7. Model Selection

Choosing the correct linear regression model can be difficult. After all, the real world and how it works is complex. Trying to model it with only a sample does not make it any easier. As George E P Box said: *"All models are wrong, but some are useful"*.

Here, we will review some common statistical methods for selecting models, discuss complications we may face and provide some practical advice for choosing the best regression model.

A process for selecting models might start with all the variables in the data set and eliminate the least significant terms, based on p-values. For a small data set, a final model can be developed in a reasonable amount of time. If you start with a large model, however, eliminating one variable at a time can take an extreme amount of time. You would have to continue this process until only terms with *p*-values lower than some threshold values, such as 0.10 or 0.05, remain.

Selecting Candidate Model

Candidate models can be identified by any combinations of the following criteria.

- Knowledge of the subject-matter.
- Information collected from data exploration.
- Residuals plots and analyses to evaluate model assumptions and model fit.
- Automatic selection criteria available in statistical software, such as R^2, adjusted R^2, Mallows C_p, stepwise selections and Information Criteria's such as AIC and SBC.

Note: SPSS does not provide these selection criteria through the **Regression** dialogue box. But they can be requested through command syntax. In the **Regression** dialogue box, do not click **OK**, instead click **Paste**, add SELECTION to /STATISTICS subcommand and in **Run** menu click on **All**.

Mallows C_p

This criterion is calculated by:

$$C_p = p + \frac{\left(MSE_p - MSE_{full}\right)\left(n-p\right)}{MSE_{full}}$$

In the formula above, p is the number of parameters in the model being evaluated (the number of variables plus one). MSE_p is the mean squared error for the model with p parameters. MSE_{full} is the mean squared error for the full model used to estimate the true residual variance. Mallows C_p is a simple indicator of model misspecification. When C_p is much larger than p, it usually indicates model under-specification.

To see the details of the rule, another representation of the computational formula for C_p is $C_p = \frac{SSE_p}{MSE_{full}} + 2p - n$, where SSE_p is the error (residual) sum of squares for the model with p-1 variables and MSE_{full} is the error (residual) mean square when using all the independent variables .

When the model is correctly specified the residual sum of squares is an unbiased estimate of $(n-p)\sigma^2$, and C_p is an unbiased estimate of $\frac{(n-p)\sigma^2}{\sigma^2} + 2p - n = p$. So C_p is approximately equal to p when the model is correctly specified. When important variables are omitted from the model, the residual sum of squares will be increased by the amount of variability that would be explained by those terms had they been included in the model. Therefore C_p will increase and $C_p > p$.

Information Criteria

- **AIC**, Akaike's Information Criterion : $\text{AIC} = n\ln\left(\frac{SSE}{n}\right) + 2k$
- **AICC**, corrected AIC: $\text{AICC} = \text{AIC} + \frac{2k(k+1)}{n-k-1}$
- **SBC** (**BIC**), Schwarz's Bayesian Criterion: $\text{BSC} = n\ln\left(\frac{SSE}{n}\right) + k\ln(n)$

All these criteria, for which smaller values indicate better models, are commonly suggested for model selection among nested models. Although some statisticians believe this is only a myth.

Stepwise Selections

The principle objective of this method is to select the best sub-set of regressors that meet the selection criteria. The selection criteria, however, can differ from one statistical software to another. Here we describe the method used by SPSS, however, the process is the same for most statistical software, including R.

Enter: A default procedure for variable selection in which all variables in a block are entered in a single step.

Forward Selection: It starts with an empty model. The first variable considered for entry into the equation is the one with the largest positive or negative correlation with the dependent variable. This variable is entered into the equation only if it satisfies the criterion for entry. When the first variable is entered, the independent variable not in the equation that has the largest partial correlation is considered next. The procedure stops when there are no variables that meet the entry criterion.

Backward Elimination: All variables are entered into the equation and then sequentially removed. The variable with the smallest partial correlation with the dependent variable is considered first for removal. If it meets the criterion for elimination, it is removed. After the first variable is removed, the variable remaining in the equation with the smallest partial correlation is considered next. The procedure stops when there are no variables in the equation that satisfy the removal criteria.

Stepwise Selection: This is similar to forward selection in that it starts with an empty model and incrementally builds a model one variable at a time. However, the method differs from forward selection in that variables already in the model do not necessarily remain. At each step, the independent variable not in the equation that has the smallest p-value of F (significance level) is entered, if that probability is sufficiently small. Variables already in the regression equation are removed if their probability of F becomes sufficiently large. The method terminates when no more variables are eligible for inclusion or removal.

Remove: A procedure for variable selection in which all variables in a block are removed in a single step.

All variables must pass the tolerance criterion to be entered in the equation, regardless of the entry method specified. The default tolerance level is 0.0001. Also, a variable is not entered if it would cause the tolerance of another variable already in the model to drop below the tolerance level.

All independent variables selected are added to a single regression model. However, you can specify different entry methods for different subsets of variables. To add a second block of variables to the regression model, click Next.

Note: Stepwise selections (Forward, Backward, and Stepwise), in general, have some serious shortcomings. The significance values in output are based on fitting a single model. Therefore, the significance values are generally invalid when a stepwise method is used. Simulation studies evaluating variable selection techniques has also found the following problems:

- The degree of collinearity among the predictor variables affected the frequency with which authentic independent variables found their way into the final model.
- The number of candidate predictor variables affected the number of noise variables that gained entry to the model.
- The size of the sample was of little practical importance in determining the number of authentic variables contained in the final model.

One recommendation is to use the variable selection methods to create several candidate models and then use subject-matter knowledge to select the variables that result in the best model within the context of the problem. Therefore, you are simply using these methods as a useful tool in the model-building process.

Are the *p*-Values Correct?

Statisticians give warnings and cautions about the over-interpretation of *p*-values from models chosen using any automated variable selection technique. Re-fitting many sub-models in terms of an optimum fit to the data distorts the significance levels of conventional statistical tests. However, many researchers and users of statistical software neglect to report that the models they ended up with were chosen using automated methods. They report statistical quantities such as standard errors, confidence limits, *p*-values, and R^2 as if the resulting models were entirely pre-specified. These inferences are inaccurate, tending to err on the side of over-stating the significance of predictors and making predictions with overly optimistic confidence. This problem is very evident when there are many iterative stages in model building. When there are many variables and you use stepwise selection to find a small subset of variables, inferences become less accurate.

Although, the problem of biasing the inferences can be avoided by pre-specifying the model, when there is a large number of variables, however, pre-specifying the model is not feasible. Consequently, the *p*-values calculated in the model selection techniques are not *p*-values in the traditional hypothesis-testing context. Instead, they should be viewed as indicators of relative importance among variables. Because the biased *p*-values over-state the significance of the predictor variables, the traditional cut-off of .05 is not very useful unless the sample size is small (30-50). For large sample sizes, much smaller *p*-values are required to imply that the data provide evidence for the effect of interest.

Specification Bias

The discussion in the previous section is closely related to a problem denoted as specification (or in fact misspecification) bias. An implicit assumption underlying the regression model is that the estimated model is the *true* model. Therefore, an over-specified or an under-specified estimated model is not reliable. Here, we discuss the latter case of *Omitted Variables* in details.

Consider the following equations:

(1) $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

(2) $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}$

where, both ***X*** and ***Z*** are matrices of independent variables. The estimated equations are:

$$\hat{\boldsymbol{y}} = \boldsymbol{Xb}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X\hat{\beta}} + \boldsymbol{Z\hat{\gamma}}$$

Now suppose model (2) is the true model, while we wrongly estimate model (1). The LS estimator is therefore:

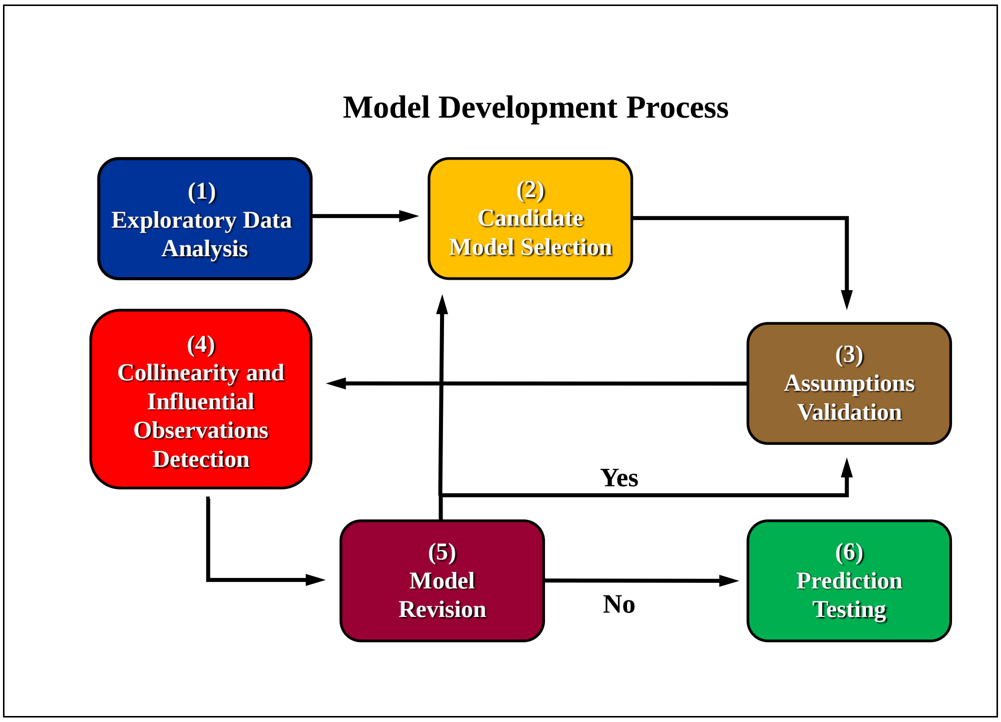
$$\begin{aligned} \boldsymbol{b} &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'Z\gamma} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'\varepsilon} \end{aligned}$$

$$E(\boldsymbol{b}) = \boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'Z\gamma}$$

$$Var(\boldsymbol{b}) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\, Var(\boldsymbol{\varepsilon})\, \boldsymbol{X}(\boldsymbol{X'X})^{-1} = \sigma^2 (\boldsymbol{X'X})^{-1}$$

That is, the LS estimator ***b*** becomes biased, although the variance of ***b*** is correctly calculated.

On the other hand, if the true model is (1) but an over-specified model (2) is estimated, it can be shown that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, ie the LS estimator remains unbiased but now $Var(\hat{\boldsymbol{\beta}}) \geq Var(\boldsymbol{b})$.


Model Development Process
(1) Exploratory Data Analysis
(2) Candidate Model Selection
(3) Assumptions Validation
(4) Collinearity and Influential Observations Detection
(5) Model Revision
(6) Prediction Testing
Yes
No

(1) Exploratory Data Analysis: This step includes the use of descriptive statistics, graphs, and correlation analysis to identify those variables that might be useful in the regression model.

(2) Candidate Model Selection: This step uses the information gathered from the exploratory data analysis to identify one or more candidate models. Potential models can be evaluated by comparing the adjusted coefficients of determination, Mallows C_p statistic and information criteria. You can also produce the plot of residuals versus the predicted values plots of the residuals versus the regressors to assess the model fit.

(3) Model Assumptions Validation: This step includes graphs of the residuals versus the predicted values. It also includes tests for normal residuals, constant variances, and independent observations.

(4) Collinearity and Influential Observation Detection: The presence of multicollinearity can be detected by the use of the VIF statistic, condition indices, and variance proportions. Influential observations can be detected by examining plots of Studentised residuals, Cook’s D statistics, DFFITS statistics, DFBETAS statistics, covariance ratio statistics, leverage statistics and partial leverage plots.

(5) **Model Revision:** If steps (3) or (4) indicate the need for model revision, generate a new model that is more appropriate. For example, you might need to transform the response variable to meet the equal variance assumption. Based on the nature of the refinement, you might need to return to step (1) or (2) to identify new candidate models for the transformed response variable.

(6) **Prediction Testing:** This final step is to evaluate the model's predictive capability with data not used to build the model. In other words, you would build the model with part of your data and use the remainder of the data to determine how well the model fits the data.

Example 7.1

Based on data gathered from a social survey, we will investigate the effect of the respondent's **age**, **sex**, **education** and **spouse's education** on the **occupational prestige** score, using stepwise selection method.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Highest Year of School Completed	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	Age of Respondent	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	Respondent's Sex	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: R's Occupational Prestige Score

The **Variables Entered/Removed** table tells us that at the first stage, the variable **education** was entered into the model; then **age** was added; finally, **sex** was added to the previous two. At no point was any variable removed due to becoming insignificant in the model, and **spouse's education** was not found to satisfy the criteria for inclusion in the model.

The inclusion and removal thresholds are set at 0.05 for inclusion and 0.1 for removal, but these can be changed if required by clicking on the **Options** button.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.553[a]	.306	.305	10.940
2	.571[b]	.326	.324	10.786
3	.574[c]	.330	.327	10.763

a. Predictors: (Constant), Highest Year of School Completed

b. Predictors: (Constant), Highest Year of School Completed, Age of Respondent

c. Predictors: (Constant), Highest Year of School Completed, Age of Respondent, Respondent's Sex

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	39799.799	1	39799.799	332.515	.000[b]
	Residual	90368.457	755	119.693		
	Total	130168.256	756			
2	Regression	42455.767	2	21227.884	182.481	.000[c]
	Residual	87712.489	754	116.330		
	Total	130168.256	756			
3	Regression	42945.906	3	14315.302	123.586	.000[d]
	Residual	87222.350	753	115.833		
	Total	130168.256	756			

In the **Model Summary** table, we can see how the ***R*²** and **adjusted *R*²** are increasing with each adjustment to the variables included in the model, indicating a better fit. The **ANOVA** table also shows that the combination of variables in each model has a significant **Linear Relationship** with **occupational prestige**.

Coefficients

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	11.585	1.829		6.335	.000
	Highest Year of School Completed	2.474	.136	.553	18.235	.000
2	(Constant)	3.627	2.454		1.478	.140
	Highest Year of School Completed	2.639	.138	.590	19.104	.000
	Age of Respondent	.125	.026	.148	4.778	.000
3	(Constant)	6.646	2.855		2.328	.020
	Highest Year of School Completed	2.620	.138	.586	18.963	.000
	Age of Respondent	.119	.026	.140	4.511	.000
	Respondent's Sex	-1.624	.789	-.062	-2.057	.040

Excluded Variables[a]

						Collinearity Statistics
Model		Beta In	t	Sig.	Partial Correlation	Tolerance
1	Age of Respondent	.148[b]	4.778	.000	.171	.938
	Respondent's Sex	-.078[b]	-2.578	.010	-.093	.999
	Highest Year School Completed, Spouse	.032[b]	.846	.398	.031	.629
2	Respondent's Sex	-.062[c]	-2.057	.040	-.075	.984
	Highest Year School Completed, Spouse	.053[c]	1.398	.163	.051	.622
3	Highest Year School Completed, Spouse	.054[d]	1.441	.150	.052	.621

Looking at the table of **Excluded Variables**, we can see how, at each stage, the next variable to be included in the model is selected. At the first model stage, it is **age** that is most significant, and so in the **Coefficients** table we find **age** in the second model stage. Both **education** and **age** remain significant in the model, so no variables are chosen for removal. This leaves **spouse's education** and **sex** as the excluded variables at the second model stage, and now **sex** is the more significant of the two; it is inserted in the third model stage. Once again, all variables in the model remain significant, but the excluded variable, **spouse's education**, has not satisfied the 5% inclusion criteria (the significance is 0.150) and so SPSS exits the procedure, having found the best fitting model from these variables.

The final model is therefore displayed as **Model 3** in the **Coefficients** table. Therefore the estimated equation is:

$$\textbf{Occupational prestige} = 6.646 + (2.620 \times \textbf{educ}) + (0.119 \times \textbf{age}) - (1.624 \times \textbf{sex})$$

Practical Week 7

In week 3 you used **Bodyfat.sav** data set to predict **PctBodyFat** as a function of the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.

Use stepwise regression methods to select a candidate model, trying FORWARD, STEPWISE and BACKWARD.

Calculate appropriate criteria to select the best model.