Chapter 8

Building the Regression Model I: Selection of Predictor Variables

We take a first look at the first step in building a regression model, at selecting a "best" subset of predictors from a large group of predictors.

8.1 Overview of Model–Building Process

There a number of steps in building an adequate regression model.

1. *Collect the data.*

   (a) *Controlled Experiment or Observed Study?*
   The collection of data depends very much on whether you (the statistician) planned it out (controlled experiment[1]) or whether this data was given to you and you are trying to make the best of it (observed study). Typically, more information can be derived from a controlled experiment, rather than an observed study, because you can set key quantities, such as sample size, or design the experiment in such a way as to anticipate problems related to multicollinearity or nonlinearity or non–normality, for example.

   (b) *Data diagnostics*[2].
   However the data is collected, diagnostic procedures such as scatter plots and correlation matrices should be applied to the data to find out not only if there are problems such as outliers, non–constant variance or multicollinearity in the data, but also to provide direction, if necessary (in

[1] Although the techniques we will be looking at in STAT 512 can be used in controlled experiments, they tend to used more for observed studies. The central technique discussed in STAT 512, linear regression, in general, tends to be used for observed studies. The central technique discussed in STAT 514, analysis of variance, in general, tends to be used controlled experiments.

[2] We looked at some diagnostic procedures previously in chapters 4 and 6.

observed studies), as to what model (what explanatory variables) to use to fit the data.

(c) *Data remedial measures*[3].
Remedial measures such as transforming the data or (carefully) eliminating outliers should be used to try and fix any problems identified in the previous data diagnostic stage. Another round of data diagnostics should then be applied to the revised data set.

2. *Develop a model to fit the data.*

(a) *Reduction of explanatory variables*[4].
In observed studies[5], it is often necessary to reduce the number of explanatory variables to a more manageable (three of four, say) number. After this is done, there is often several models with different explanatory variables to choose from.

(b) *Model/Data diagnostics*[6].
Diagnostic procedures such as residual plots, scatter plots and correlation matrices should be used to determine not only how well the model(s) fit(s) the data but also, if necessary (in observed studies), to choose a "best" model.

(c) *Model/Data remedial measures*[7].
Remedial measures such as introducing curvature or interaction effects, or transforming the data should be used to try and fix any problems identified in the previous model/data diagnostic stage. Another round of model/data diagnostics should then be applied to the revised model.

This chapter describes the different procedures available to select a "good" subset of predictor variables; in other words, topic 2(a) listed above.

8.2 Surgical Unit Example

This is an interesting study where an initial look at the data is conducted and where residual plots are used to check for interaction variables.

[3]We looked at some remedial procedures previously in chapters 4 and 6.

[4]We look at this topic in this chapter.

[5]Typically, this is not often done in controlled experiments because in this case, the explanatory variables have been identified in the initial part of the analysis.

[6]We look at this topic in the next chapter, chapter 9.

[7]We look at this topic in chapter 10.

8.3 All–Possible–Regression Procedure for Variables Reduction

SAS program: att7-8-3-read-all-pos-reg

There are two ways of identifying the "best" possible subset of predictor variables from a larger groups of candidate predictor variables.

1. The all–possible–regression procedures considers *all* possible subsets and then choose a few "good" sets from all these subsets. This "parallel" method is used when there is medium number (less than 10, say) number of candidate predictor variables under consideration.

2. The stepwise procedures add and/or delete candidate predictors until one "best" subset of predictor variables is achieved. This "serial" method is used when there is large (between 40 and 60, say) number of candidate predictor variables under consideration.

We look at the all–possible–regression procedure in this section. We use four criteria to decide how to reduce the candidate predictor variables to a "good" subset: R_p^2, R_a^2, C_p and $PRESS_p$.

Exercise 8.3 (All–Possible–Regression Procedure for Variables Reduction)

illumination, X_{i1}	9	7	11	16	21	19	23	29	31	33
noise, X_{i2}	15	20	17	22	24	26	28	30	29	37
eyesight, X_{i3}	3	2	3	2	4	5	5	4	5	3
ability to read, Y	70	70	75	88	91	94	100	92	90	85

A table of the R_p^2, R_a^2, C_p and $PRESS_p$ values with plots for R_p^2, R_a^2 and C_p for all of the possible candidate regression models are given below.

X variables in model	R_p^2	R_a^2	C_p	$PRESS_p$
X_1	0.4790	0.4139	3.5454	810.58160
X_2	0.4180	0.3452	4.6637	1173.60069
X_3	0.5307	0.4721	2.5983	771.22031
X_1, X_2	0.4791	0.3303	5.5434	1335.76975
X_1, X_3	0.6529	0.5538	2.3589	741.36495
X_2, X_3	0.5307	0.5789	2.0011	916.64877
X_1, X_2, X_3	0.6725	0.5088	4.0000	1108.46803

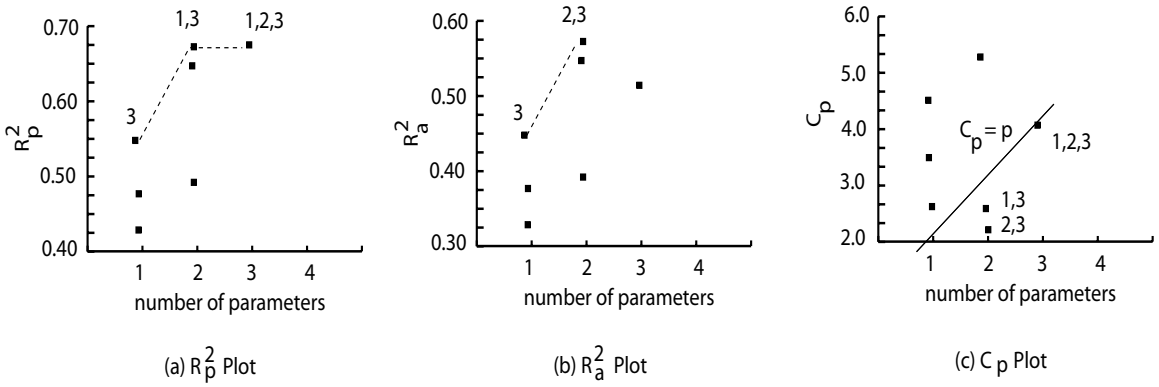

Figure 8.1 (R^2_p, R^2_a and C_p Plots)

1. *What Are We Trying To Do?*
   The best fitting model is *most likely* (choose one)

   (a) $\hat{Y} = 52.04 + 0.02X_1 + 0.62X_2 + 4.88X_3$

   (b) $\hat{Y} = 51.64 + 0.65X_2 + 4.92X_3$

   (c) $\hat{Y} = 70.01 + 0.77X_1$

   because this model has the most predictor variables (X_1, X_2 and X_3) to describe the data. Is it possible that one of the other two *simpler* models listed here (or, indeed, any of the other models with fewer predictor variables), fits the data "almost" as well, but with fewer predictor variables?

2. R^2_p *Criterion, Figure (a).*
   The R^2_p criterion[8] is given by

$$R^2_p = 1 - \frac{SSE_p}{SSTO} = \frac{SSR}{SSTO}$$

   Large values of the coefficient of multiple determination, R^2_p, indicate that the model is a *good* fit to the data. Looking at the table and figure above, two appropriate models are (pick *two*) $\boldsymbol{X_3}$ / $\boldsymbol{X_1, X_3}$ / $\boldsymbol{X_1, X_2, X_3}$

3. R^2_p *Criterion, Revisited.*
   According to the R^2_p criterion, the model (X_1, X_3) is an appropriate one; in other words, the model (*choose none, one or more!*)

   (a) $\hat{Y} = 60.23 + 0.47X_1 + 4.42X_3$

   (b) $(1, 3)$

[8]Instead of the R^2_p criterion, it is possible to equivalently use the SSE_p criterion where large values indicate a (circle one) **poor** / **good** fit since the two are inversely proportional to one another.

(c) $(2, 3)$

4. R_a^2 *Criterion, Figure (b).*
   The R_a^2 criterion[9] is given by

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SSTO} = 1 - \frac{MSE_p}{\frac{SSTO}{n-1}}$$

   Large values of the adjusted coefficient of multiple determination, R_a^2, indicate not only that the model is a *good* fit to the data but also the increasing number of parameters is taken into account[10]. In this case, only one model seems appropriate: (circle one) $\boldsymbol{X_1, X_3}$ / $\boldsymbol{X_2, X_3}$ / $\boldsymbol{X_1, X_2, X_3}$

5. C_p *Criterion, Figure (c).*
   The C_p criterion[11] is given by

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

   Small values of C_p which are also *close* to the $C_p = p$ line (small bias, $E\{\hat{Y}\} = \mu$) indicate not only that the model is a *good* fit to the data but also the fit is *unbiased*. In this case, two models seem appropriate:
   (circle two) $\boldsymbol{X_1, X_3}$ / $\boldsymbol{X_2, X_3}$ / $\boldsymbol{X_1, X_2, X_3}$

6. $PRESS_p$ *Criterion.*
   The $PRESS_p$ criterion[12] is given by

$$PRESS_p = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2$$

   Small values of $PRESS_p$ indicate that not only is the model is a *good* fit to the data, but also the fit accounts for the ith observation missing. In this case, looking at the table above, two models seem appropriate:
   (circle two) $\boldsymbol{X_3}$ / $\boldsymbol{X_1, X_3}$ / $\boldsymbol{X_1, X_2, X_3}$

7. *Comparing the Criteria.*
   Match the criterion with the most appropriate description.

[9]Instead of the R_a^2 criterion, it is possible to equivalently use the MSE_p criterion where large values indicate a **poor** fit since the two are inversely proportional to one another.

[10]As the number of parameters, p, increases, R_a^2 decreases; in other words, R_a^2, unlike R_p^2, accounts for an increase in the number of parameters.

[11]The criterion C_p is an estimate of the total mean squared error divided by the the true (unbiased) error variance. In other words, the C_p criterion not only measures the fit of the model, but also if it is unbiased or not.

[12]The $PRESS_p$ criterion is an estimate of the total squared *prediction* errors (errors based on fitted values missing the ith observation).

Criterion	Description
$\boldsymbol{R_p^2}$	(a) correlation between data and model
$\boldsymbol{R_a^2}$	(b) correlation, adjusted for number of parameters
$\boldsymbol{C_p}$	(c) standardized *SSE* and accounts for bias
$\boldsymbol{PRESS_p}$	(d) *SSE* and accounts for outliers

Criterion	R_p^2	R_a^2	C_p	$PRESS_p$
Description				

8. *More comparing the Criteria.*
   The four criteria used in the all–regression–procedure, R_p^2, R_a^2, C_p and $PRESS_p$, will (choose one) **always** / **sometimes** lead to the *same* collection of "best" reduced models. In this case, although the four criteria give collections of different models, they all seem to agree the single best reduced model to be (circle two) $\boldsymbol{X_2, X_3}$ / $\boldsymbol{X_1, X_3}$ / $\boldsymbol{X_1, X_2, X_3}$

9. *All–Regression–Procedure Chooses Between Additive Models.*
   Possible candidate models that the all–regression–procedure chooses from, include (choose one)

   (a) $\hat{Y} = 69.50 + 0.75X_1 + 0.04X_2$

   (b) $\hat{Y} = 51.64 + 0.65X_2 + 4.92X_1X_3$

   (c) $\hat{Y} = 70.01 + 0.77X_1^2$

8.4 Forward Stepwise Regression and Other Automatic Search Procedures for Variables Reduction

SAS program: att7-8-4-read-forward-step

We look at stepwise procedures add and/or delete candidate predictors until one "best" subset of predictor variables is achieved.

1. The *forward stepwise* procedure add and/or delete candidate predictors until one "best" subset of predictor variables is achieved.

2. The *forward* procedure adds (but does not delete) candidate predictors until one "best" subset of predictor variables is achieved.

3. The *backward* procedure deletes (but does not add) candidate predictors until one "best" subset of predictor variables is achieved.

Exercise 8.4 (Forward Stepwise Regression and Other . . .)

illumination, X_{i1}	9	7	11	16	21	19	23	29	31	33
noise, X_{i2}	15	20	17	22	24	26	28	30	29	37
eyesight, X_{i3}	3	2	3	2	4	5	5	4	5	3
ability to read, Y	70	70	75	88	91	94	100	92	90	85

1. *Forward Stepwise.*
   From SAS, the forward stepwise procedure chooses the model
   (circle one) $\boldsymbol{X_3}$ / $\boldsymbol{X_2, X_3}$ / $\boldsymbol{X_1, X_2, X_3}$

2. *Steps of Forward Stepwise.*

   (a) The
   $$F^* = \frac{MSR(X_k)}{MSE(X_k)}$$
   statistic for each simple linear regression (for X_1, X_2 and X_3) are calculated and the most significant one is chosen to enter the model.
   Variable (choose one) $\boldsymbol{X_1}$ / $\boldsymbol{X_2}$ / $\boldsymbol{X_3}$
   is entered (added) and so the model is
   $$\hat{Y} = 62.27 + 6.45X_3$$

   (b) The partial F^* statistic
   $$F^* = \frac{MSR(X_k|X_3)}{MSE(X_3, X_k)}$$
   for all regression models with pairs of variables that include X_3 (for X_1, X_3 and X_2, X_3 only) is determined. The most significant, one, (X_2, X_3), is chosen; in other words,
   variable (choose one) $\boldsymbol{X_1}$ / $\boldsymbol{X_2}$ / $\boldsymbol{X_3}$
   is entered (added) and so the model now becomes
   $$\hat{Y} = 51.64 + 0.65X_2 + 4.92X_3$$

   (c) A "backward" check is then made to see if any variables should be *dropped* from the model. In this case, since there are only two variables in the model, (X_2, X_3), and X_2 was just added, this means that the following partial F^* statistic
   $$F^* = \frac{MSR(X_3|X_2)}{MSE(X_3, X_2)}$$
   is check to see if it is now insignificant. It is not and so variable X_3 (circle one) **remains in the model / is dropped from the model**.

(d) The partial F^* statistic

$$F^* = \frac{MSR(X_k|X_3, X_2)}{MSE(X_3, X_2, X_1)}$$

for all regression models with triples of variables is determined (there is only one in this case, (X_1, X_2, X_3)) for significance. Since X_1 is not significant, it (circle one) **is added to the model / is not added to the model**. Consequently, the model remains one with X_2 and X_3.

3. *Forward.*
   A forward procedure is the same as a forward stepwise procedure, except the former does *not* drop variables.
   From SAS, the forward procedure chooses the model
   (circle one) $\boldsymbol{X_3}$ / $\boldsymbol{X_2, X_3}$ / $\boldsymbol{X_1, X_2, X_3}$

4. *Backward.*
   A *backward* procedure starts with all the variables and then drops the least significant variables, one at a time (and does not every add any variables).
   From SAS, the backward procedure chooses the model
   (circle one) $\boldsymbol{X_3}$ / $\boldsymbol{X_2, X_3}$ / $\boldsymbol{X_1, X_2, X_3}$

5. *Comparing Stepwise Procedures.*
   The three stepwise procedures, forward stepwise, forward and backward, gave (choose one) **the same / different** best models. However, two out of the three procedures gave the best model with predictor variables (X_2, X_3).

6. *Comparing Stepwise Procedures To All–Possible–Regression Procedures.*
   The three stepwise procedures and the all–possible–regression procedures gave (choose one) **the same / different** best models. It appears, though, that the two sets of procedures favor the two models, (X_1, X_3) and (X_2, X_3).

8.5 Some Final Comments on Model Building for Exploratory Observational Studies

Some guides to model building are:

1. Less is more in models; use the least number of regression variables possible to keep the model as simple as possible.

2. Pick predictors that make sense; that are related to the response variable.