in Y as possible by including all relevant predictors**

  and
  - **Minimizing variance of the resulting estimates , that is minimizing s^2 by keeping the number of predictors small.**

$$C_p = p + \frac{(n-p)(s_p^2 - \widehat{s^2})}{\widehat{s^2}},$$

p=number of predictors in the model,

s_p^2 **=MSE of a model with p predictors,**

$\widehat{s^2}$ **="best" estimate of the true error σ^2 usually estimated by the minimum MSE among ALL 2^k models.**

As "best" model we choose one with the smallest Cp.

Cp is reported in MINITAB in the output of the "best subsets regression".

NOTE: PRESS and Cp methods usually agree as to the "best" model.

THE CHOICE

- **What if R_a^2, PRESS, and Cp are very close for two models?**

- **The scientist/statistician must choose one model: Use common sense and be practical:**
  - **A model with less expensive predictors is more practical than a model with expensive predictors;**
  - **A model with a clear interpretation is better than one that is complex and difficult to interpret.**
  - **Etc.**

Stepwise procedures versus overall quality based selection

- **Overall quality selection:**
  - **(+) compares all models**
  - **(+) flexible selection criteria**
  - **(-) computationally expensive or impossible in practice for many predictors.**

Generally we look at the overall quality measures when possible and practical.

Summary of model selection criteria based on overall quality measures

- **The coefficient of multiple determination.**

  $$R^2 = SSR/SS_Y$$

  Choose models with high R^2. However, R^2 increases every time more predictors are added, regardless of their importance in predicting Y.

- **Adjusted R^2.**

  $$\text{Adj } R^2 = 1 - (n-1)(1 - R^2)/(n-k-1)$$

  Choose models with high adj. R^2. Better suited for model selection than R^2. It increases/decreases only when important/unimportant predictors are added to the model.

- **Residual mean square (mean square error).**

  $$MSE = SSE_k/(n-k-1)$$

  Choose models with low MSE.

- **Mallows' C_p statistic. Choose models with small C_p .**

- **PRESS statistic: choose models with small PRESS statistic.**

Example of Model Selection-class data

Regression Analysis: y versus x1, x2

Forward Selection of Terms

α to enter = 0.25

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	157.381	78.6904	99.59	0.000
x1	1	76.931	76.9311	97.37	0.000
x2	1	45.747	45.7471	57.90	0.000
Error	7	5.531	0.7901		
Total	9	162.912			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.888893	96.60%	95.63%	92.93%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.530	0.671	0.79	0.456	
x1	2.766	0.280	9.87	0.000	1.05
x2	-2.117	0.278	-7.61	0.000	1.05

Regression Equation

y = 0.530 + 2.766 x1 - 2.117 x2

Backward elimination-classdata

Regression Analysis: y versus x1, x2

Backward Elimination of Terms

Candidate terms: x1, x2

-----Step 1-----

	Coef	P
Constant	0.530	
x1	2.766	0.000
x2	-2.117	0.000

S	0.888893
R-sq	96.60%
R-sq(adj)	95.63%
R-sq(pred)	92.93%
Mallows' Cp	3.00

α to remove = 0.1

Regression Equation

$y = 0.530 + 2.766\, x1 - 2.117\, x2$

Best subsets regression: classdata

Best Subsets Regression: y versus x1, x2

Response is y

		R-Sq		R-Sq	Mallows		x	x
Vars	R-Sq	(adj)	PRESS	(pred)	Cp	S	1	2
1	68.5	64.6	69.8	57.2	58.9	2.5317	X	
1	49.4	43.1	108.1	33.6	98.4	3.2106		X
2	96.6	95.6	11.5	92.9	3.0	0.88889	X	X