Applied Survival Analysis Using R
Chapter 6: Model Selection and Interpretation

Qi Guo

Department of Mathematical Sciences
The University of Texas at Dallas

April, 13 2019

1 Covariate Adjustment

2 Categorical and Continuous Covariates

3 Nested Models

4 The Akaike Information Criterion for Comparing Non-nested Models

Covariate Adjustment

- Most of our study is a randomized clinical trial, the focus will be on *comparing the effectiveness of different treatments*.
- A successful randomization procedure should ensure that *confounding covariates are balanced* between the treatments.
- Goal:Use methods to sift through a potentially large number of potential explanatory variables to find the important ones.
- Example(`coxph` model of the effect of treatment on survival unadjusted for the genetic mutation status of the patients)

```
> coxph(Surv(ttAll, status) ~ trt)
      coef    exp(coef)     se(coef)     z         p
trt 0.464         1.59        0.117    3.96    7.6e-05

Likelihood ratio test=15.5 on 1 df,p=8.2e-05
```

Conclusion and stratify

- Conclusion:
  - The estimate of the log hazard ratio treatment effect, $\hat{\beta}$, is 0.464, so higher hazards are associated with the treatment than with the control, unfortunate result.
  - The value of $e^{\hat{\beta}} = 1.59$, suggesting (incorrectly, as we know) that the treatment is associated with a 59% additional risk of death over the control.
- Stratify on **`genotype`**:

```
> coxph(Surv(ttAll, status) ~ trt + strata(genotype))
coef    exp(coef)     se(coef)     z         p
trt -0.453        0.636        0.164    -2.76    0.0058

Likelihood ratio test=7.66 on 1 df,p=0.00566
```

- Conclusion: The coefficient is negative, indicating that, within each genotype, *the treatment is effective*.

Explicitly estimating the genetic effect

```
> coxph(Surv(ttAll, status) ~ trt + genotype)
                coef      exp(coef)    se(coef)     z         p
trt           -0.453          0.636       0.164    -2.76     0.0058
genotypewt    -1.568          0.209       0.183    -8.59     0.0000
Likelihood ratio test=93.4 on 2 df, p=0
```

- Conclusion:The **wild type genotype** has lower hazard than the **reference (mutant) genotype**, and thus that the mutant genotype incurs additional risk of death.

Indicator or Dummy variable

- The previous sections considered a partial likelihood for comparing two groups, indexed by a covariate z. Since z can take the values 0 or 1 depending on which of two groups a subject belongs to, this covariate is called an `indicator` or `dummy variable`.
- When *categorical variables* with three or more variables, we will need `multiple dummy variables`. For example:
  - If a research question is how survival in `non-white` groups compares to survival in `whites`, one would select "`white`" as the reference variable. Since there are four levels, we need to create three dummy variables, say, z_2, z_3, and z_4 to represent "`race`". Then for a `white` patient, all three would take the value zero. For an Asian person, we would have $z_2 = 1$, and $z_1 = z_3 = 0$.

k covariates model and enhance

- k covariates model:

$$\log(\psi_i) = z_{1i}\beta_1 + z_{2i}\beta_2 + \cdots + z_{ki}\beta_k. \quad (1)$$

- For each covariate, the parameter β_j is the log hazard ratio for the effect of that parameter on survival, adjusting for the other covariates.
- Matrix form: $\log(\psi_i) = z_i'\beta$ (for Patient i), where z_i' (the transpose of z_i) is a 1 *x* *k* matrix (i.e. a row matrix) of covariates, and β is a *k* *x* 1 matrix (i.e. a column matrix) of parameters.
- Enhance the model:
  - ⊗ If a continuous variable is not linearly related to the log hazard, transform it using, eg: a logarithmic or square root function.
  - ⊗ "discretize" a variable, eg: split the "age" into three pieces, "under 50" and "50-64", and "65 and above" and entered into the model as a categorical variable.

Difference

- ⊙ But it is incorporate for *interaction* terms.
- Difference with linear and logistic regression model:
  - Survival data can evolve over time, there is a possibility that some covariate values may also *change as time passes*. eg:Time-related variables like `age` must also be defined and fixed by taking their value at the beginning of the trial, even though patients will age as the trial progresses(Chapter 8).
  - There is no intercept term in proportional hazards models, if there were one, it would be absorbed into the baseline hazard(canceled out in `num` and `den`).

Example

Tidy the data

Suppose that we have two black patients,two white patients,and two patients of other races, with ages 48, 52, 87, 82, 67, and 53, respectively. We may enter these data values as follows:

```
> race <- factor(c("black", "black", "white", "white", "other", "other"))
age <- c(48, 52, 87, 82, 67, 53)
```

- Create matrix using "`model.matrix`" function(In my "`R for data science` presentation)

```
> model.matrix(~ race + age)[,-1]
     raceother racewhite age
1           0         0  48
2           0         0  52
3           0         1  87
4           0         1  82
5           1         0  67
6           1         0  53
```

Example

- If we need to use `whites` as the reference, we can change the race factor to have "`whites`" as the reference level

```
> race <- relevel(race, ref="white")
> model.matrix(~ race + age)[,-1]
     raceblack     raceother   age
1            1             0    48
2            1             0    52
3            0             0    87
4            0             0    82
5            0             1    67
6            0             1    53
```

- Three covariates, say, z_1, z_2, and z_3, the first two of which are `dummy variables` for black race and other race, and the third a `continuous variable`, age.
- For a black 48-year old person, the log hazard ratio is:

$$\log(\psi_1) = z_{11}\beta_1 + z_{12}\beta_2 + z_{13}\beta_3 = 1x\beta_1 + 0x\beta_2 + 48x\beta_3. \quad (2)$$

Example

- β_1 represents the log hazard ratio for blacks as compared to whites, and β_3 represents the change in log hazard ratio that would correspond to a one-year change in age.
- Add Interaction:

```
> model.matrix(~ race + age + race:age)[,-1]
    raceblack raceother age raceblack:age raceother:age
1           1         0  48           48            0
2           1         0  52           52            0
3           0         0  87            0            0
4           0         0  82            0            0
5           0         1  67            0           67
6           0         1  53            0           53
```

Example

Simulation

Generate a small survival data set and show how models are incorporated into a survival problem

- Generate 60 ages between 40 and 80 at random and categorize,and make “white” the reference category:

```
> age <- runif(n=60, min=40, max=80)
> race <- factor(c(rep("white", 20), rep("black", 20),
rep("other", 20)))
> race <- relevel(race, ref="white")
```

- The variables are exponentially distributed with a particular rate parameter that depends on the covariates, specified the log rate parameter to have baseline -4.5, and “age” increase the log rate by 0.05 per year:

```
> log.rate.vec <- -4.5 + c(rep(0,20), rep(1,20), rep(2,20))
+ age*0.05
```

Example

- No censoring

```
>tt <- rexp(n=60, rate=exp(log.rate.vec))
>status <- rep(1, 60)
```

- Fit a *Cox proportional hazards model*

```
> library(survival)
> result.cox <- coxph(Surv(tt, status) ~ race + age)
> summary(result.cox)
n= 60, number of events= 60
               coef    exp(coef)  se(coef)     z      Pr(>|z|)
raceblack   1.15154    3.16305    0.36752    3.133     0.00173 **
raceother   2.49905   12.17087    0.42936    5.820    5.87e-09 ***
age         0.07798    1.08110    0.01448    5.385    7.24e-08 ***
---
Signif. codes:  0 ***  0.001  **  0.01 * 0.05 . 0.1     1
```

- Conclusion:
  - The coefficient estimates, 1.15, 2.50, and 0.08, are close to the true values from the simulation, (1, 2, and 0.05).
  - The `blacks` have 3.16 times the risk of death as do `whites`.

Nested Models

Nested Models

When comparing two models, the covariates of one model must be a subset of the covariates in the other.

- Example:
  Model A: ageGroup4
  Model B: employment
  Model C: ageGroup4 + employment
  Here, Model A is nested in Model C, and Model B is also nested in Model C, so these models can be compared using statistical tests

```
> levels(ageGroup4)
[1] "21-34" "35-49" "50-64" "65+"
> levels(employment) # "ft" refers to full-time, "pt" is part-time
[1] "ft" "other" "pt"
```

Coxph model

- Model A:

```
> modelA.coxph <- coxph(Surv(ttr, relapse) ~ ageGroup4)
> modelA.coxph
                   coef    exp(coef)   se(coef)      z        p
ageGroup435-49   -0.453      0.636       0.164    -2.76    0.0058
ageGroup450-64   -1.568      0.209       0.183    -8.59    0.0000
ageGroup465+     -0.3173     0.728       0.444   -0.7153   0.470

Likelihood ratio test=12.2 on 3 df, p=0.00666
```

- Model B:

```
> modelB.coxph <- coxph(Surv(ttr, relapse) ~ employment)
> modelB.coxph
                   coef     exp(coef)   se(coef)      z        p
employmentother  -0.453       0.636       0.164    -2.76    0.0058
employmentpt     -1.568       0.209       0.183    -8.59    0.0000

Likelihood ratio test=2.06 on 2 df, p=0.357
```

Coxph model

- Model C:

```
> modelC.coxph <- coxph(Surv(ttr, relapse) ~ ageGroup4 +
employment)
> modelC.coxph
                    coef    exp(coef)    se(coef)        z        p
ageGroup435-49   -0.130      0.878       0.321    -0.404   0.6900
ageGroup450-64   -1.024      0.359       0.359    -2.856   0.0043
ageGroup465+     -0.782      0.457       0.505    -1.551   0.1200
employmentother   0.526      1.692       0.275     1.913   0.0560
employmentpt      0.500      1.649       0.332     1.508   0.1300

Likelihood ratio test=16.8 on 5 df, p=0.00492
```

- The log-likelihoods:

```
> logLik(modelA.coxph)
'log Lik.'  -380.043 (df=3)
> logLik(modelB.coxph)
'log Lik.' -385.1232 (df=2)
> logLik(modelC.coxph)
'log Lik.' -377.7597 (df=5)
```

Compare

- Determining if "employment" belongs in the model by comparing Models A and C, the hypothesis test:
  - H_0: The three coefficients for "employment" are zero.
    H_A: Not all zero.
- The likelihood ratio statistic is:

$$2(\ell(\hat{\beta}_{full}) - \ell(\hat{\beta}_{reduced})) = 2(-377.7597 + 380.043) = 4.567 \quad (3)$$

- Compare this to a chi-square distribution with $5 - 3 = 2$ degrees of freedom.

```
1 > pchisq(4.567, df=2, lower.tail=F)
2 [1] 0.1019268
```

- Conclusion: The effect of "employment" is not statistically significant when "ageGroup4" in the model.

ANOVA

- In our **STAT 6337:Advanced Statistic Method I** class, we use "anova" function, which is more direct.

```
> anova(modelA.coxph, modelC.coxph)
Analysis of Deviance Table
Cox model: response is Surv(ttr, relapse)
Model 1: ~ ageGroup4
Model 2: ~ ageGroup4 + employment
   loglik   Chisq    Df   P(>|Chi|)
1 -380.04
2 -377.76   4.5666    2   0.1019
```

AIC

- The Akaike Information Criterion, or *AIC*:

$$AIC = -2 \cdot \ell(\hat{\beta}) + 2 \cdot k \tag{4}$$

- where $\ell(\hat{\beta})$ denotes the value of the partial log likelihood at the M.P.L.E. for a particular model, and k is the number of parameters in the model.
- The AIC balances two quantities which are properties of a model:
  - Goodness of fit: $-2 \cdot \ell(\hat{\beta})$,this quantity is smaller for models that fit the data well
  - The number of parameters is a measure of complexity.
- Conclusion: Smaller is better.

AIC

- Compute the AIC for model A:
  $AIC = 2x380.043 + 2x2 = 766.086$
- Use the "`AIC`" function:

```
> AIC(modelA.coxph)
[1] 766.086
> AIC(modelB.coxph)
[1] 774.2464
> AIC(modelC.coxph)
[1] 765.5194
```

- Conclusion: The best fitting model from among these three, using the AIC criterion, is Model C.

stepwise procedure

- Using the AIC criterion:

```
> modelAll.coxph <- coxph(Surv(ttr, relapse) ~ grp + gender +
race + employment + yearsSmoking + levelSmoking +
ageGroup4 + priorAttempts + longestNoSmoke)
> result.step <- step(modelAll.coxph, scope=list(upper=~ grp +
gender + race + employment + yearsSmoking +
levelSmoking + ageGroup4 + priorAttempts + longestNoSmoke,
lower=~grp) )
```

```
Start: AIC=770.2 Surv(ttr, relapse) ~ grp + gender + race +
employment + yearsSmoking + levelSmoking + ageGroup4 +
priorAttempts + longestNoSmoke
- race                Df       AIC
- yearsSmoking         3    766.98
- gender               1    768.20
- priorAttempts        1    768.20
- levelSmoking         1    768.24
- longestNoSmoke       1    768.47
<none>                     770.20
- employment           2    772.45
- ageGroup4            3    774.11
```

stepwise procedure

- Conclusion:The terms ordered from the one which, when deleted, yields the greatest AIC reduction ("`race`" in this case) to the smallest reduction ("`ageGroup4`"). Thus, "`race`" is deleted.
- Last step:

```
Step: AIC=758.42 Surv(ttr, relapse) ~ grp + employment
+ ageGroup4
                        Df     AIC
<none>                        758.42
+ longestNoSmoke         1    759.10
- employment             2    760.31
+ yearsSmoking           1    760.34
+ gender                 1    760.39
+ priorAttempts          1    760.40
+ levelSmoking           1    760.41
+ race                   3    761.53
- ageGroup4              3    767.24
```

- The "+" sign shows the effect on AIC of adding certain terms. This table shows that no addition or subtraction of terms results in further reduction of the AIC.

BIC

- An alternative to the AIC is the "`Bayesian Information Criterion`", sometimes called the "`Schwartz criterion`", *BIC* is given by:

$$BIC = -2 \cdot \log(L) + k \cdot \log(n) \tag{5}$$

- BIC penalizes the number of parameters by a factor of $\log(n)$ rather than by a factor of 2 as in the AIC.
- The BIC in model selection will tend to result in models with fewer parameters as compared to AIC.