Topic 13 - Model Selection

STAT 525 - Fall 2013

Outline

- Variable Selection
  - R^2
  - C_p
  - Adjusted R^2
  - PRESS
- Automatic Search Procedures

Predicting Survival - Page 350

- Surgical unit wants to predict survival in patients undergoing a specific liver operation
- Has random sample of 108 patients - use only 54 patients
- Response Y is survival time (days)
- Eight predictor variables
  - X_1 blood clotting score
  - X_2 prognostic index
  - X_3 enzyme function score
  - X_4 liver function score
  - X_5 age
  - X_6 gender
  - X_7 and X_8 history of alcohol use

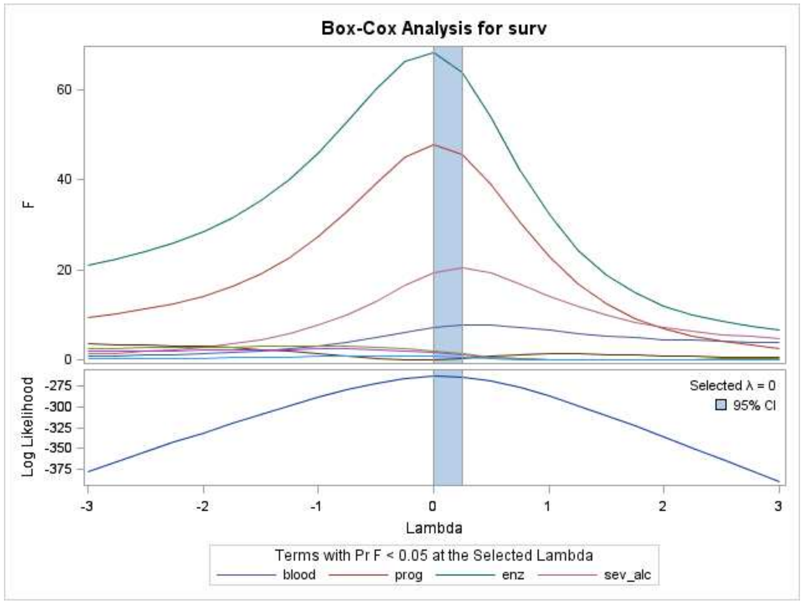
Survival Time as a Response

- Conditional distribution often highly skewed to the right
- Times can be censored if study stopped prior to all deaths
- Survival analysis techniques should be used when censoring is present
- In this case, we observe all survival times so we will investigate transformation using Box-Cox transformation

SAS Commands

```
Data a1;
 infile 'U:\.www\datasets525\Ch09ta01.txt' dlm='09'x;
 input blood prog enz liver age female mod_alc sev_alc surv;
run;

proc transreg;
 model boxcox(surv) = identity(blood) identity(prog) identity(enz)
                      identity(liver) identity(age)  identity(female)
                      identity(mod_alc) identity(sev_alc);
run;
```

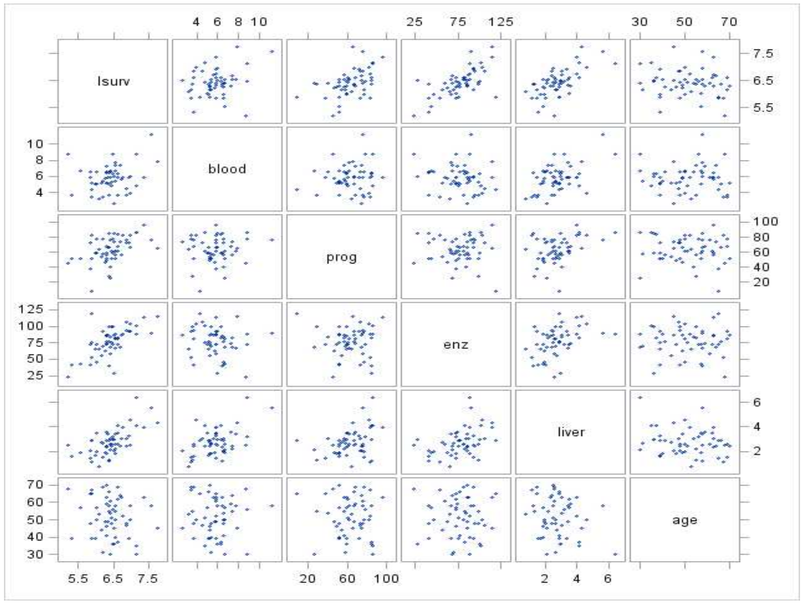
Continuing the Analysis

```
data a1; set a1;
   lsurv=log(surv);
run;

proc sgscatter;
 matrix lsurv blood prog enz liver age;
run;

proc corr;
 var lsurv blood prog enz liver age;

proc reg data=a1;
   model lsurv=blood prog enz liver age female mod_alc sev_alc /
   selection= rsquare adjrsq cp aic sbc best=2 b;
run;
```

Output

```
                    Simple Statistics
Variable  N       Mean    Std Dev       Sum     Minimum   Maximum
lsurv    54    6.43054    0.49152  347.24929   5.19850   7.75919
blood    54    5.78333    1.60303  312.30000   2.60000  11.20000
prog     54   63.24074   16.90253       3415   8.00000  96.00000
enz      54   77.11111   21.25378       4164  23.00000 119.00000
liver    54    2.74426    1.07036  148.19000   0.74000   6.40000
age      54   51.61111   11.12267       2787  30.00000  70.00000

                Pearson Correlation Coefficients, N = 54
            lsurv     blood      prog       enz     liver       age
lsurv     1.00000   0.24633   0.47015   0.65365   0.64920  -0.14505
                     0.0726    0.0003    <.0001    <.0001    0.2953
blood               1.00000   0.09012  -0.14963   0.50242  -0.02069
                               0.5169    0.2802    0.0001    0.8820
prog                          1.00000  -0.02361   0.36903  -0.04767
                                         0.8655    0.0060    0.7321
enz                                     1.00000   0.41642  -0.01290
                                                    0.0017    0.9262
liver                                              1.00000  -0.20738
                                                              0.1324
```


Variable Selection

- Two distinct questions
  1. What is the appropriate subset size?

     adjusted R^2, C_p, MSE, PRESS, AIC, SBC
  2. What is the best model for a fixed size?

     R^2 and any of the above measures

C_p Criterion

- Compares total mean squared error with σ^2
- Squared error

$$\begin{aligned}(\hat{Y}_i - \mu_i)^2 &= (\hat{Y}_i - \mathrm{E}(\hat{\mathrm{Y}}_\mathrm{i}) + \mathrm{E}(\hat{\mathrm{Y}}_\mathrm{i}) - \mu_\mathrm{i})^2 \\ &= (\mathrm{E}(\hat{\mathrm{Y}}_\mathrm{i}) - \mu_\mathrm{i})^2 + (\hat{\mathrm{Y}}_\mathrm{i} - \mathrm{E}(\hat{\mathrm{Y}}_\mathrm{i}))^2 \\ &= \mathrm{Bias}^2 + (\hat{\mathrm{Y}}_\mathrm{i} - \mathrm{E}(\hat{\mathrm{Y}}_\mathrm{i}))^2\end{aligned}$$

- Mean value is $(\mathrm{E}(\hat{\mathrm{Y}}_\mathrm{i}) - \mu_\mathrm{i})^2 + \sigma^2(\hat{\mathrm{Y}}_\mathrm{i})$
- Total mean value is $\sum(\mathrm{E}(\hat{\mathrm{Y}}_\mathrm{i}) - \mu_\mathrm{i})^2 + \sum\sigma^2(\hat{Y}_i)$
- Criterion measure

$$\Gamma_p = \frac{\sum(\mathrm{E}(\hat{\mathrm{Y}}_\mathrm{i}) - \mu_\mathrm{i})^2 + \sum\sigma^2(\hat{Y}_i)}{\sigma^2}$$

C_p Criterion

- Do not know σ^2 nor numerator
- For σ^2, use MSE$(X_1, X_2, ..., X_{p-1})$=MSE(F) as estimate
- For numerator:
  - Can show $\boldsymbol{\sigma}^2(\hat{\mathbf{Y}}) = \sigma^2\mathbf{H}$
  - This means $\sum\sigma^2(\hat{Y}_i) = \sigma^2\mathrm{Trace}(\mathbf{H}) = \sigma^2 p$ (Trace(idempotent matrix)= rank)
  - Can show E(SSE)=$\sum(E(\hat{Y}_i) - \mu_i)^2 + (n-p)\sigma^2$
  - Note: when model correct, $\sum(E(\hat{Y}_i) - \mu_i)^2 = 0$

$$\begin{aligned}C_p &= \frac{(\mathrm{SSE_p} - (\mathrm{n-p})\mathrm{MSE(F)}) + \mathrm{pMSE(F)}}{\mathrm{MSE(F)}} \\ &= \frac{\mathrm{SSE_p}}{\mathrm{MSE}(\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_{\mathrm{p}-1})} - (n-2p)\end{aligned}$$

C_p Criterion

- p is number of predictors + intercept
- When model correct, there is no bias
- $E(C_p) \approx p$
- When plotting models against p
  - Biased models will fall above $C_p = p$
  - Unbiased models will fall around line $C_p = p$
  - By definition: C_p for full model equals p

Adjusted R^2 Criterion

- Takes into account the number of parameters in model
- Switches from SS's to MS's

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

- Choose model which maximizes R_a^2
- Same approach as choosing model with smallest MSE

PRESS_p Criterion

- Looks at the prediction sum of squares which quantifies how well the fitted values can predict the observed responses
- For each case i, predict Y_i using model generated from other $n-1$ cases
- $\text{PRESS} = \sum (Y_i - \hat{Y}_{i(i)})^2$
- Want to select model with small PRESS
- Can calculate this in one fit (Chpt 10)

Other Approaches

- Criterion based on minimizing -2log(likelihood) plus a penalty for more complex model
- AIC - Akaike's information criterion

$$n \log\left(\frac{\text{SSE}_\text{p}}{n}\right) + 2p$$

- SBC - Schwarz Bayesian Criterion

$$n \log\left(\frac{\text{SSE}_\text{p}}{n}\right) + p\log(n)$$

- Can use to compare non-nested models

Selection in SAS

- Helpful options in model statement
  - selection= to choose criterion and method
    - forward (step up)
    - backward (step down)
    - stepwise (forward with backward glance)
  - include=n forces first n variables into all models
  - best=n limits output to the best n models
  - start=n limits output to models with $\geq n$ X's
  - b will include parameter estimates

Models of same subset size

- Can also use R^2 or SSE
- May result in several worthy models
- Use knowledge on subject matter to make final decision
- Decision not that important if goal is prediction

Output

R-Square Selection Method

Number in Model	R-square	Adjusted R-square	C(p)	AIC	SBC
1	0.4273	0.4162	117.4783	-103.8110	-99.83305
1	0.4215	0.4103	119.1712	-103.2679	-99.28994
2	0.6632	0.6500	50.4918	-130.4785	-124.51159
2	0.5992	0.5835	69.1967	-121.0890	-115.12206
3	0.7780	0.7647	18.9015	-151.0021	-143.04620
3	0.7572	0.7427	24.9882	-146.1614	-138.20545
4	0.8299	0.8160	5.7340	-163.3759	-153.43101
4	0.8144	0.7993	10.2633	-158.6694	-148.72443
5	0.8375	0.8205	5.5282	-163.8257	-151.89179
5	0.8359	0.8188	5.9990	-163.2934	-151.35953
6	0.8435	0.8235	5.7725	-163.8583	-149.93537
6	0.8392	0.8186	7.0288	-162.3961	-148.47317
7	0.8460	0.8226	7.0288	-162.7428	-146.83088
7	0.8436	0.8198	7.7214	-161.9186	-146.00670
8	0.8461	0.8187	9.0000	-160.7773	-142.87649

Background Reading

- KNNL Sections 9.1-9.5
- knnl350.sas
- KNNL Chapter 10