Biostatistics I: Variable Selection

Dimitris Rizopoulos
Department of Biostatistics, Erasmus University Medical Center

d.rizopoulos@erasmusmc.nl

@drizopoulos

Contents

Chapter 1

Variable Selection

1.1 Aims of Statistical Models

Scientific questions revolve around the *understanding* of phenomena

▷ How does the age of patients affect their blood pressure?

▷ Are patients with gene mutations more likely to develop cancer?

▷ . . .

1.1 Aims of Statistical Models (cont'd)

- These questions translate to the understanding of random variables
  - ▷ **however,** the relationships between these variables are quite complex

To make progress, we make a simplification of reality

$\Downarrow$

Statistical Regression Models

1.1 Aims of Statistical Models (cont'd)

- Regression models describe the relationships between
  - ▷ an outcome variable
  - ▷ a set of explanatory variables / predictors / covariates

- The outcome defines the model
  - ▷ continuous outcomes → linear regression model
  - ▷ binary outcomes → logistic regression model
  - ▷ count outcomes → Poisson regression model
  - ▷ survival outcomes → proportional hazards models
  - ▷ ...

1.1 Aims of Statistical Models (cont'd)

- Statistical models are developed for three main purposes
  - ▷ etiology
  - ▷ prediction
  - ▷ description

- Explanatory models: used in etiology research to explain differences in outcome values by differences in explanatory variables
  - ▷ estimate (causal) effects of risk factors or exposures
  - ▷ confounders, colliders and mediators
  - ▷ aim to minimize bias

1.1 Aims of Statistical Models (cont'd)

- Predictive Models: aim to accurately predict an outcome using a set of predictors
  - ▷ expected prediction error is the quantity of main concern

- Descriptive Models: capture the association between covariates and an outcome
  - ▷ elements of both explanatory and predictive models
  - ▷ focus on size of effects, but not causal relationships

1.2 Overfitting and Effective Sample Size

- For all three types of models, a relevant and difficult question is

Which explanatory variables / predictors / covariates to include in the model?

1.2 Overfitting and Effective Sample Size (cont'd)

- Secondary question: **How to include these variables in the model?**

- Linearity: Including a continuous variable as-is in a model assumes linearity
  ⇒ *Many times not reasonable*
  - ▷ polynomials & splines
  - ▷ transformations

- Additivity: Including two variables in a model assumes that their effects on the outcome are independent
  - ▷ interaction terms

1.2 Overfitting and Effective Sample Size (cont'd)

- Fitting too complex models (i.e., models with too many parameters) may result in *Overfitting*

- Overfitting has two important consequences
  - ▷ estimated effects have increased variance $\Rightarrow$ influences confidence intervals and p-values
  - ▷ predicted values from the model do not agree with observed values from future data sets (from the same population) $\Rightarrow$ the model does not validate well

1.2 Overfitting and Effective Sample Size (cont'd)

- To avoid overfitting, we need to restrict the model's complexity
  - ▷ the number of coefficients to estimate

- Note: the number of coefficients is not, in general, equal to the number of covariates
  - ▷ categorical covariates with k levels are represented by $k-1$ dummy variables
  - ▷ nonlinear & interaction terms

1.2 Overfitting and Effective Sample Size (cont'd)

- A *rule of thumb* is to include up to

$$\text{Effective Sample Size} = \frac{n^*}{10}$$

  coefficients in a model

- The value of n^* depends on the information available in the outcome

1.2 Overfitting and Effective Sample Size (cont'd)

▷ Continuous outcome / linear regression:

$$n^* = n, \text{ the sample size}$$

▷ Dichotomous outcomes / logistic regression:

$$n^* = \min\{\# \text{ number of 0s}, \ \# \text{ number of 1s}\}$$

▷ Event times / Cox regression:

$$n^* = \{\# \text{ number of events}\}$$

1.3 Variable Selection Strategies

- For all three types of models, a relevant and difficult question is

Which explanatory variables / predictors / covariates to include in the model?

1.3 Variable Selection Strategies (cont'd)

Due to its practical importance, this question has received a lot of attention

- There are many algorithms available to tackle this problem, ranging from
  - ▷ automatic: the computer does the work for you
  - ▷ manual: the user/researcher needs to do the work

- Automatic Algorithms

  ▷ *Backward elimination*

  * Start: a global model
  * Repeat: remove the most insignificant covariate and re-estimate the model
  * Stop: if no insignificant covariate is left

  ▷ *Forward Selection*

  * Start: the most significant univariable model
  * Repeat: Evaluate the added value of each covariate that is currently not in the model; include the most significant covariate and re-estimate the model
  * Stop: if no significant covariate is left to include

1.3 Variable Selection Strategies (cont'd)

- Automatic Algorithms
  - ▷ *Stepwise flavors*
    - * Combinations of backward elimination & forward selection
  - ▷ *Univariable selection*
    - * Estimate all univariable models
    - * Fit a multivariable model including only the significant covariates from the previous step

1.3 Variable Selection Strategies (cont'd)

- Automatic Algorithms – **Advantages**
  - ▷ *we don't have to think* ⇒ the computer does the work for us automatically
  - ▷ we can consider as many variables as we like

1.3 Variable Selection Strategies (cont'd)

- Automatic Algorithms – **Disadvantages**
  - ▷ *we don't have to think* ⇒ the computer does the work for us automatically
  - ▷ we can consider as many variables as we like

1.3 Variable Selection Strategies (cont'd)

- Automatic Algorithms – **Disadvantages**
  - ▷ yield coefficients that biased high in absolute value
  - ▷ yield p-values that are too small
  - ▷ provide confidence intervals that are too narrow
  - ▷ they suffer even more from collinearity

1.3 Variable Selection Strategies (cont'd)

- Manual Algorithms

  ▷ Make a list of candidate variables using background knowledge

    * critically question the role and further properties of each variable, i.e.,
    * chronology of measurement collection, costs of collection, quality of measurement, availability

  ▷ Make a grouping of variables of primary and secondary interest

  ▷ For the variables of primary interest consider

    * nonlinear terms for continuous variables
    * relevant interaction terms

- Manual Algorithms

  ▷ Setting I: The number of coefficients is *smaller* than the effective sample size

$$\Downarrow$$

Fit the multivariable model containing all terms

1.3 Variable Selection Strategies (cont'd)

- Manual Algorithms

  ▷ Setting II: The number of coefficients is *larger* than the effective sample size

  * reduce the set of secondary variables by eliminating variables with narrow distributions and large number of missing data

  * use data reduction methods (e.g., principal component analysis, clustering)

$\Downarrow$

Fit the multivariable model containing the reduced terms

1.3 Variable Selection Strategies (cont'd)

- Manual Algorithms

  ▷ evaluate the model assumptions using residuals, and appropriately refit the model

  * e.g., consider transformations of the outcome variable

  ▷ consider dropping the complex terms, i.e., the interaction and nonlinear terms

  * perform an omnibus test for all interaction (nonlinear) terms
  * if the p-value > 0.15, you could eliminate all of them
  * otherwise, find which of them seem to play a role

1.3 Variable Selection Strategies (cont'd)

- Manual Algorithms

  ▷ if you build a Descriptive Model ⇒ stop

  * you do **not** need to drop non-significant variables
  * p-values and confidence from the full model (containing non-significant variables) are of better quality

  ▷ if you build a Predictive Model

  * you could drop variables, provided that the predictive accuracy is not compromised
  * the model needs to be 'practical', i.e., easy to use in clinical practice

1.3 Variable Selection Strategies (cont'd)

- Manual Algorithms

  ▷ present the results of the analysis

  * interpret the size (i.e., point estimate) and uncertainty (95% CIs) of the coefficients

  * if necessary (i.e., when you have interaction and nonlinear terms), use effect plots to communicate the results

1.3 Variable Selection Strategies (cont'd)

We have presented a procedure with general guidelines for model-building.

It should be stressed that in some settings, adaptations and exceptions of some of these steps could be relevant.

1.4 Nested vs. Non-Nested Models

When we compare (two) statistical models, an important consideration is whether these models are **nested** or **non-nested**

- **Note:** Model A is nested in Model B, when Model A is a special case of Model B
  - ▷ i.e., by setting some of the parameters of Model B at some specific value we obtain Model A

1.4 Nested vs. Non-Nested Models (cont'd)

- Example 1:

$$M_A : \log(\texttt{serBilir}_i) \; = \; \beta_0 + \beta_1 \texttt{Sex}_i + \beta_2 \texttt{Age}_i + \beta_3 \texttt{Age}_i^2 + \varepsilon_i$$
$$M_B : \log(\texttt{serBilir}_i) \; = \; \beta_0 + \beta_1 \texttt{Sex}_i + \beta_2 \texttt{Age}_i + \varepsilon_i$$

- Model M_B is nested in model M_A
  - ▷ because if we set $\beta_3 = 0$ in model M_A, we get model M_B

1.4 Nested vs. Non-Nested Models (cont'd)

- Example 2:

$$M_A : \log(\texttt{serBilir}_i) \;=\; \beta_0 + \beta_1 \texttt{Sex}_i + \beta_2 \texttt{Age}_i + \beta_3 \texttt{Age}_i^2 + \varepsilon_i$$
$$M_B : \log(\texttt{serBilir}_i) \;=\; \beta_0 + \beta_1 \texttt{Sex}_i + \beta_2 \texttt{Age}_i + \beta_3 \texttt{BMI}_i + \varepsilon_i$$

- Models M_A and M_B are not nested
  - ▷ we *cannot* set some coefficients to a particular value in the one model to get the other

1.4 Nested vs. Non-Nested Models (cont'd)

- Example 3:

$$M_A : \texttt{serBilir}_i \; = \; \beta_0 + \beta_1 \texttt{Sex}_i + \beta_2 \texttt{Age}_i + \beta_3 \texttt{Age}_i^2 + \varepsilon_i$$
$$M_B : \log(\texttt{serBilir}_i) \; = \; \beta_0 + \beta_1 \texttt{Sex}_i + \beta_2 \texttt{Age}_i + \varepsilon_i$$

- Models M_A and M_B are not nested
  - ▷ if we set $\beta_3 = 0$ in the linear predictor of M_A, we get the linear predictor of M_B
  - ▷ *however,* model M_A has outcome variable $\texttt{serBilir}_i$ while model model M_B has outcome variable $\log(\texttt{serBilir}_i)$

1.4 Nested vs. Non-Nested Models (cont'd)

- Most often we compare **nested** models using the likelihood ratio test (LRT):

$$\text{LRT} = -2 \times \{\ell(\hat{\theta}_0) - \ell(\hat{\theta}_a)\} \sim \chi^2_p$$

where

▷ $\ell(\hat{\theta}_0)$ the value of the log-likelihood function under the null hypothesis, i.e., the special case model

▷ $\ell(\hat{\theta}_a)$ the value of the log-likelihood function under the alternative hypothesis, i.e., the general model

▷ p denotes the number of parameters being tested

<u>Note:</u> We can also compare nested model using the Wald and Score tests

1.4 Nested vs. Non-Nested Models (cont'd)

- When we have **non-nested** models we **cannot** use standard tests anymore

- As an alternative for this case we use information criteria – the two standard ones are:

$$\begin{aligned}\text{AIC} &= -2\ell(\hat{\theta}) + 2n_{par}\\ \text{BIC} &= -2\ell(\hat{\theta}) + n_{par}\log(n)\end{aligned}$$

where

▷ $\ell(\hat{\theta})$ is the value of the log-likelihood function

▷ n_{par} the number of parameters in the model

▷ n the number of subjects (independent units)

1.4 Nested vs. Non-Nested Models (cont'd)

When we compare two **non-nested** models we choose the model that has the **lowest** AIC/BIC value