PubH 7405:
REGRESSION ANALYSIS

MLR: MODEL BUILDING

In order to provide more comprehensive prediction of a specific dependent variable Y – say the outcome of certain treatment, it is very desirable to **consider a large number of factors** – with available data **-** and **sort out which ones are most closely related that outcome.** We **do not want to miss** identifying any important predictor/covariate.

NUMBER OF COVARIATES

- In biomedical research, the independent variables are covariates representing patients' characteristics and, in many cases of clinical research, one of them represents the treatment.
- There may be more "**potential predictors**" than we can manage to investigate; sometimes the number of factor is even larger than the sample size.
- Multiple Regression could help us to "investigate" the factors with available data but, in the end, may be only a few of the potential explanatory factors have predictive power. The "core" of the process is **to "build" a model.**

A simple strategy for the building of a regression model consists of some, most, or all of the following **five steps** or phases:

(1) Data collection and preparation,

(2) Preliminary model investigation,

(3) Reduction of the predictor variables,

(4) Model refinement and selection, and

(5) Model validation

#1: DATA COLLECTION

The data collection phase separates studies into two types:

(1) **Controlled experiments**, and

(2) **Observational Studies**

#2: **PRELIMINARY** MODEL INVESTIGATION

Once data have been collected, the process begins with steps/actions employed to identify:

(1) Functional **form for predictor variables**; whenever possible, one should rely on investigator/statistician's prior knowledge and or similar previous studies to suggest appropriate **data transformations** – such as taking logs.

(2) Important **interactions** that should be included in the list of variables from which to narrow down in the next step.

#3: **REDUCTION** OF EXPLANATORY VARIABLES

Major reason? To avoid “multiple decision problem”; in addition, with factors include – some are unnecessarily – the error degree of freedom is reduced which weakens subsequent statistical decisions:

Very often, inexperienced investigators might screen a set of explanatory variables by **fitting the full model containing the entire set of potential predictor variables, then simply drop those "not statistically significant" factors** using individual t-tests. It first seems reasonable but one may drop important intercorrelated predictor variables (which changes the results for the remaining factors – which is good but has to be done cautiously – with a "control" strategy).

#4: **MODEL REFINEMENT** & SELECTION

At this stage, a “tentative” regression model or models need to be checked in details for curvature and interaction effects; residual plots are helpful here. In addition, **efforts are needed to identify outliers and further reduction is needed due to multicollinearity.**

#5: **MODEL VALIDATION**

Validation is an useful and necessary final phase of the model-building process. It refers to the stability and reasonableness of the regression coefficients, the **plausibility and usability** of the regression function, as well as and the **ability to generalize** references drawn to the population situations.

To deal with that very **large number of models**:

In the early phase, the reduction phase, we usual decide on **a criterion** or **a set of criteria**, then the selection/elimination process proceeds "automatically" according to that selected criterion in order to reduce the list candidate models to a smaller number – say 3 to 6.

In the later phase, the refinement phase, a detailed examination can then be made to that more limited number of the more promising or "candidate" models, leading to the selection of the final model or models for validation.

There are more than one criteria for "model selection"; none is clearly dominating. The decision on which criterion to use is still a subjective judgment. Even for a given criterion, it is still possible that **more than one "good" models are found**, and that the choice can only be made on the basis of additional considerations. For example, we settle on certain specific criterion with the goal of identifying a model with "high predictive power" (high according to that selected criterion); but how high is high? Or how much "higher" is enough to justify the selection of a larger model?

Generally:

(1) When the pool of explanatory variables are small enough, one could examine most of possible subsets of explanatory variables and identify those subsets that are “good” according to the selected criterion,

(2) For larger pools of explanatory variables, one has to rely on automatic search procedures to arrive at a single subset of the explanatory variables; the only step left is to validate it for possible acceptance – most of the middle steps in this automatic search are very much hidden.

R^2 CRITERION

The "**R^2 Criterion**" calls for the use of the "coefficient of multiple determination" R^2 in the effort to identify "good" subset or subsets of predictor variables; subsets for which R^2 is "high".

Coefficient of multiple determination R^2 represents the portion of total variation in y-values attributable to difference in values of independent variables or covariates – that is attributable to "the model"; it is a measure of **"predictive power" of a regression model.**

The intent in using the R^2 criterion is to find a point where adding more predictor variables is "not worthwhile" because it leads to a very small increase in R^2. But "**how small is small?**"; clearly the determination of where diminishing returns set in is a **judgmental** one. One could use some conventional "floor value" like less than 3% or 5% - just like the use of .05 cutoff for p-values.

Since R^2 does not take into account the number of parameters in the regression model, **it never reaches a maximum** value; it keeps increasing as additional predictor variables are added to the model. Therefore, forming a "stopping rule" is always subjective.

R_a^2 CRITERION

The "**adjusted coefficient of multiple determination**" R_a^2 has been suggested as an alternative; adjusted coefficient takes the number of parameters in the regression model into account through the degrees of freedom:

$$R_a^2 = 1 - \left(\frac{df_T}{df_E} \right) \frac{SSE}{SST}$$

$$= 1 - \frac{MSE}{MST}$$

$$R_a^2 = 1 - \left(\frac{df_T}{df_E}\right)\frac{SSE}{SST}$$

$$= 1 - \frac{MSE}{MST}$$

MST = SST/(n-1) which is not often mentioned. When we keep adding more and more predictor variables, the decrease in SSE becomes so small, not enough to offset the decrease in the error degree of freedom) that **R_a^2 can**, indeed, reach a **maximum** value and start to decrease. That may be a "natural" point where we should stop the search process.

AIC CRITERION

Another well-known alternative criterion that also provide penalties for adding predictors is the "Akaike's Information Criterion (AIC)". In AIC, the first term (nln(SSE)) **decreases** as the number of predictor variables increases, the second term (nln(n)) is **fixed**, and the third term **increases**. We **choose model with small AIC value**:

$$AIC = n\ln(SSE) - n\ln n + 2(k+1)$$

When the number of factors in the model, k, increases, up to a point, Akaike's Information Criterion (AIC) starts to go up! There is an optimal choice: the model with the smallest AIC value.

In the general case, the AIC is

$$\mathbf{AIC = 2k - 2\ln(L)}$$

where k is the number of parameters, and L is the likelihood function.

The AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. **This penalty discourages overfitting**. The preferred model is the one with the lowest AIC value. The AIC methodology attempts to find the model that best explains the data with a minimum number of free parameters.

There are other criteria besides the three popular ones that have been introduced here; the textbook introduces three others : **Mallow's C_p** (p is the number of parameters, p=k+1), **SBC**, and **PRESS** (pages 357-361). **One can also use the p-value** associated with the "marginal contribution" of the newly added predictor variable; but the interpretation is a bit more obscure because the p-value is affected more by the sample size n.

$$C_p = \frac{SSE}{s^2} - (n - 2p)$$

In this formula, s^2 is the MSE of the full model. C_p is proportional to the sum of the squared bias plus the variance of the predicted; if there is no bias, it is approximately p (the number of parameters); **Good models have (i) small C_p value and (ii) C_p is near p.**

For larger k (the number of predictor variables), the total number of possible models 2^k increases quickly (1024 models for 10 variables) and the task of fitting and evaluating all models becomes **impossible –** at least not practically feasible. We have no choice but to rely only computer packages to perform certain "**automatic computer-search procedure**".

There are two automatic search processes:

(1) To specify a criterion, say R_a^2. The computer has **to fit all possible models** but the "selection step" is automatic; the result may be one or one small group of good models. This is usually done with **smaller number of potential predictors**.

(2) To evaluate, "one-at-a-time", the marginal contribution of one predictor variable at a time and the process is stopped when all and only "good" ones are in the model. In theory it's still a long process but in reality, **the process usually stops after a few or several rounds**.

ONE-AT-A-TIME AUTOMATIC SEARCH PROCESS

- Two important steps are still:

(1) Specifying **criterion or criteria** for the automatic selection step (i.e. threshold p-values)

(2) Specifying a **strategy** for applying the criteria.

- The basic strategy is that of adding into or removing from the current model one predictor variable at a time according to certain order of relative importance with respect to meeting the selected p-values' thresholds.

#1: ONE-AT-A-TIME STRATEGIES

Basic “moving” Strategies are:

(i) **Forward** Selection Procedure,

(ii) **Backward** Elimination Procedure,

(iii) **Stepwise** Regression Procedure.

In practice “forward selection” and “backward elimination” are usual included as steps of a “stepwise” strategy. **Stepwise Regression** is perhaps the most popular search strategy, especially when the number of predictor variables is rather large (say, 30-40) – even larger than sample size - and fitting all possible models may not be feasible.

An **important characteristic**: The stepwise algorithm identifies not a group of models but a single regression model as “**the best**”

FORWARD SELECTION

The forward-selection phase starts with no variables in the model. It fits all one-variable models, calculates the "F statistics"; the one with largest F serves as a "**candidate**".

(1) If the p-value for this candidate is smaller than a pre-determined level, called SLENTRY or SLE (for example, we choose SLE = .05), that candidate is the first variable to be included in the model;

(2) Otherwise, the process stops without any variable in the final model).

Let say the result is X_7.

Assume that X_7 "entered" the model in step 1. The routine now fits all models with two predictor variables, one of the two must be X_7; for example, models with (X_7,X_1), (X_7,X_2), etc…For each two-variable model, it calculates the F statistic associated with the marginal contribution of the newly added variable; for example, MSR(X_2|X_7)/MSE(X_2,X_7). **The one with largest F serves as a "candidate".**

(1) If p-value for this candidate is smaller than SLENTRY, that candidate is the second variable to be included in the model. Let say the result is X2; we now have 2 variables in the model,
(2) Otherwise, the process stops with one variable in the final model). **The process is repeated to evaluate all three-variables model.**

If “FORWARD” is used as the strategy, once a variable is in the model, it stays.

BACWARD ELIMINATION

The backward elimination strategy begins with the "full model" (including all predictor variables). Then the variables are deleted from the model one at a time and it calculates the F statistics associated with the marginal contribution of each deleted variable- just **like it is one newly added** , say

$MSR(X_4|X_1,X_2,X_3,X_5,X_6,X_7)/MSE(X_1,X_2,X_3,X_4,X_5,X_6,X_7)$

The variable with smallest F serves as a **candidate**

(1)If the p-value for this candidate is greater than a pre-determined level, called **SLSTAY** or **SLS** (for example, we choose SLS = .15), that candidate is the first variable to be eliminated from the model;

(2) Otherwise, the process stops and all variables stay to form the final model.

The process continues searching for the next elimination.

STEPWISE REGRESSION

Stepwise regression combines forward selection and backward elimination , one step forward followed by one step backward, continues until no more variables can be added or removed. We need both SLENTRY and SLSTAY; **default values** for SAS are both ".15" (default value for SLE is .50 for FORWARD alone and SLS is .10 for BACKWARD alone).

#2: IMPLEMENTATION OF ALL-POSSIBLE MODELS

The “all-possible models” search algorithm, one can put in one or more than one criteria in a SAS program:

```
MODEL Y = X1 X2 X3 X4 X5/
SELECTION = RSQUARE ADRSQ AIC;
```

For each of these criteria, the routine performs all possible regression models then display the results in the increasing order of the number of predictor variables (all one-variable models, followed by all two-variable models, etc…) then in the decreasing order of the statistic within each group of models of the same size. When more than one criteria are specified, the decreasing order is the one of the first criterion listed.

To reduce computing time/effort and/or computer output, we have the following options:

(1) STOP = s specifies the largest number of predictor variables to be evaluated and reported,

(2) BEST = b specifies that only b best models are displayed for each group of models of the same size.

And instead of “SELECTION = RSQUARE”, the specification “SELECTION = MAXR” would result in a more simple computer output listing of only “the best one-variable model”, “the best two-variable model”, etc…(instead the list of ALL models of the same size). Actually, MAXR has a hidden backward elimination step similar to that of STEPWISE but checking the effect of elimination/switching on R^2 instead of p-value.

```
PROC REG data = SURV;
model Y = X1 X2 X3 X4/selection=stepwise
SLentry = .15 SLstay = .30;
```

Summary of Stepwise Procedure for Dependent Variable Y

Step	Variable Label	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X4 Liver Test	1	0.5218	0.5218	93.5454	56.7310	0.0001
2	X2 Prognostic	2	0.0956	0.6174	66.8411	12.7473	0.0008
3	X3 Enzyme	3	0.1449	0.7623	25.3353	30.4939	0.0001
4	X1 Blood Clotting	4	0.0744	0.8367	5.0000	22.3353	0.0001

(NOTE: No step #5; No removal – because SLstay was set at ".30")

PROC REG data = SURV;

model Y = X1 X2 X3 X4/ selection = rsquare adjrsq cp;

Number in Model	R-square	Adjusted R-square	C(p)	Variables in Model
1	0.52175569	0.51255868	93.54544	X4
1	0.33668289	0.32392679	149.09519	X3
1	0.30688939	0.29356034	158.03774	X2
1	0.13877015	0.12220803	208.49887	X1
2	0.65911506	0.64574702	54.31690	X2 X3
2	0.61738863	0.60238426	66.84113	X2 X4
2	0.61623041	0.60118063	67.18877	X3 X4
2	0.55251128	0.53496270	86.31412	X1 X3
2	0.52187923	0.50312940	95.50835	X1 X4
2	0.41180870	0.38874237	128.54612	X1 X2
3	0.83253178	0.82248369	4.26573	X1 X2 X3
3	0.76233507	0.74807518	25.33533	X2 X3 X4
3	0.64519769	0.62390956	60.49421	X1 X3 X4
3	0.61970441	0.59688667	68.14604	X1 X2 X4
4	0.83674875	0.82342212	5.00000	X1 X2 X3 X4

PROC REG data = SURV;

model Y = X1 X2 X3 X4/ selection = rsquare Stop = 3;

Number in Model	R-square	Adjusted R-square	C(p)	Variables in Model
1	0.52175569	0.51255868	93.54544	X4
1	0.33668289	0.32392679	149.09519	X3
1	0.30688939	0.29356034	158.03774	X2
1	0.13877015	0.12220803	208.49887	X1
2	0.65911506	0.64574702	54.31690	X2 X3
2	0.61738863	0.60238426	66.84113	X2 X4
2	0.61623041	0.60118063	67.18877	X3 X4
2	0.55251128	0.53496270	86.31412	X1 X3
2	0.52187923	0.50312940	95.50835	X1 X4
2	0.41180870	0.38874237	128.54612	X1 X2
3	0.83253178	0.82248369	4.26573	X1 X2 X3
3	0.76233507	0.74807518	25.33533	X2 X3 X4
3	0.64519769	0.62390956	60.49421	X1 X3 X4
3	0.61970441	0.59688667	68.14604	X1 X2 X4

NOTE: There is no model with 4 predictor variables with “Stop = 3”

PROC REG data = SURV;

model Y= X1 X2 X3 X4/selection= rsquare Stop=3 Best=2;

Number in Model	R-square	Adjusted R-square	C(p)	Variables in Model
1	0.52175569	0.51255868	93.54544	X4
1	0.33668289	0.32392679	149.09519	X3
2	0.65911506	0.64574702	54.31690	X2 X3
2	0.61738863	0.60238426	66.84113	X2 X4
3	0.83253178	0.82248369	4.26573	X1 X2 X3
3	0.76233507	0.74807518	25.33533	X2 X3 X4

NOTE: There is no model with 4 predictor variables with “Stop = 3”
& only 2 models are listed in each group because of “Best = 2”

```
PROC REG data = SURV;
model Y = X1 X2 X3 X4/selection=maxr;
```

Maximum R-square Improvement for Dependent Variable Y

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-71.92123497	38.30031842	36287.70264864	3.53	0.0660
X4	98.05483965	13.01843184	583808.87323857	56.73	0.0001

The above model is the best 1-variable model found.

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-424.56413841	63.74961983	331720.68573475	44.35	0.0001
X2	4.88261191	0.70299436	360779.51674699	48.24	0.0001
X3	4.05843991	0.55907160	394116.40334444	52.70	0.0001

The above model is the best 2-variable model found.

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-659.17941990	55.67408110	525372.16182084	140.18	0.0001
X1	38.32274412	5.32589144	194041.43220438	51.78	0.0001
X2	4.56773202	0.49955913	313323.76018378	83.60	0.0001
X3	4.48503622	0.40017410	470760.31016036	125.61	0.0001

The above model is the best 3-variable model found.

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-621.59755029	64.80042601	343025.80578100	92.02	0.0001
X1	33.16382813	7.01727463	83263.81220448	22.34	0.0001
X2	4.27185982	0.56338454	214332.51418236	57.49	0.0001
X3	4.12573829	0.51116093	242857.75291128	65.15	0.0001
X4	14.09156259	12.52532754	4718.50099971	1.27	0.2661

The above model is the best 4-variable model found.

NOTE: WE could impose a limit “STOP = s” on the MAXR routine/strategy

“Type I” SUM OF SQUARES

Type I fits the reduction in residual sum of squares achieved by adding that variable at whatever stage it is added. Type I *SS* are order-dependent (hierarchical, sequential). Each effect is adjusted for all other effects that appear earlier – or enter before it - in the model, but not for any effects that appear later – or enter after it - in the model. For example, it is "adjusted" for no other variable if it is first in the model, for one other variable if it is second, for two others if it is third, etc.

“Type III” SUM OF SQUARES

For Type III , each variable is credited with the reduction in residual sum of squares achieved by entering it on top of all other variables; each is adjusted for all others. When you fit just one multiple regression model – or when options such as RSQUARE, ADJRSQ, and/or Cp are specified, “SSR” is of type III.

“Type II” SUM OF SQUARES

Type II *SS* are the reduction in the *SSE* due to adding the effect to a model that contains all other effects except effects that “contain the effect being tested”. An effect is contained in another effect if it can be derived by deleting terms in that effect—for example, A, B, C, A*B, A*C, & B*C are all contained in A*B*C. For example, in evaluating the effect of A, it is not adjusted for A*B. Type II Sum of Squares is sequential like type I, the effect of each variable is not adjusted for variables not yet entered the model. However, type II Sum of Squares is like type III – for each model being investigated at the time, the effect of each variable is adjusted for all others regardless of the entering order.

In sequential model building (with option FORWARD, BACKWARD, or STEPWISE), SAS uses – as “default” – Type II Sum of Squares. However, if preferred – regardless of selection strategy, you can always request your choice of Sum of Squares. For example, even if you fit only one multiple regression model, you could request for “SS1”. In that case the effect of each variable is adjusted for all variables listed before it (on left side) in the model statement –but not variables listed after it. Of course, there is no compelling reason to do that.

Readings & Exercises

- Readings: A thorough reading of the text's sections 9.1-9.4 (pp.343-367) is recommended.
- Exercises: The following exercises are good for practice, all from chapter 9: 9.13(c), 9.15(c), 9.17(c,d), 9.19, and 9.30.

Due As Homework

#18.1 Refer to dataset "Infants", let Y = Birth Weight and consider a model with four independent variables, X_1 = Gestational Weeks, X_2 = Mother's Age, X_3 = Toxemia (toxemia = 1 is a pregnancy condition resulting from metabolic disorder), and X_4 = Length:

a) Fit the MLR model containing all 4 variables; does it appear that all four predictor variables should be retained?

b) Find the best subset of predictors using the stepwise strategy.

c) Find the best subset of predictors using the R_a^2 criterion; does the result agree with that in (b)?

d) Find the best subset of predictors using the AIC criterion; does the result agree with that in (b)? And with that in (c)?

#18.2 We have data on the conduct of a number of cancer clinical trials from "ClinicalTrials.gov" (File: Minority Enrollment); the aim is to is to investigate potential factors which might affect the enrollment of black patients.

There were n=113 trials and the (response) variable under investigation is the percent of black patients (“Black”) among those recruited for each trial. To provide possible explanations, we’ll investigate 9 possible exploratory factors: Age (1= under 18, 2 = 18 and above), Gender (1 = Male, 2 = Female, 3 = both), Funder (1 = Government, 2 = Industry, 4 = Combination), Trial Duration (in months), Allocation (1 = Randomized, 2 = Non-randomized), Intervention Model (or Design; 1 = Parallel (multiple arms), 2 = Single group, 3 = Cross-over), Primary Purpose (1 = Therapeutic, 2 = Non-therapeutic), Masking (1 = Open Label, 2 = Double Blind). The final factor, Trial Size, is represented by two variables: Actual enrollment, and Accrual Percentage which expressed accrual as percentage of Planned Accrual.

a) Define indicator/dummy variables to represent categorical factors; how may independent variables are there?

b) Fit the MLR model containing all variables; does it appear that all independent variables should be retained? How many are significant?

c) Find the best subset of predictors using the stepwise strategy.