Chapter 9: Building the regression model I: model selection and validation

Goal:
(1) Selection of the predictor variables for exploratory observational studies
(2) Methods for validating regression models

9.1 Overview of the model building process

1) Data collection and preparation
2) Reduction of independent variables
3) Model refinement and selection
4) Model validation

See Figure 9.1 on p. 344

1) Data collection and preparation

Data Collection

- Controlled experiments
    i. The experimenter controls the levels of the explanatory variables
    ii. And assigns a treatment (a combination of levels of the explanatory variables which are often called **factors** or **control variables**) to each experiment unit

- Controlled experiments with covariates
    i. There are **uncontrolled variables or covariates** (such as characteristic of the experimental units)
    ii. Cannot be incorporated into the design of the experiment
    iii. Can be in incorporated into the regression model

- Confirmatory observational studies
    i. Test (to confirm or not to confirm) hypotheses derived from pervious studies
    ii. Collect data for explanatory variables which have shown to affect the response variables in previous studies and new variables

- **Exploratory observational studies**
    i. Widely used in social, behavioral, health sciences, management and other fields
    ii. Not controlled experiments

After data is collected, one should do "edit checks" to make sure no extreme data entry errors are made.

2) Reduction of independent variables

Often there are a large number of independent variables under consideration. Since we want the most parsimonious model, it is important to remove independent variables that may not be important to predicting the dependent variable.

In addition, if prior knowledge of the data suggests that interaction and quadratic terms are important, these should be included in this step.

- Controlled experiments: Not an important issue since the explanatory variables are controlled by the experimenter

- Controlled experiments with covariates: Some reduction of the covariates may take place but the number of covariates is usually small

- Confirmatory observational studies: Generally no reductions

- **Exploratory observational studies: Reduction is needed**
    i. The number of explanatory variables is usually large
    ii. Many of the variables are highly correlated which may substantially increase the sampling variation of the regression ocefficients
    iii. A regression model with numerous explanatory variables may be difficult to work with and understand

3) Model refinement and selection

The tentative regression model, or the several "good" regression models, need to be checked in detail for interactions and curvature effects. Diagnostic checks can be helpful to determine if changes need to be made to the model because of model assumption violations.

4) Model validation
Model validity refers to the stability and reasonableness of the regression coefficients, the plausibility and usability of the regression function, and the ability to generalize inferences drawn from the regression analysis.

9.2 Surgical unit example

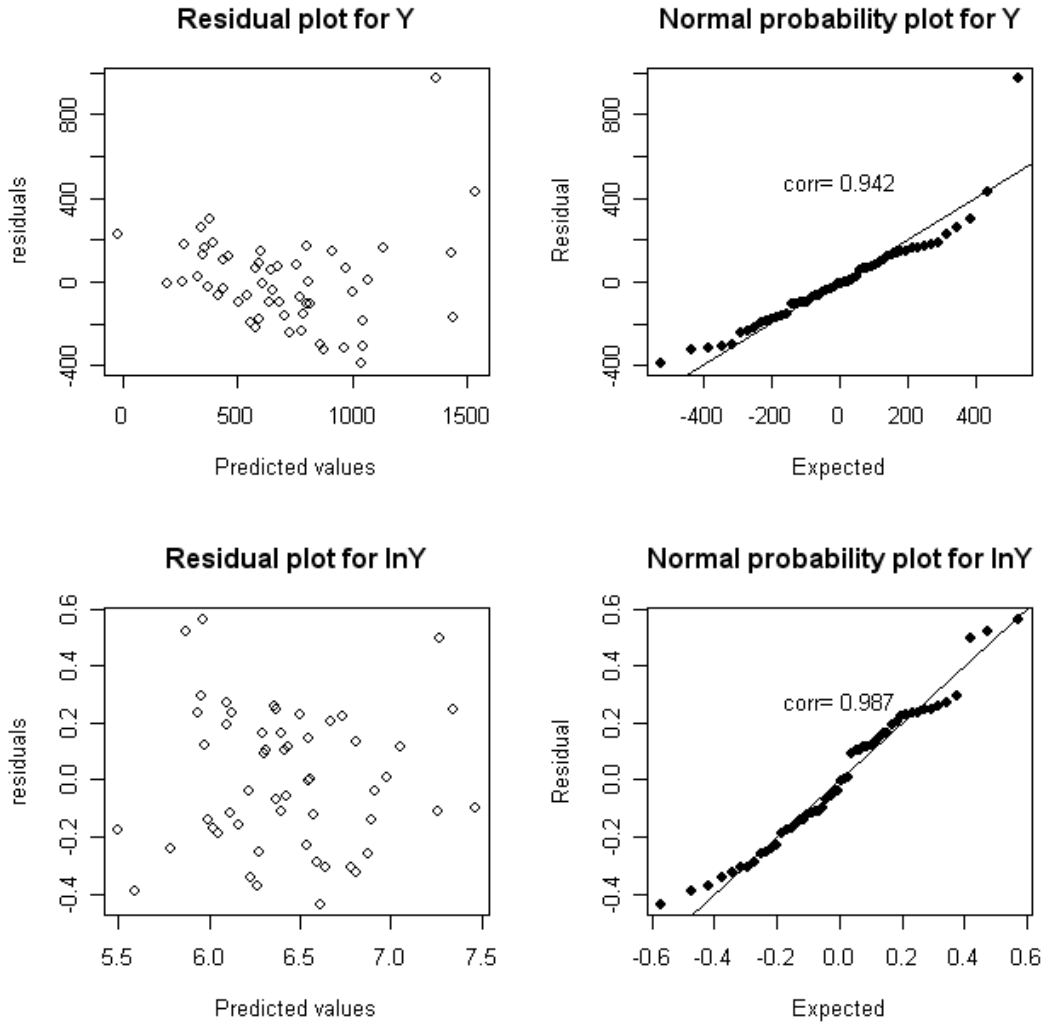
A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 108 patients was available for analysis. From each patient record, the following information was extracted from the pre-operation evaluation:

X_1 blood clotting score
X_2 prognostic index
X_3 enzyme function test score
X_4 liver function test score
X_5 age in years
X_6 indicator variable for gender (0=male, 1=female)
X_7 and X_8 indicator variable for history of alcohol use

Alcohol Use	X_7	X_8
None	0	0
Moderate	1	0
Severe	0	1

The response variable is survival time. A proportion of the data (the first 54 patients) is listed in Table 9.1. For illustration purpose, we only use the first four explanatory variables.

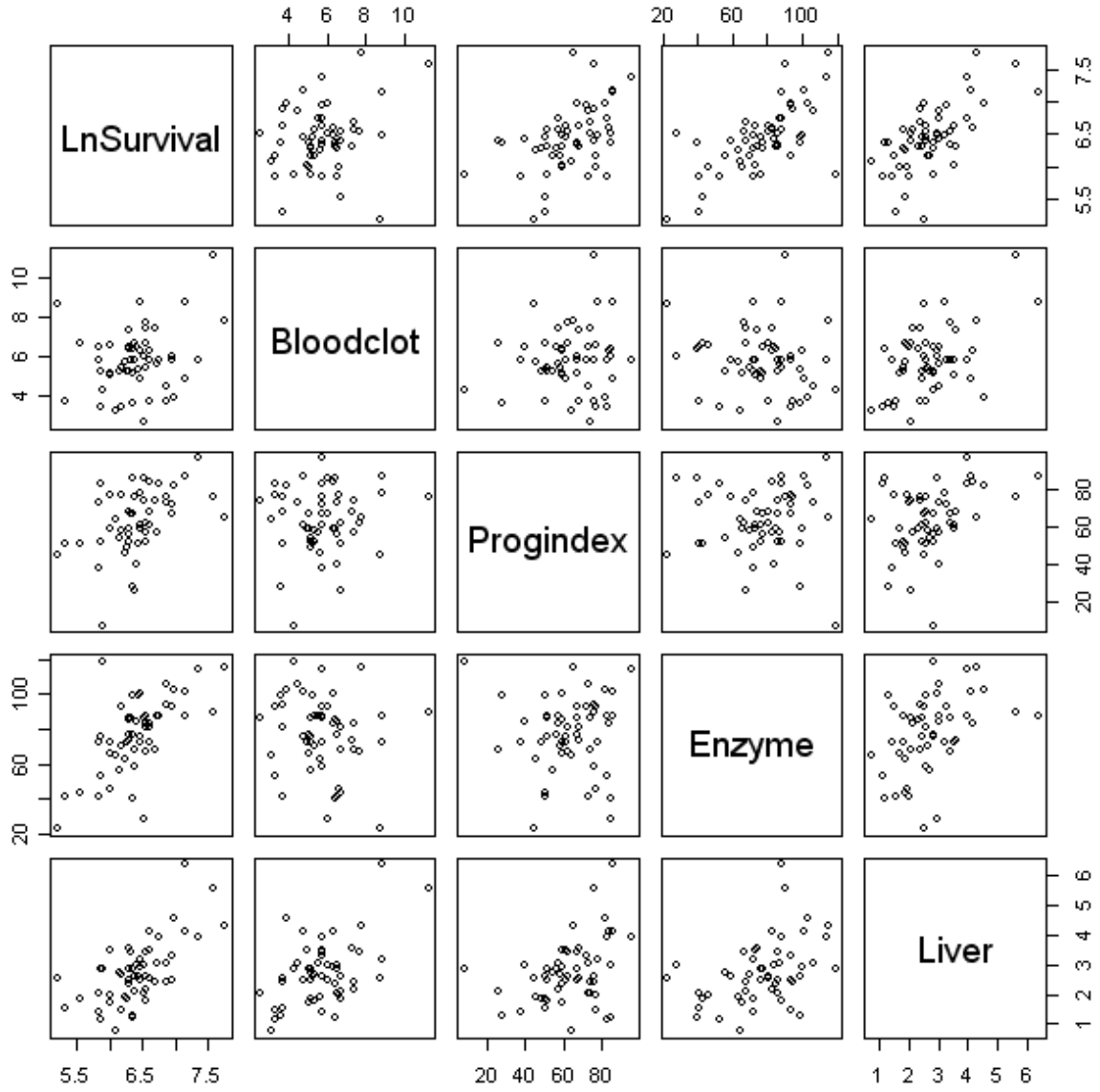
1) Log transformation



2) Correlation matrix and scatter plot matrix

	LnSurvival	Bloodclot	Progindex	Enzyme	Liver
LnSurvival	1.0000000	0.24618787	0.46994325	0.65388548	0.6492627
Bloodclot	0.2461879	1.00000000	0.09011973	-0.14963411	0.5024157
Progindex	0.4699432	0.09011973	1.00000000	-0.02360544	0.3690256
Enzyme	0.6538855	-0.14963411	-0.02360544	1.00000000	0.4164245
Liver	0.6492627	0.50241567	0.36902563	0.41642451	1.0000000

Each of the predictor variables is linearly associated with lnY, with X_3 and X_4 showing the highest degrees of association and X_1 the lowest.
X_4 has moderately high pairwise correlations with X_1, X_2 and X_3.



On the basis of these analyses, the investigators concluded to use, at this stage of model-building process, the log-transformed response variable, to represent the predictor variables in linear terms, and not to include any interaction terms. The next stage in the model-building process is to examine whether all of the potential predictor variables are needed or whether a subset of them is adequate.

9.3 Criteria for model selection

Let P-1 = the total number of independent variables under consideration.

Let p-1 = number of independent variables in a model under consideration.

Examine ALL possible regression models to determine a set of models that are "good" to consider further.

> Suppose there are the independent variables X_1, X_2, and X_3. ALL possible models include the models of: no independent variables, X_1 only, X_2 only, X_3 only, X_1 and X_2, X_1 and X_3, X_2 and X_3, and X_1, X_2 and X_3.

Note:
1) No transformations or interactions are considered at this stage of the model building process (unless they are believed to be important based on knowledge of the data set).
2) The number of possible models is 2^{P-1}

Five different criteria are examined to determine which models are "good".

R_p^2 or SSE_p criterion

> R_p^2 is the coefficient of determination. The subscript p denotes the p-1 variables in the model for which R^2 is calculated. Models with a "large" R_p^2 are considered "good".
>
> SSE_p is the sum of squared errors. The subscript p denotes the p-1 variables in the model for which SSE is calculated. Models with a "small" SSE_p are considered "good".
>
> Notes:
> 1) There is an equivalence between examining these two measures since
> $$R^2 = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO}$$
> 2) Remember that R^2 always increases (or stays the same) as variables are added to the model. Because of this, one should look for a point where the R_p^2 starts to level off as more independent variables are added to the model. A similar discussion can be made about SSE_p.

MSE_p or $R_{a,p}^2$

> Since R_p^2 never decreases, $R_{a,p}^2$ (equivalently MSE) can be used instead.
>
> $$R_{a,p}^2 = 1 - \frac{n-1}{n-p}(1-R^2) = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SSTO} = 1 - \frac{SSE_p/(n-p)}{SSTO/(n-1)} = 1 - \frac{MSE_p}{SSTO/(n-1)}$$

Mallows' C_p criterion

Provide penalties for adding predictors

Steps:

1) Compute MSE($X_1,\ldots,X_{P-1}$). This serves as the "best" estimate of σ^2.
2) Compute SSE($X_1,\ldots,X_{p-1}$) for each subset model
3) Note that **IF** MSE($X_1,\ldots,X_{p-1}$)=SSE($X_1,\ldots,X_{p-1}$)/(n-p) is about the same as MSE($X_1,\ldots,X_{P-1}$), then

$$\frac{SSE(X_1,\ldots,X_{p-1})}{MSE(X_1,\ldots,X_{P-1})} = \frac{(n-p)MSE(X_1,\ldots,X_{p-1})}{MSE(X_1,\ldots,X_{P-1})} \approx n-p$$

where ≈ means "approximately".

4) Compute:

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

where n-2p serves as a penalty for the number of variables in the model.

If $X_1,\ldots,X_{p-1}$ represents a "good" model, then $C_p \approx p$ (see #3 above). Models are also considered to be "good" if $C_p \leq p$ (sampling error causes C_p not to be bigger than p).

Notes:

1) Depends on the set of P-1 variables under consideration.
2) See KNN for a more in depth discussion about the derivation of C_p.

***AIC*$_p$ and *SBC*$_p$ criteria**

Provide penalties for adding predictors. We search for models that have small ***AIC*$_p$** or SBC$_p$.

Akaike's information criterion: $AIC_p = n\ln SSE_p - n\ln n + 2p$

Schwarz' Bayesian criterion: $SBC_p = n\ln SSE_p - n\ln n + [\ln n]p$

For both measures, the first term decreases as p increases, the second term is fixed for a given sample size n, and the third term increases with the number of parameters.

PRESS$_p$ criterion

PRESS = PREdiction Sum of Squares

Measures the SSE obtained when the i^{th} observation is deleted.

PRESS prediction error = $Y_i - \hat{Y}_{i(i)}$

To obtain $\hat{Y}_{i(i)}$, remove the ith observation from the data set, and then estimate the regression function using the remaining (n-1) data points and. $\hat{Y}_{i(i)}$ is the predicted value for the ith observation.

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

Models with a low $PRESS_p$ are "good".

Example: Table 9.2: For all possible regression models – surgical unit example. (Note here "Number in Model" represents the number of variables. To get p, the number of parameters, you have to add 1 to the first column).

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	MSE	SBC	SSE	Variables in Model
1	0.4276	0.4166	66.4889	-103.8269	0.14099	-99.84889	7.33157	Enzyme
1	0.4215	0.4104	67.7148	-103.2615	0.14248	-99.28357	7.40873	Liver
1	0.2208	0.2059	108.5558	-87.1781	0.19191	-83.20011	9.97918	Progindex
1	0.0606	0.0425	141.1639	-77.0788	0.23137	-73.10079	12.03147	Bloodclot
2	0.6633	0.6501	20.5197	-130.4833	0.08456	-124.51634	4.31249	Progindex Enzyme
2	0.5995	0.5838	33.5041	-121.1126	0.10058	-115.14561	5.12970	Enzyme Liver
2	0.5486	0.5309	43.8517	-114.6583	0.11335	-108.69138	5.78096	Bloodclot Enzyme
2	0.4830	0.4627	57.2149	-107.3236	0.12984	-101.35663	6.62201	Progindex Liver
2	0.4301	0.4078	67.9721	-102.0669	0.14312	-96.09998	7.29905	Bloodclot Liver
2	0.2627	0.2338	102.0313	-88.1622	0.18515	-82.19528	9.44267	Bloodclot Progindex
3	0.7573	0.7427	3.3905	-146.1609	0.06217	-138.20494	3.10854	Bloodclot Progindex Enzyme
3	0.7178	0.7009	11.4237	-138.0232	0.07228	-130.06723	3.61413	Progindex Enzyme Liver
3	0.6121	0.5889	32.9320	-120.8442	0.09936	-112.88823	4.96782	Bloodclot Enzyme Liver
3	0.4870	0.4562	58.3917	-105.7477	0.13140	-97.79178	6.57020	Bloodclot Progindex Liver
4	0.7592	0.7396	5.0000	-144.5895	0.06294	-134.64461	3.08396	Bloodclot Progindex Enzyme Liver

9.4 Automatic search procedures for model selection

When there are MANY independent variables, fitting all possible regression models and calculating the measures discussed in Section 9.3 may not be feasible.

This section discusses search algorithms that try to find the "best" regression model.

"Best" subsets algorithms

Provide several "good" subsets according to the specific criterion.
Example: Find the best three subsets using adjusted R-squared value (SAS)

Number in Model	R-Square	Adjusted R-Square	Variables in Model
6	0.8234	0.8434	Bloodclot Progindex Enzyme Age Gender AlcHeavy
7	0.8226	0.8460	Bloodclot Progindex Enzyme Age Gender AlcMod AlcHeavy
5	0.8205	0.8374	Bloodclot Progindex Enzyme Gender AlcHeavy

Stepwise regression methods

End with the identification of a single regression models as "best".

Forward selection

Start with a model with no independent variables.

Perform the following steps:

1. Fit a simple linear regression model for each independent variable (i.e., estimate $E(Y_i)=\beta_0+\beta_1 X_{ij}$ for j=1,…,P-1). Pick the best model.
2. Starting with the model in #1, fit all 2 independent variable models and pick the best model which includes the variable chosen in #1. For example, suppose the model $E(Y_i)=\beta_0+\beta_1 X_{i3}$ was chosen in #1. Then all 2 variable models that include X_3 are fit.
3. Starting with the model in #2 fit all 3 independent variable models and pick the best model which includes the variables chosen in #1 and #2.

⋮

The procedure stops when no more variables are important enough to add to the model.

Notes:

1) The criteria used to determine the best model at each step is the maximum F^* (or lowest p-value) from the partial F-test. This is equivalent to using a t-test (for quantitative variables) since only one variable is being tested at a time.
2) The procedure stops when F^* is less than the partial F-tests critical value (or p-value>α).
3) A recommended level of significance to use is α=0.15. This helps to ensure all possible important variables are allowed to enter the model.

Backward elimination

Start with all of the independent variables in the model.

Perform the following steps:

1. Fit the model with all of the independent variables in the model. Remove the variable that is least important.
2. Starting with the model in #1, Remove the variable that is least important.

⋮

The procedure stops when all of the remaining model variables are important.

Notes:

1) The criteria used to determine the variables to remove from the model is the partial F-test. Again, this is equivalent to using a t-test since only one variable is being tested at a time.
2) The smallest F^* value (least significant variable) is equivalent to the largest p-value
3) The procedure stops when all of the F^*'s are greater than the partial F-tests critical value (or p-value<α).

4) A recommended level of significance to use is α=0.30. This helps to ensure that no possibly important variables are removed model.

Stepwise forward selection

Combination of forward selection and backward elimination.

Start with a model with no independent variables.

Perform the following steps:

1. Fit a simple linear regression model for each independent variable. Pick the best model.
2. Starting with the model in #1, fit all 2 independent variable models and pick the best model which include the variable chosen in #1.
3. After adding a variable, determine if any variables should be removed.
4. Using the resulting model in #3, determine if any variables should be added.

⋮

Notes:

1) The criteria used to determine the variables to remove or add to the model is the partial F-test.
2) The procedure stops when no more variables can be added or removed.
3) A recommended level of significance to use for adding is α=0.15 and removal is α=0.30.

Note that these model selection procedures do not always end up with the same model!

Example of Grocery data set (Problem 6.9; see handout)

A large, national grocery retailer tracks productivity and costs of its facilities closely. Data were obtained from a single distribution center for a one-year period. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped (X1), the indirected costs of the total labor hours as a percentage (X2), a qualitative predictor called holiday that is coded 1if the week has a holiday and 0 otherwise (X3), and the total labor hours (Y).

9.6 Model validation

Model validation involves checking a candidate model against independent data. There are three ways:

1. Collection of new data to check the model and its predictive ability
   a. Re-estimate the model form chosen earlier using the new data.
   b. Use the model chosen earlier to predict each case in the new data set and then to compute the mean of the squared prediction errors (MSPR for mean squared prediction error):

$$MSPR = \frac{\sum_{i=1}^{n^*}(Y_i - \hat{Y}_i)^2}{n^*},$$

   Where, Y_i is the value of the response variable in the ith validation case, $\hat{Y}_i$ is the predicted value for the ith validation case based on the model-bulding data set, and n* is the number of cases in the validation set.

2. Comparison of results with theoretical expectations, earlier empirical results, and simulation results

3. Use a holdout sample to check the model and its predictive ability.

   Data splitting: (1) the **model-building se**t or the **training sample** used to develop the model, and (2) the **validation** or **prediction** set used to evaluate the reasonableness and predictive ability of the selection model.
   This validation procedure is called **cross-validation**.

   Comments:
   1) Double cross validation: The model is built for each half of the split data and then tested on the other half of the data
   2) K-fold cross-validation for smaller data set. The data are first split into K roughly equal parts. For k=1, …, K, we use the kth part as the validation set, fir the model using the other k-1 parts, an obtain the predicted sum of squares for error. The K estimates of prediction error are then combined to produce the K-fold cross-validation estimate. When K=n, it is identical to PRESS.

   Example: Page 373 (Look at the surgical unit example again).