Statistical Modelling Assignment - KU Leuven

Claudio Olivelli

March 27, 2021

To define which model should be used for this data, considering the regression purpose, I am going to investigate the empirical distribution and the type of response variable. Since the response variable Y is a counting variable, I can model it as a Poisson random variable:

$$Y \sim POI(\lambda_i) \tag{1}$$

and through the GLM class of parametric regression models, I can model the expected value of the dependent variable Y as:

$$g(E[Y_i|x_i]) = x_i^T\beta$$

where $\mathbf{g()}$ is the link function, x_i is the vector of explanatory variables,β is the vector of coefficients which can be estimated via MLE. Since the distribution of Y_i is a Poisson, I am going to use a Poisson Regression model with logarithmic(canonical) link function assuming that

$$\xi_i = exp(x_i^T\beta)$$

such that:

$$g(E[Y_i|x_i]) = g(\xi_i) = log(exp(x_i^T\beta)) = x_i^T\beta$$

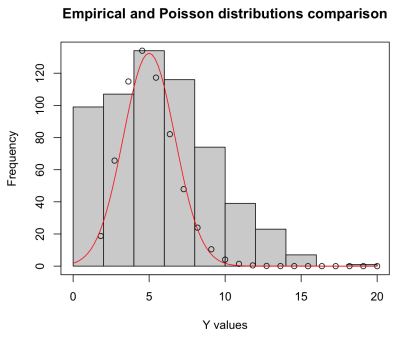

Figure 1: Empirical distribution(grey) and Poisson($\lambda = 3.5$, $\sigma = 3.5$)

In the variable selection I am going to look among the set of models which has as a model with the lowest number of regression coefficients, the one with just the intercept, and as the model with the highest regression coefficients, the one with all the main effects and all the pairwise interactions. I chose this strategy because both AIC and BIC are error estimation methods that takes into account the number of estimated parameters as a penalisation, so to have a wide range of models containing a different number of parameters, allows to have an idea about the penalisation impact on these prediction error estimations. For the AIC the penalisation term is $2 \cdot p$ and for the AIC is $p \cdot log(n)$ where n = 600 and k is the number of parameters in the model.

Among all the possible model, the ones chosen through the AIC and BIC minimization criteria are:

AIC

- $Y \sim X1 + X2 + X3 + X4 + X1 : X2 + X3 : X4$,with $AIC = 2865.99$
- $Y \sim X1 + X2 + X3 + X4 + X1 : X2 + X3 : X4 + X5$,with $AIC = 2867.29$

- $Y \sim X1 + X2 + X3 + X4 + X1:X2 + X3:X4 + X5 + X4:X5$,with $AIC = 2867.38$

BIC

- $Y \sim X1 + X2 + X3 + X4 + X1:X2 + X3:X4$,with $BIC = 2896.77$
- $Y \sim X1 + X2 + X3 + X4 + X1:X2 + X3:X4 + X5$,with $BIC = 2902.46$
- $Y \sim X1 + X2 + X3 + X4 + X1:X2 + X3:X4 + X5 + X4:X5$,with $BIC = 2906.95$

This results has been obtained through the *stepAIC* R function, setting the $k = 2$ for AIC and $k = log(n)$ for BIC. Using both the criteria we obtain the same Poisson regression models with the same covariates. The BIC results, as expected, lower in value than the AIC because the penalisation factor has more impact in it.

The **Focussed Information Criterion** states that "[...] a best model should depend on the parameter under focus[...]"(Claeskens and Hjort, 2003). Indeed, through this method I first define two estimands, also called "focuses" , which will be the mean and the variance of all the covariates and their interactions, and then calculate the best model in order to minimize the MSE.adj of the estimators focused.
Through an initial correlation analysis of the covariates, I figured out to delete those interactions with a correlation ρ>0.90, avoiding in this way every matrix singularity problem. The following interactions among covariates have been removed from the wide model : X1:X4, X1:X5 , X2:X4, X3:X4, X3:X5, X4:X5.
So the final wide model is :

$$Y \sim X1 + X2 + X3 + X4 + X5 + X1:X2 + X1:X3 + X2:X3 \tag{2}$$

Among all the possible model combination, from the model containing only the intercept(narrow model) to the wide model, I am going to choose the best 3 models that minimize the rmse.adj, which contain both the estimated variance and the estimated bias for every model, calculated with the *fic* R function. The FIC value doesn't contain the constants term, so the list of the best models may result in a different order than through rmse.adj .

Mean of the covariates

- $Y \sim X4$, $rmse.adj = 0.09267694, focus = 6.039617$
- $Y \sim X3$, $rmse.adj = 0.0.09294290, focus = 6.024322$
- $Y \sim X5$, $rmse.adj = 0.0.09294608, focus = 6.023316$

Variance of the covariates

- $Y \sim X4$, $rmse.adj = 0.2061807, focus = 6.047795$
- $Y \sim X3$, $rmse.adj = 0.2133233, focus = 5.945052$
- $Y \sim X3 + X4$, $rmse.adj = 0.2135859, focus = 5.950950$

The focus output of the *fic* function, reported before, express the estimates of the focuses in these reported models, respectively the mean and the variances of the covariates and their "cleaned" interaction(excluding the high correlated ones). From a first overview, it comes out that both the mean and the variance focuses takes as a best two model the ones with only X4(the width of the lane on which the cars drive) and X3(the width of the "median" which separates the traffic in opposite directions) covariates while the 3rd best one changes in behaviour, indeed for the mean includes just X5 but for the variance include the sum of X3 + X4.
It is interesting to see how the focus values do not follow the rmse.adj progression, which confirm the choice to use rmse.adj as a metrics and not the focus which would have given a complete different result.
To summarise, there is a complete different behaviour between BIC/AIC and FIC criterias. The models chosen by BIC/AIC are the same, or in a similar case would be of a similar magnitude, because the two criterias approximately work in the same way(MLE and penalisation terms on the whole model). Instead, the FIC definetely reduces the number of covariates of the optimal model, taking into account X4,X5,X3 which are the most suitable variable to estimate the mean of the covariates(variance between variables is high and variance in the singles variables is even higher especially in X1,X2).

```
library(MASS)
library(fic)
studentnumber =  0787524
fulldata = read.table("dataHW.txt",header=T)
set.seed(studentnumber)
rownumbers = sample(1:nrow(fulldata),600,replace=F)
mydata = fulldata[rownumbers,]
hist(mydata$Y,main = "Empirical␣and␣Poisson␣distributions␣comparison", xlab="Y␣values",)
par(new=TRUE)
plot(0:20, dpois( x=0:20, lambda=3.5), xlim=c(-2,20),xaxt='n', ann=FALSE, yaxt='n',main = paste(
"Empirical␣and␣Poisson␣distributions␣comparison"))
normden <- function(x){dnorm(x, mean=3.5, sd=sqrt(3.5))}
curve(normden, from=-2, to=20, add=TRUE, col="red")
hw.full=glm(Y~.^2, data = mydata, family = poisson())
summary(hw.full)
#use stepAIC to find the best models based on AIC
hw.AIC=stepAIC(hw.full, k=2, scope =list(upper=~.^2,lower=~1),direction = "backward")
#fit the best three models selected by AIC(the ones with the LOWEST AIC)
fit1.aic=glm(Y~X1+X2+X3+X4+X1*X2+X3*X4, data = mydata, family = poisson())
fit2.aic=update(fit1.aic,.~.+ X5)
fit3.aic=update(fit2.aic,.~.+ X4*X5)
#calculate aic
AIC1=AIC(fit1.aic,k=2)
AIC2=AIC(fit2.aic,k=2)
AIC3=AIC(fit3.aic,k=2)
#use stepAIC to find the best models based on BIC
hw.BIC=stepAIC(hw.full, k=log(nrow(mydata)), scope =list(upper=~.^2,lower=~1) ,
               direction = "backward")
### fit 3 best models according to 3 final steps of stepAIC based on BIC
fit1.bic=glm(Y~X1+X2+X3+X4+X3*X4+X1*X2, data = mydata, family = poisson())
fit2.bic=update(fit1.bic,.~.+X5)
fit3.bic=update(fit2.bic,.~.+X5*X4)
#calculate bic : in both BIC and AIC the lower the IC value the best is the model
BIC1=AIC(fit1.bic,k=log(nrow(mydata)))
BIC2=AIC(fit2.bic,k=log(nrow(mydata)))
BIC3=AIC(fit3.bic,k=log(nrow(mydata)))
library(fic)
focus=function(par,X) exp(X%*%par)
variable=model.matrix(hw.full)
matrix.cor=cor(variable)
library('corrplot')
corrplot(matrix.cor, method = "circle")
hw.full.singular=glm(Y~.^2-X1:X4-X1:X5-X2:X4-X3:X4-X3:X5-X4:X5-X2:X5, data = mydata, family = poisson())
inds0=c(1,rep(0,length(hw.full.singular$coefficients)-1))
combs=all_inds(wide = hw.full.singular, inds0 = inds0)
#exclude models with interactions that do not include both main effects
combs.excl=with(combs,combs[!(combs[,2]==0 &(combs[,8]==1|combs[,7]==1)|combs[,3]==0
&(combs[,7]==1|combs[,9]==1)|combs[,4]==0 & (combs[,8]==1|combs[,9]==1 )),])
# specify X matrix, which contains the focus we'll use for the evaluation
X_eval1=round(c(1,mean(mydata$X1),mean(mydata$X2),mean(mydata$X3),mean(mydata$X4),mean(mydata$X5)
,mean(mydata$X1*mydata$X2),mean(mydata$X1*mydata$X3), mean(mydata$X2*mydata$X3)),digits = 2)
X_eval2=round(c(1,var(mydata$X1),var(mydata$X2),var(mydata$X3),var(mydata$X4),var(mydata$X5)
,var(mydata$X1*mydata$X2),var(mydata$X1*mydata$X3),var(mydata$X2*mydata$X3)),digits = 2)
fic1=fic(wide = hw.full.singular, inds = combs.excl, inds0 = inds0, focus = focus, X=X_eval1)
index1=order(fic1$rmse.adj)[1:3]
# Implement fic for focus 2
fic2=fic(wide = hw.full.singular, inds = combs.excl, inds0 = inds0, focus = focus, X=X_eval2)
index2=order(fic2$rmse.adj)[1:3]
fic2[index2,]
```