

EXAM 2023

Problem 1

- C. - The BIC subset is not included in the AIC subset because all subsets of each size are considered, and this is also an indication of correlation between covariates.
- The different sizes of p-values indicates that some of the covariates are correlated.

d. In a GAM model with spline terms, we may write

$$\hat{y} = Sy$$

for a matrix S . Then the number of degrees of freedom is given by

$$df = \text{tr}(S).$$

Roos.Board seems to have a linear effect, which is confirmed by AIC and BIC almost the same as for the full model.

Expend is almost linear in the area with most observations, and the BIC is lower, and the AIC almost the same as for the full model.

The same arguments might be applied to Invol.

c. Different points for the same number of hidden nodes correspond different random start values in the optimisation.

- The RMSE tends to drop for increasing number of hidden nodes up to 120-140, where it flattens. This could be an argument for stopping at 120.

- Better RMSE than for the other considered models, as long as the number of hidden nodes is at least 140. This is an argument for choosing at least 140 nodes if the aim is purely prediction.

- The model V should also be assessed on an independent test set when choosing the final model.

Problem 2

- a. Binomial trial :
- fixed number of trials n
 - independent outcome in each trial
 - two possible outcomes in each trial:
Success or failure
 - the probability of success is the same (p) in each trial

→ The number Y of successes is the binomial distributed with parameters n and p .

This is reasonable as long as the number of available spots in the college is large. If only a limited number of applications will be accepted, the independence assumption and the assumption of equal probabilities of success will be violated.

b. We have $y_i = N_i^{10} + N_i^{11}$ and

$$\hat{y}_i = N_i^{01} + N_i^{11} \quad \text{We get:}$$

$$\begin{aligned} y_i - \hat{y}_i &= N_i^{10} + \cancel{N_i^{11}} - (N_i^{01} + \cancel{N_i^{11}}) \\ &= N_i^{10} - N_i^{01} \end{aligned}$$

The error on individual level is $N_i^{10} + N_i^{01}$.

The accumulated error $y_i - \hat{y}_i$ is smaller because the errors (to some extent) cancel each other out.

C - The likelihood is a product of the marginal point mass functions, which assumes that the Z_{ij} 's are independent.

- We have $\hat{p}_{ij} = h(x_{ij})$, where x_{ij} are the covariates of Z_{ij} . Since x_{ij} is the same for all j , $\hat{p}_{ij}$ will be the same for all j , i.e. $\hat{p}_{ij} = \hat{p}_i$.

- We have:

$$L = \prod_{i=1}^N \prod_{j=1}^{n_i} p_i^{z_{ij}} (1-p_i)^{1-z_{ij}}$$

$$= \prod_{i=1}^N \prod_{j=1}^{n_i} \left(\frac{p_i}{1-p_i} \right)^{z_{ij}} \cdot (1-p_i)$$

$$= \prod_{i=1}^N \prod_{j=1}^{n_i} \left(\frac{p_i}{1-p_i} \right)^{z_{ij}} \cdot \underbrace{\prod_{j=1}^{n_i} (1-p_i)}_{(1-p_i)^{n_i}}$$

$$= \prod_{i=1}^N (1-p_i)^{n_i} \cdot \underbrace{\left(\frac{p_i}{1-p_i} \right)^{z_{i1}} \cdots \left(\frac{p_i}{1-p_i} \right)^{z_{in_i}}}_{\left(\frac{p_i}{1-p_i} \right)^{\sum_{j=1}^{n_i} z_{ij}}}$$

$$= \prod_{i=1}^N (1-p_i)^{n_i} \cdot p_i^{\sum_{j=1}^{n_i} z_{ij}} \left(\frac{1}{1-p_i} \right)^{\sum_{j=1}^{n_i} z_{ij}}$$

$$= \prod_{i=1}^N (1-p_i)^{n_i - \underbrace{\sum_{j=1}^{n_i} z_{ij}}_{y_i}} \cdot p_i^{\underbrace{\sum_{j=1}^{n_i} z_{ij}}_{y_i}}$$

$$= \prod_{i=1}^N (1-p_i)^{n_i - y_i} \cdot p_i^{y_i}$$