Lesson 3: Variable selection

Zacharie Naulet
zacharie.naulet@math.u-psud.fr

Contents

Motivations

- Recall our favorite example: O_3 concentration. We have a lot of explanatory covariates:
  - The quantitative covariates of the dataset (T12, Ne12, Vx, ...)
  - The categorical[1] variables (1 covariate per possible outcomes[2])
  - All the possible transformations of the covariates ($\sqrt{\text{T12}}$, T12^2, $\log(\text{Vx})$, ...)
  - Interactions between covariates ($\text{T12} \cdot \text{Vx}$, T12/Ne12, ...)
- Using all these covariates (including their possible transformations and combinations), we can build a lot of models.
- **But**, most of those models are more complicated than necessary (most of β_j's are zero). Moreover:
  - It is easier to understand a phenomenon if we use a small model;
  - Too many covariates can cause issues if they are too correlated (think about T12 and T15 in TP5);
  - If $\text{rank}(\boldsymbol{X}^T\boldsymbol{X}) = n$ (which can happens if $p \geq n$), then we ca fit the data perfectly, *i.e* even if the covariates are irrelevant, we can have $\hat{\varepsilon} = 0$.
- Another example: Trying to explain blood sugar as a function of the number of alleles in a given position $j = 1, \ldots, p$. Often, $p \gg 1000$ while $n \approx 10$ at best. It is necessary to select the relevant positions.

Notations and assumptions

- We write the complete model as
$$\underbrace{Y}_{n\times 1} = \underbrace{\boldsymbol{X}}_{n\times p}\underbrace{\beta}_{p\times 1} + \underbrace{\varepsilon}_{n\times 1}.$$
- We assume $p \leq n$ and our usual assumptions H_A and H_B (not necessarily H_C).
- We also assume that there is a $J_* \subset \{1, \ldots, p\}$, corresponding to the "true" model, that is
$$\beta_j \neq 0 \Leftrightarrow j \in J_*$$
In other words, J_* contain the indices corresponding to the relevant explanatory variables.
- As in Lesson 2, we user the notations:
  - $\beta_J = (\beta_j)_{j\in J} \in \mathbb{R}^{|J|}$ the vector constructed from β by keeping only the coordinates with index in J (Here, $|J|$ means the number of elements in J).
  - $\boldsymbol{X}_J$ is the $n \times |J|$ matrix constructed from $\boldsymbol{X}$ by keeping only the columns with indices in J.
  - $\hat{\beta}_J = (\boldsymbol{X}_j^T\boldsymbol{X}_J)^{-1}\boldsymbol{X}_J^T Y$ (the least squares estimator corresponding to the model J)
  - $\hat{Y}_J = \boldsymbol{X}_J\hat{\beta}_J = \boldsymbol{X}_J(\boldsymbol{X}_J^T\boldsymbol{X}_J)^{-1}\boldsymbol{X}_J^T Y$ (adjusted values using the model J).

[1] In french: qualitatives.
[2] In french: modalités.

1 Goals of variable selection

We note that there is not one unique goal, and variable selection could be used in different contexts. We give some examples of applications in the subsections below.

1.1 Selecting relevant variables

- The goal here is to identify J_*.
- We can be mistaken in two ways:
  - We forget a relevant variable $j \in J_*$.
  - We keep a non relevant variable $j \notin J_*$.
- Ideally we want to make no mistake, but this is usually not the case. Thus, it is important to identify if those two errors have consequences in our application. Issues can be:
  - False discovery if we keep the variable j while in truth $j \notin J_*$.
  - Elimination of a relevant variable, which we may not continue to collect in further experiments.
- Note that the two errors are not symmetric (reminiscent to H_0/H_1 in hypothesis testing).

1.2 Estimating parameters

- We want to estimate as accurately as possible the true coefficients β_j.
- The basic idea is that even if all the variables are relevant, some of them may have only a little impact on the response and it might be a good idea to not estimate the corresponding parameters β_j.
- Procedure: we select a model J (which may or not be J_*, indeed in this scenario it could be the case that J_* corresponds to the complete model if all the variables are relevant) and
  - we estimate β_j by $\hat{\beta}_j = 0$ if $j \notin J$;
  - we estimate the other coefficients β_j by $\hat{\beta}_j = \hat{\beta}_{J,j}$ for $j \in J$;
- The error in estimating β is then

$$\mathbb{E}\Big[\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j)^2\Big] = \mathbb{E}\Big[\sum_{j\in J}(\hat{\beta}_{J,j} - \beta_j)^2\Big] + \sum_{j\notin J}\beta_j^2$$

- Why it is a good idea? If the true β_j's are small for $j \notin J$, then we might only incur a low error by estimating that $\beta_j = 0$. Moreover, the variance of $\hat{\beta}$ depends on $|J|$, and it is lower if $|J|$ is small. In other word, our estimator is now biased by the fact that we set $\hat{\beta}_j = 0$ if $j \notin J$, but is has a lower variance. In the case where the bias is small, this might be a good idea (which will be discussed later).

1.3 Estimating adjusted values

- We may also be only interested in estimating the adjusted values in the complete model

$$\boldsymbol{X}\beta = \begin{pmatrix} \mathbb{E}[Y_1] \\ \vdots \\ \mathbb{E}[Y_n] \end{pmatrix} \in \mathbb{R}^n.$$

- For the same reason as in Section 1.2, it might be a good idea to bias a bit the model by considering a fewer number of covariates.
- Indeed, using the model J instead, and using again:
  - $\hat{\beta}_j = 0$ if $j \notin J$;
  - $\hat{\beta}_j = \hat{\beta}_{J,j}$ for $j \in J$;

  The error of estimating $\boldsymbol{X}\beta$ by $\boldsymbol{X}\hat{\beta}$ is,

$$\begin{aligned} \mathbb{E}[\|\boldsymbol{X}\beta - \boldsymbol{X}\hat{\beta}\|^2] &= \mathbb{E}\Big[\sum_{i=1}^{n}(\hat{Y}_{J,i} - \mathbb{E}[Y_i])^2\Big] \\ &= \underbrace{\sigma^2 \cdot |J|}_{\text{variance}} + \underbrace{\|(I_n - P_{\boldsymbol{X}_J})\boldsymbol{X}\beta\|^2}_{\text{bias}}. \end{aligned} \tag{1}$$

- Again, there is a competition between reducing the variance (proportional to $|J|$ and increasing the bias (determined by the fraction of $\boldsymbol{X}\beta$ which is not in $\mathrm{Im}(\boldsymbol{X}_J)$, and hence can be small if J is well chosen).

1.4 Prediction

- Similar to Sections 1.2 and 1.3: we hope to reduce the error by increasing a bit the bias but reducing the variance.
- If $x_{n+1} \in \mathbb{R}^p$, we predict Y_{n+1} with $\hat{y}^p_{J,n+1} = x_{n+1}^T\hat{\beta}_J$.
- The error is then:

$$\mathbb{E}[(\hat{y}^p_{J,n+1} - Y_{n+1})^2] = \sigma^2 + \mathbb{E}[(\hat{Y}_{J,n+1} - \mathbb{E}[Y_{n+1}])^2].$$

- Remark that this is essentially the same problem as for estimating the adjusted value $\mathbb{E}[Y_{n+1}]$ (we get an extra σ^2 because of the variance of Y_{n+1}, which is classical).

2 Consequences of a bad choice of variable

2.1 Missing a relevant variable

That is, we have chosen a set J such that there exists $j_0 \in J_*$ but $j_0 \notin J$.

Proposition 1. *Assume H_A, H_B and there is $j_0 \in J_*$ but $j_0 \notin J$. Then,*

1. *$\mathbb{E}[\hat{\beta}^J] \neq \beta$ and $\mathbb{E}[\hat{Y}_J] \neq \boldsymbol{X}\beta$ (bias).*
2. *$\mathbb{E}[\hat{\sigma}^2_J] > \sigma^2$. That is, our estimator of the residual variance over estimate the residual variance in average.*

2.2 Addition of non-relevant variable (overfit)

- In this scenario, we have chosen a set J such that $J_* \subset J \subseteq \{1, \dots, p\}$, and J is strictly larger than J_*. In other words, J contains at least one irrelevant variable.
- Consequence ?
  - No bias for $\hat{\beta}_J$, $\hat{Y}_J$ or $\hat{\sigma}_J$ (consequence of Lesson 2).
  - However, the variance is larger than if we used J_*.

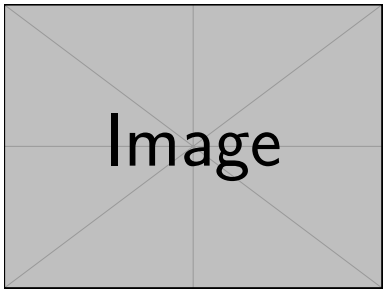
  Proposition 2. *Under H_A and H_B:*

  1. $\mathrm{var}(\beta_{J_*,j}) \leq \mathrm{var}(\hat{\beta}_{J,j}$ *for all j;*
  2. $\mathrm{var}(\hat{Y}_{J_*,i}) \leq \mathrm{var}(\hat{Y}_{J,i})$ *for all i.*

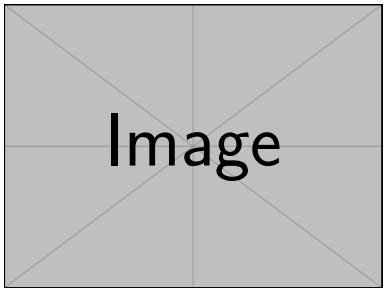
  In general those inequality are strict.

2.3 Summary: Bias-variance trade-off

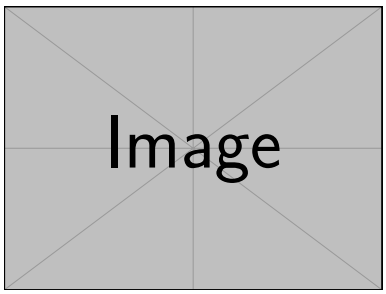
- If we want a small bias: we need to take J as large as possible.
- If we want a small variance: we need to take J as small as possible.
- Then there is a competition between bias and variance, and the goal is to find the right compromise between bias and variance, *i.e.* to choose a good set J.
- This is emphasize for instance by equation (1).
- The optimal J_{opt} is the one which minimizes the risk, *i.e.* bias^2 + variance. Interestingly, J_{opt} is not necessarily J_* but $J_{opt} \subseteq J_*$ always.
- **Example 1**: Low noise (small σ^2), n large, strong signal (*i.e.* large β_j's)



- **Example 2**: High noise (large σ^2), n small, weak signal (*i.e.* small β_j's)



- **Example 3**: Extreme version of the previous example: very high noise, small n and weak signal.



Here the signal is so weak compared to the noise that the best we can do is to estimate β by $\hat{\beta} = 0$ (so that the risk is only the bias).

3 Criteria for variable selection

- Until now we only have investigated what happens if we do not choose carefully our set or relevant variables J, but we don't know how to do the selection.
- The question we want to as is: can we choose a good set J only from the data ? (note that we have to define what good means, in particular, in view of the previous section, we are not necessarily looking for J_*).
- There exist a lot of selection mechanisms, we refer to CML (or any other textbook on the subject).
- We expose in this section a few of them.

3.1 Testing nested models

- If we want to choose between J_0 and J_1 such that $J_0 \subseteq J_1$, we can do Fisher's test between nested model (see Lesson 2).
  - Reject $H_0 \Rightarrow$ we prefer the larger model J_1 (typically we add 1 variable)
  - Accept $H_0 \Rightarrow$ we prefer the smaller model J_0.
- Remark: There is an asymmetry H_0/H_1. Indeed, this procedure will always favor J "small". This is not problematic if the goal is to identify J_* or a subset of J_*.
- Remark: This can also be only used when the model are nested, which should be checked carefully. In situations when the model are not nested, do something else.

3.2 Coefficient of determination

- Recall:

$$R^2_J = \frac{SCM_J}{SCT} = \frac{\|\hat{Y}_J - \bar{y} \cdot \mathbf{1}\|^2}{\|Y - \bar{y}\mathbf{1}\|^1} = 1 - \frac{SCR_J}{SCT} = 1 - \frac{\|Y - \hat{Y}_J\|^2}{\|Y - \bar{y}\mathbf{1}\|^2}.$$

- Note that the large J is, and the smaller $\|Y - \hat{Y}^J\|^2$ gets, regardless of the relevance of the selected covariates.
- Consequence: R_J^2 is always maximal for $J = \{1, \ldots, p\}$.
- **Don't use it!**
- We can use the adjusted coefficient of determination $R_a^2 \equiv R_a^2(J)$

$$R_a^2(J) = 1 - \frac{n-1}{n-|J|}\frac{SCR_J}{SCT}$$

- The $R_a^2(J)$ solves the previous issue.
- We can choose a subset J such that

$$\hat{J} \in \arg\max_J R_a^2(J).$$

that is $\hat{J}$ maximizes the $R_a^2(J)$ (there might be not only one subset maximizing this criterion).

3.3 Penalized criteria

- The idea is to choose J which minimizes a criterion of the form

$$\{\text{goodness of fit of model } J\} + \{\text{penalty } \nearrow \text{ with } |J|\}$$

- Why the penalty term? Because if we put only a criterion on the goodness of fit, we will always select $J = \{1, \ldots, p\}$, because the more covariates we use, and the better we fit the observed data.
- Non exhaustive list of popular penalized criteria:
  - **Mallows' C_p criterion**: Choose J minimizing the criterion

$$C_p(J) := \frac{SCR_J}{\hat{\sigma}^2} + 2|J| = \frac{\|Y - \hat{Y}_J\|^2}{\hat{\sigma}^2} + 2|J|.$$

    Here in general we compute $\hat{\sigma}^2$ using the complete model (but not the only possibility).
  - **Akaike's Information Criterion (AIC)**: Choose (β_J, J) minimizing

$$(\beta_J, J) \mapsto -2\ell_J(\beta_J) + 2|J|,$$

    where ℓ_J is the log-likelihood of the model J (which we need to know, hence we need a model). Note that if the model is Normal, then this is essentially equivalent to Mallows' C_p criterion.
  - **Bayesian Information Criterion (BIC)**: Choose (β_J, J) minimizing

$$-2\ell_J(\beta_J) + \log(n) \cdot |J|.$$

    Note that $\log(n) > \log(2)$ whenever $n > 7$, and thus the BIC criterion tends to select smaller model than the AIC criterion (we pay a stronger penalty by choosing a larger model).

3.4 Comparison between criterions

- Assuming H_C, they all minimize $SCR_J = \|Y - \hat{Y}_J\|^2 +$ a penality term.
- Writing $\hat{J}$ the subset of variable selected by each method, we compare them in Table 1.

F-test or BIC	Smaller $\lvert\hat{J}\rvert$ (more conservative)	Ok to identify J_* if n is large
AIC ou C_p	$\downarrow$	
$R^2_a(J)$	Larger $\lvert\hat{J}\rvert$ (less conservative)	Ok to estimate $\boldsymbol{X}\beta$, at least if the number of models considered is not too large.

Table 1: Comparison of criteria for selecting variables: ranging from most conservative to least conservative.

3.5 Cross-Validation

[...]

4 Algorithms for selection

- Exhaustive search: Among all the 2^p possible models (feasible only when p is not too large).
- Greedy algorithm: Start from a model J_0 and add or remove the covariates one by one.
  - Forward algorithm: Start from the model with $J = \varnothing$ and add the new variable optimizing the chosen criterion (C_p, AIC, BIC, etc.). Stop the procedure when the criterion start to getting worse.
  - Backward: Start from the complete model and remove the covariates.
- Note that using greedy algorithm we are not guaranteed to find the set J minimizing the chosen criterion, only the exhaustive search can guarantee that (but most of time is too complicated and we don't have the choice to give up the guarantee of finding the optimal J).