Contents

Model Selection and Validation (Ch. 9)

1. Overview

The strategy or criteria to select the "best" model depends on several things:

- Objectives or goals of the study
- Previously acquired knowledge
- Availability of data
- Availability of software
- Read Ch.9.1 for more details

2. Model Selection Criteria

Consider a data set with (P – 1) predictors and 1 response variable. Without considering polynomial terms or interactions, we select (*p* – 1) predictors to build a regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

a. Coefficient of multiple determination, R^2

$$R^2 = SSR / SSTO = 1 - SSE/SSTO$$

- For 2 models with the same number of predictors, the one with the larger R^2 (i.e., smaller SSE) is preferred.
- Maximizing R^2 is equivalent to minimizing SSE.
- R^2, by its construction, is guaranteed to increase with the number of parameters (*p*). Hence, it is only appropriate for comparing two models with the same number of parameters.

b. Adjusted-R^2

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

- Want to find model that maximizes adjusted-R^2.
- Maximizing Adjusted-R^2 is equivalent to minimizing MSE.
- Adjust R^2 with the sample size and the number of parameters.

c. Akaike Information Criterion (AIC) and Schwarz Bayesian Criterion (SBC)

$$AIC = n\log_e(SSE) - n\log_e(n) + 2p$$

$$SBC = n\log_e(SSE) - n\log_e(n) + (\log_e(n))\,p$$

- Look for models with smaller AIC or SBC.
- SBC is also known as the Bayesian Information Criterion (BIC).
- There is a trade-off between model fitness (SSE) and complexity (p).
- SBC (BIC) tends to select "smaller" model than AIC does because SBC (BIC) penalizes the model complexity (p) more heavily. Why?

d. Mallows' C (or Mallows' C_p)

$$C_p = \frac{\text{SSE}_p}{\text{MSE}(all\ P-1\ predictors)} - (n-2p)$$

- For prediction purposes, a model is good if $C_p \le p$. Either pick the smallest model (minimal p) for which $C_p \le p$, or pick the model that minimizes C_p and $C_p \le p$.
- In addition, Hocking (1976, Biometrics, 32, p.1-49) suggested for estimation purposes, use $C_p \le 2p - (P-1)$.
- There is a trade-off between possible bias and error variance.
- Need to carefully develop the pool of "all possible predictors".

e. Prediction Sum of Squares (PRESS)

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

where $\hat{Y}_{i(i)}$ is the predicted value of the i-th observation when regression model is estimated without the i-th observation. $e_{i(i)} = Y_i - \hat{Y}_{i(i)}$ is called the "deleted residual."

- A model with small PRESS is preferred.
- Look at SS of the observed Y vs. predicted.
- Not available in all software packages.

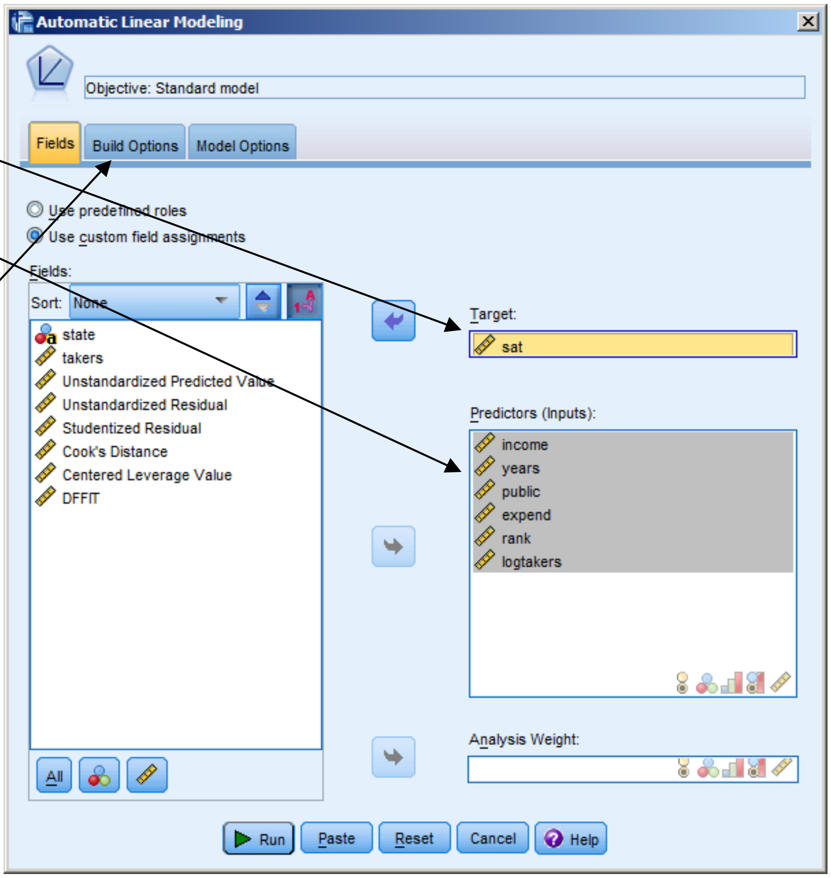
3. Automatic Search Procedures

a. Best Subset Algorithm

- Given the number of parameters (predictors), find the best in the subset by fitting all possible models with the same number of parameters. Then evaluate the criteria (see above) of those models.
- The Best Subset Algorithm is computationally intensive. It is not available in all software packages.
- Newer versions of SPSS may have this function built in:
  Analyze → Regression → Automatic Linear Modeling…

Add your response variable (Y) under Target, and all predictors (X's)

Click Build Options



Objective: Standard model

○ Use predefined roles

● Use custom field assignments

Fields:

Sort: None

- state
- takers
- Unstandardized Predicted Value
- Unstandardized Residual
- Studentized Residual
- Cook's Distance
- Centered Leverage Value
- DFFIT

Target: sat

Predictors (Inputs):

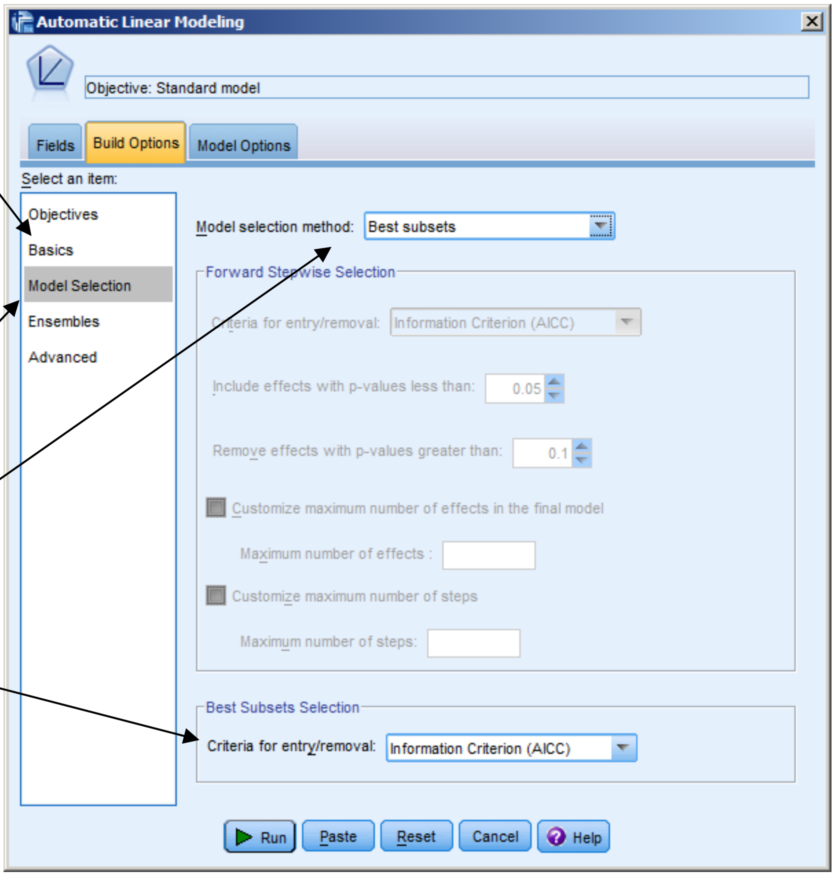
- income
- years
- public
- expend
- rank
- logtakers

Analysis Weight:

Select Basics and un-check "Automatically prepare data".

Select Model Selection.

Select "Model Selection methods" and selection criteria.



Objective: Standard model

Model selection method: Best subsets

Forward Stepwise Selection

Criteria for entry/removal: Information Criterion (AICC)

Include effects with p-values less than: 0.05

Remove effects with p-values greater than: 0.1

□ Customize maximum number of effects in the final model

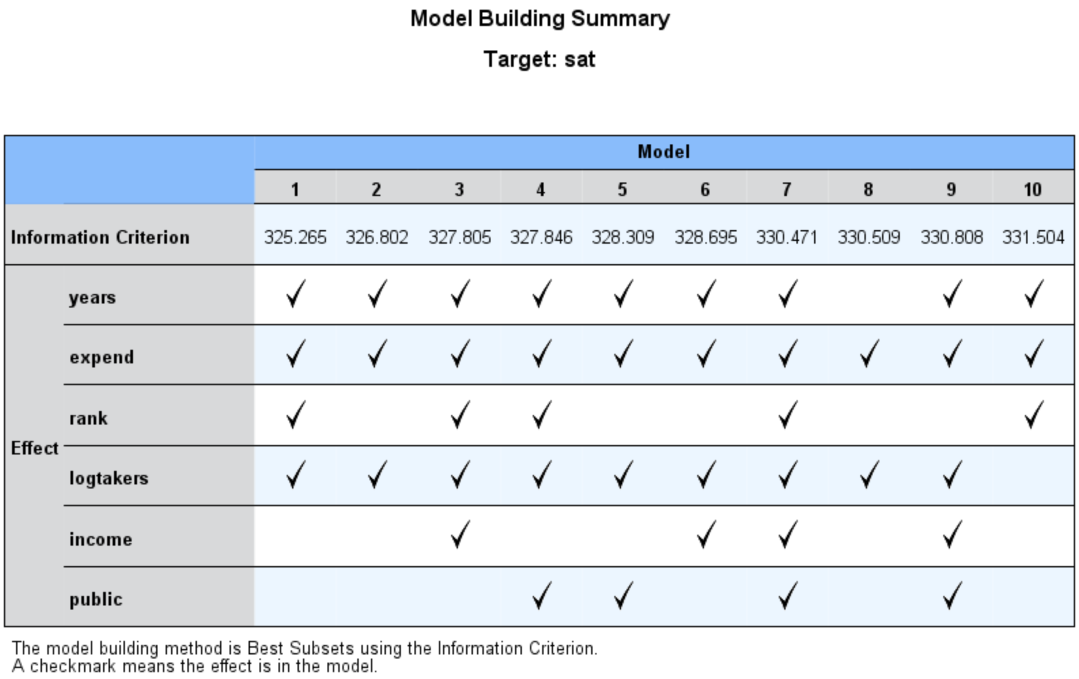
Maximum number of effects :

□ Customize maximum number of steps

Maximum number of steps:

Best Subsets Selection

Criteria for entry/removal: Information Criterion (AICC)



Model Building Summary

Target: sat

		Model									
		1	2	3	4	5	6	7	8	9	10
Information Criterion		325.265	326.802	327.805	327.846	328.309	328.695	330.471	330.509	330.808	331.504
Effect	years	✓	✓	✓	✓	✓	✓	✓		✓	✓
	expend	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	rank	✓		✓	✓			✓			✓
	logtakers	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	income			✓			✓	✓		✓	
	public				✓	✓		✓		✓	

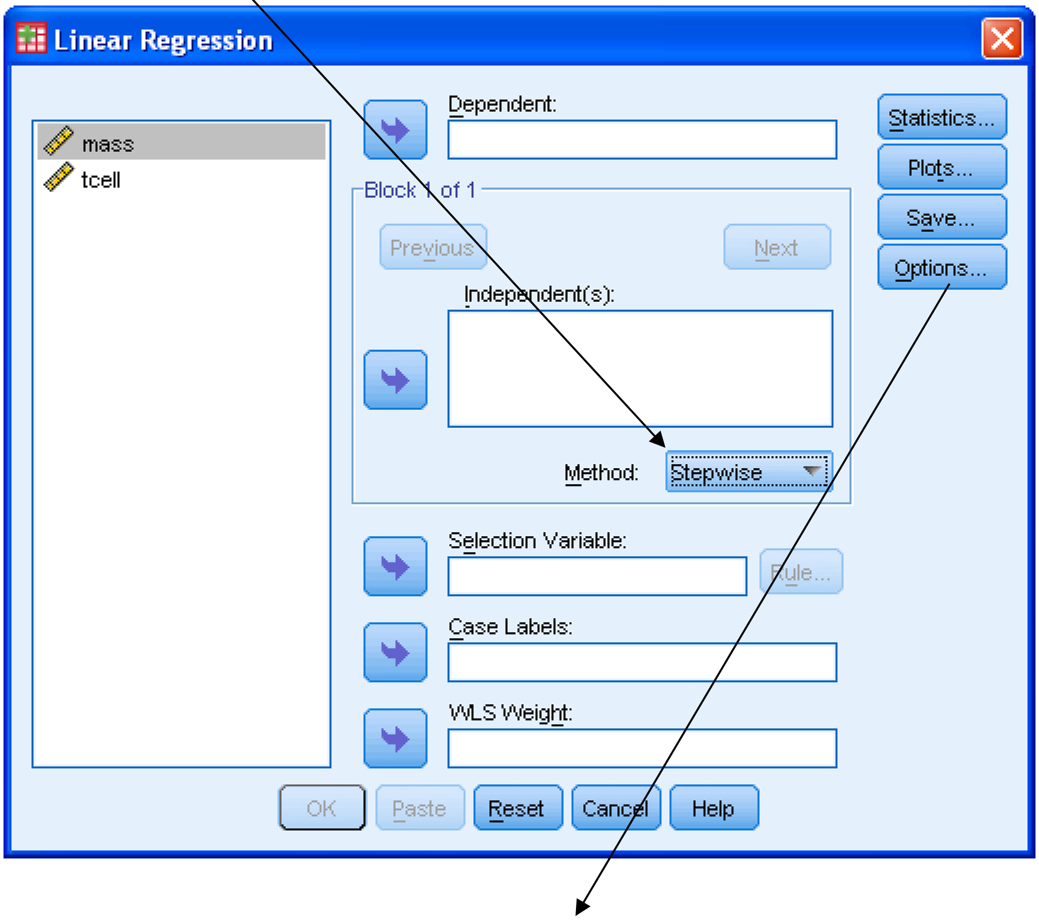
The model building method is Best Subsets using the Information Criterion.
A checkmark means the effect is in the model.

(The above output is for demonstration purposes. We did not use this data set.)

b. Stepwise selection

- Forward (selection) + Backward (elimination)
- One predictor gets selected or removed at each step.
- Can use different criteria (SPSS uses F, R can use AIC or BIC, etc.).
- Popular choice due to intuitive computational procedure.
- Cons: It cannot guarantee that the resulting model is the "best."

In SPSS, use “Stepwise” for variable selection. Check “Options…” to change criterion.



Linear Regression

- mass
- tcell

Dependent:

Block 1 of 1

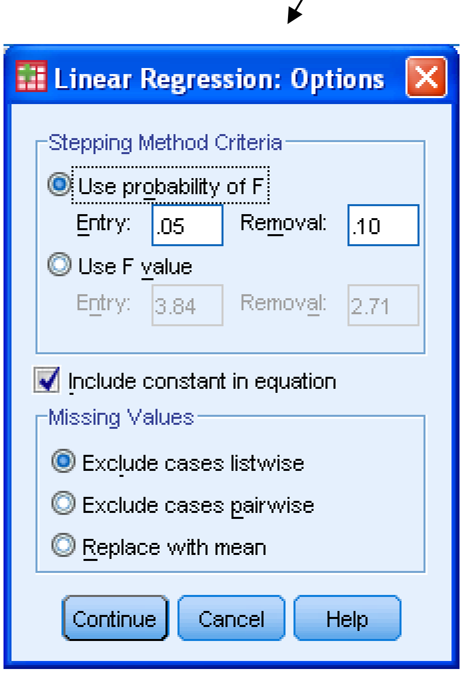
Independent(s):

Method: Stepwise

Selection Variable:

Case Labels:

WLS Weight:



Linear Regression: Options

Stepping Method Criteria

● Use probability of F

Entry: .05 Removal: .10

○ Use F value

Entry: 3.84 Removal: 2.71

☑ Include constant in equation

Missing Values

● Exclude cases listwise

○ Exclude cases pairwise

○ Replace with mean

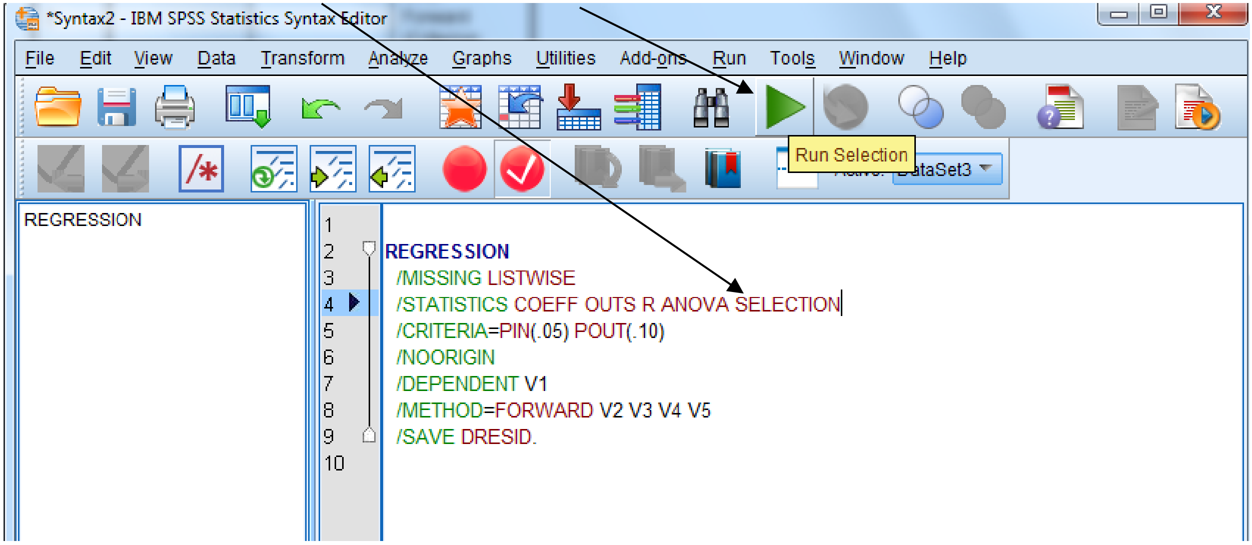
4. Model validation

- ➢ Collect more data to validate the estimated model.
- ➢ Cross-Validation: Split the current data set into training set and validation set.
- ➢ Validate the model with theories, empirical results, simulations, etc.

5. Practice (no separate Lab note for this topic)

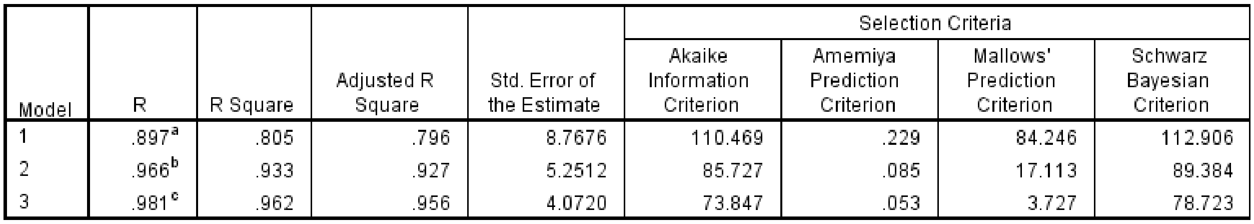
Refer to Problem 9.18 (p. 379)
In part b, do the following:
(1) Find/Compute R^2, adjusted-R^2, Mallow's C_p, AIC, SBC(i.e., BIC), PRESS for the final model.
(2) Find/Compute R^2, adjusted-R^2, Mallow's C_p, AIC, SBC(i.e., BIC), PRESS for the regression model that includes the first order terms of all 4 predictors.
Compare and comment on your results.

Note that in SPSS, you can get the software compute Mallow's C_p, AIC and SBC if you use syntax instead of "point-and-click." (Thanks to Eden Waller, Summer 2013.)
After you set up the variables in the Regression window, click "Paste"! This will open the syntax editor. Add "SELECTION" to the code and run it.



```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA SELECTION
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT V1
  /METHOD=FORWARD V2 V3 V4 V5
  /SAVE DRESID.
```

Model Summary[d]



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Selection Criteria			
					Akaike Information Criterion	Amemiya Prediction Criterion	Mallows' Prediction Criterion	Schwarz Bayesian Criterion
1	.897[a]	.805	.796	8.7676	110.469	.229	84.246	112.906
2	.966[b]	.933	.927	5.2512	85.727	.085	17.113	89.384
3	.981[c]	.962	.956	4.0720	73.847	.053	3.727	78.723