Statistical Modelling exam
Jerko Radan Cruz - r0773435

Question A

A.1

Three different estimators for the mean of *log (number of new hospital intakes +1)* were developed. Since a lm do not look appropriate, penalized splines were used, in section A.3 the models are described.

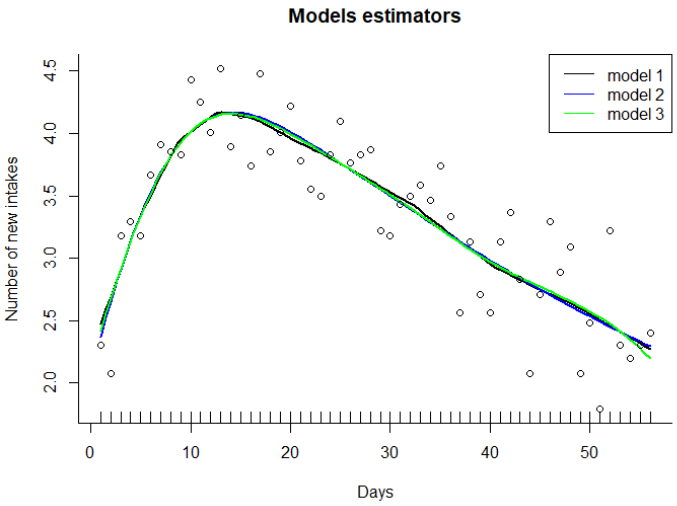

Figure 1. Models to estimate the mean of log (number of new hospital intakes +1)

A.2

A naïve pointwise 99% confidence interval is constructed for each estimate.

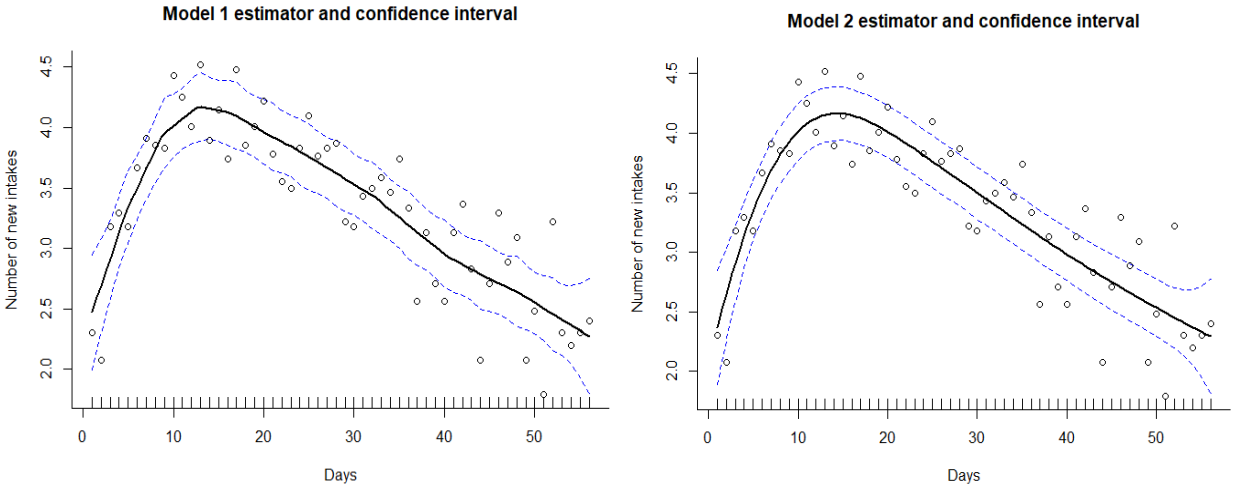

Figure 2. Model 1 (penalized spline with truncated polynomial basis, degree 1) and Model 2 (penalized spline with truncated polynomial basis, degree 2)

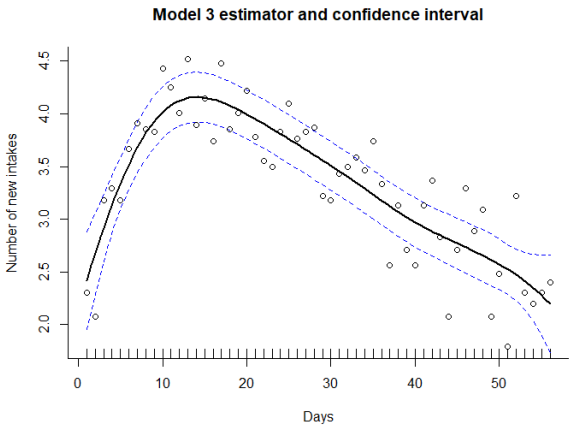

Figure 3. Model 3 (penalized spline with radial basis of degree 3)

Is important to highlight that the naïve pointwise approach is based on a t-distribution since the population variance is unknown. Additionally, it does not take into account the fact that the smoothness selection was performed, increasing the uncertainty. Several approaches have been developed for creating confidence bands for penalized spline estimators using volume-of-tube formula [1] [2], via simulation from the posterior distribution of the parameters [3] and via a smoothing parameter uncertainty corrected covariance matrix [4].

A.3

$$Y_i: log(number\ of\ new\ hospital\ intakes\ + 1)$$
$$x_i: days\ of\ COVID - 19\ reports\ (numerical\ from\ 1\ to\ 56)$$

Model 1: Penalized regression splines with truncated polynomial basis of degree 1, using a Restricted Maximum Likelihood method (REML) to estimate λ, with K nodes.

$$Y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+ + \varepsilon_i$$

The following applies to all three models: $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and $b_k \sim N(0, \sigma_b^2)$ for all k = 1, …, K. We assume that the b_k are all independent. Smooth estimator $\lambda = \sigma_\varepsilon^2 / \sigma_b^2$. The knot location $\kappa_k = \left(\frac{k+1}{K+2}\right)$, where $K = max\left(\frac{n}{4}, 20\right)$ [5], with n = 56.

Model 2: Penalized regression splines with truncated polynomial basis of degree 2, using a REML method to estimate λ, with K nodes.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + \sum_{k=1}^{K} b_k \{(x_i - \kappa_k)_+\}^2 + \varepsilon_i$$

Model 3: Penalized regression splines with radial basis of degree 3, using a REML method to estimate λ, with K nodes.

$$Y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k |x_i - \kappa_k|^3 + \varepsilon_i$$

The spm() function was used, it uses lme ("Linear Mixed-Effects model") to fit the model with random effects for the spline part of the model. The models were constructed using 13 knots, model 1 presents 7.03 degrees of freedom (spar = 5.37), meanwhile model 2 presents 5.66 df (spar = 9.5) and model 3 presents 5.87 df (spar = 15.5). Model 2 and 3 are clearly smoother than model 1. Model 2 looks to be sufficient for capturing the pattern in the data. The resulted models are the followings:

Model 1: $E[y] = 2.254 + 0.219x - 0.062 \cdot (x - 4.93)_+ - 0.098 \cdot (x - 8.86)_+ - 0.076 \cdot (x - 12.79)_+ - 0.027 \cdot (x - 16.71)_+ + 0.004 \cdot (x - 20.64)_+ - 0.006 \cdot (x - 24.57)_+ + 0.0001 \cdot (x - 28.50)_+ - 0.017 \cdot (x - 32.42)_+ + 0.003 \cdot (x - 36.36)_+ + 0.018 \cdot (x - 40.29)_+ + 0.006 \cdot (x - 44.21)_+ - 0.012 \cdot (x - 48.14)_+ + 0.0004 \cdot (x - 52.07)_+$

Model 2: $E[y] = 2.058 + 0.322x - 0.013x^2 + 5.82 \cdot 10^{-4} \cdot \{(x - 4.93)_+\}^2 + 2.267 \cdot 10^{-3} \cdot \{(x - 8.86)_+\}^2 + 3.796 \cdot 10^{-3} \cdot \{(x - 12.79)_+\}^2 + 3.543 \cdot 10^{-3} \cdot \{(x - 16.71)_+\}^2 + 2.16 \cdot 10^{-3} \cdot \{(x - 20.64)_+\}^2 + 4.60 \cdot 10^{-4} \cdot \{(x - 24.57)_+\}^2 + 4.256 \cdot 10^{-5} \cdot \{(x - 28.50)_+\}^2 - 4.13 \cdot 10^{-5} \cdot \{(x - 32.42)_+\}^2 + 5.846 \cdot 10^{-4} \cdot \{(x - 36.36)_+\}^2 - 9.057 \cdot 10^{-5} \cdot \{(x - 40.29)_+\}^2 - 3.248 \cdot 10^{-4} \cdot \{(x - 44.21)_+\}^2 + 1.908 \cdot 10^{-4} \cdot \{(x - 48.14)_+\}^2 + 7.173 \cdot 10^{-4} \cdot \{(x - 52.07)_+\}^2$

Model 2: $E[y] = 5.606 - 0.074x - 9.804 \cdot 10^{-3} \cdot |x - 4.93|^3 + 2.381 \cdot 10^{-3} \cdot |x - 8.86|^3 + 2.611 \cdot 10^{-3} \cdot |x - 12.79|^3 + 5.69 \cdot 10^{-4} \cdot |x - 16.71|^3 - 1.041 \cdot 10^{-3} \cdot |x - 20.64|^3 - 3.788 \cdot 10^{-5} \cdot |x - 24.57|^3 + 5.758 \cdot 10^{-4} \cdot |x - 28.50|^3 + 1.821 \cdot 10^{-3} \cdot |x - 32.42|^3 + 6.75 \cdot 10^{-4} \cdot |x - 36.36|^3 - 6.414 \cdot 10^{-4} \cdot |x - 40.29|^3 + 7.076 \cdot 10^{-4} \cdot |x - 44.21|^3 + 9.952 \cdot 10^{-4} \cdot |x - 48.14|^3 + 2.105 \cdot 10^{-3} \cdot |x - 52.07|^3$

References

[1] Krivobokova T, Kneib T and Claeskens G (2009). *Simultaneous Confidence Bands for Penalized Spline Estimators.* Department of Decision Sciences and Information Management (KBI).

[2] Krivobokova T and Wiesenfarth M (2018). *Adaptive Semiparametric Additive Regression with Simultaneous Confidence Bands and Specification Tests. Package 'AdaptFitOS'*. http://cran.r-project.org

[3] Simpson G (2016). *Simultaneous intervals for smooth revisited*. URL: https://fromthebottomoftheheap.net/2016/12/15/simultaneous-interval-revisited/

[4] Wood S (2019). *Mixed GAM Computation Vehicle with Automatic Smoothness Estimation R Package 'mgcv'*. http://cran.r-project.org

[5] Wand, M.P., Coull, B.A., French, J.L., Ganguli, B., Kammann, E.E., Staudenmayer, J. and Zanobetti, A. (2005). *SemiPar 1.0. R package*. http://cran.r-project.org

Question B

B.1

Based on the binary response variable *Default*, a Generalized Linear Models with a Binomial distribution (logit link function) is selected.

The variables are:
$Default$: $binary\ response\ variable$ (1: $customer\ failed\ to\ pay\ back\ the\ loan$).
$Income$: $yearly\ gross\ income$ ($numerical\ variable$).
$Education$: $categorical\ varible\ of\ the\ highest\ degree\ obtained. It\ is\ transformed\ in\ binary\ varibles$: $EducationM$, $EducationN, EducationP\ and\ EducationS$.
$Children$: $number\ of\ children$ ($numerical\ variable$).
$Employment$: $number\ of\ months\ working\ at\ the\ current\ employer$ ($numerical\ variable$).
$Phone$: $binary\ variable$ ($It\ is\ transformed\ in\ PhoneY$, 1 $if\ customer\ has\ a\ home\ phone$).

$Term$: $binary\ variable$ ($It\ is\ transformed\ in\ Term36, 1\ if\ customer\ has\ a\ 36\ months\ loan\ term$).
Age: $age\ of\ the\ customer$ ($numerical\ variable$).
$Loan$: $loan\ amount$ ($numerical\ variable$).
$Gender$: $binary\ variable$ ($It\ is\ transformed\ to\ GenderM, 1\ if\ customer\ is\ male$).
$Address$: $number\ of\ months\ the\ customer\ has\ been\ living\ in\ the\ current\ address$.
$Customer$: $binary\ variable$ ($It\ is\ transformed\ to\ CustomerN, 1\ if\ customer\ is\ new\ to\ the\ bank$).
$Limit$: $annual\ limit\ on\ the\ current\ account$ ($numerical\ variable$).

A full model with all main effects (glm.full) and a full model with all main effects and interactions (glm.full2) were constructed. For a binary response variable Y_i following a binomial distribution $Bin(m_i,p_i)$, depending of a set of k explanatory variables, $X_i = (X_1, \ldots, X_k)_i$. With $i = 1, \ldots, n\ independent\ observations$ ($sample\ of\ n = 1000$) and $m_i\ number\ of\ experiments = 1$.

$$g(p_i) = logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = x_i^T\beta, with\ E[Y_i] = p_i, \varepsilon_i \sim N(0, \sigma^2)$$

$$\Longrightarrow \xi_i = p_i = 1 - exp\{-exp(x_i^T\beta)\}$$

$$\Rightarrow g(\xi_i) = log(-log(1-p_i)) = x_i^T\beta$$

A set of models were proposed, with the following searching procedure:

- Model 1: Both directions stepwise procedure based on AIC (*stepAIC* with k =2) starting with glm.full (just main effects).
- Model 2: Backward direction stepwise procedure based on AIC (*stepAIC* with k =2) starting with glm.full (just main effects).
- Model 3: Backward direction stepwise procedure based on AIC (*stepAIC* with k =2) starting with glm.full2 (full model including interactions).

These three models present a better AIC than the full model with just main effects and the full model with all main effects and interactions. Shown below is the ranking based on AIC:

Model Ranked by AIC	Variables	AIC value
Model 1	Income, Loan, GenderM, EducationM, EducationN, EducationP, EducationS, Children, Employment, Adress, PhoneY, CustomerN, Term36, Limit, Income:Limit, Employment:CustomerN, Employment:PhoneY, PhoneY:CustomerN, Employment:Adress, Children:Term36, PhoneY:Term36, Income:PhoneY, Loan:PhoneY, EducationM:Limit, EducationN:Limit, EducationP:Limit, EducationS:Limit, Income:Term36, Term36:Limit, Loan:Term36, Loan:Limit, Income:EducationM, Income:EducationN, Income:EducationP, Income:EducationS, Loan:Adress	1168.288
Model 3	Income, Loan, Age, GenderM, EducationM, EducationN, EducationP, EducationS, Children, Employment, Adress, PhoneY, CustomerN, Term36, Limit, Income:Loan, Income:Age, Income:PhoneY, Income:CustomerN, Income:Term36, Income:Limit, Loan:Adress, Loan:PhoneY, Loan:Term36, Loan:Limit, Age:GenderM, Age:EducationM, Age:EducationN, Age:EducationP, Age:EducationS, Age:PhoneY, Age:Term36, GenderM:Children, EducationM:Limit, EducationN:Limit, EducationP:Limit, EducationS:Limit, Children:Term36, Employment:Adress, Employment:PhoneY, Employment:CustomerN, PhoneY:CustomerN, PhoneY:Term36, CustomerN:Term36, CustomerN:Limit, Term36:Limit	1169.463
Model 2	Income, EducationM, EducationN, EducationP, EducationS, Children, Employment, PhoneY, Term36	1209.734

Based on AIC, the best model for the probability of *Default* p_i will be:

$$log\left(-log(1-p_i)\right) = -12.9 + 1.88 \cdot Income - 5.35 \cdot Loan + 0.257 \cdot GenderM + 4.47 \cdot EducationM - 1.96 \cdot EducationN + 9.95 \cdot 10^{2} \cdot EducationP + 12.98 \cdot EducationS - 0.478 \cdot Children - 5.42 \cdot 10^{-2} \cdot Employment - 1.83 \cdot 10^{-2} \cdot Address + 1.1 \cdot PhoneY + 0.386 \cdot CustomerN - 7.13 \cdot Term36 + 0.41 \cdot Limit - 0.135 \cdot Income \cdot Limit + 3.24 \cdot 10^{-2} \cdot Employment \cdot CustomerN + 1.45 \cdot 10^{-2} \cdot Employment \cdot PhoneY - 1.33 \cdot PhoneY \cdot CustomerN + 2.97 \cdot 10^{-5} \cdot Employment \cdot Address + 0.425 \cdot Children \cdot Term36 + 1.93 \cdot PhoneY \cdot Term36 - 0.194 \cdot Income \cdot PhoneY + 0.57 \cdot Loan \cdot PhoneY - 0.402 \cdot EducationM \cdot Limit + 0.68 \cdot EducationN \cdot Limit - 1.07 \cdot 10^{2} \cdot EducationP \cdot Limit - 0.9 \cdot EducationS \cdot Limit - 0.41 \cdot Income \cdot Term36 + 0.969 \cdot Term36 \cdot Limit + 1.01 \cdot Loan \cdot Term36 + 0.436 \cdot Loan \cdot Limit + 2.04 \cdot 10^{-2} \cdot Income \cdot EducationM - 0.17 \cdot Income \cdot EducationN + 6.28 \cdot Income \cdot EducationP - 6.08 \cdot 10^{-2} \cdot Income \cdot EducationS + 1.41 \cdot 10^{-3} \cdot Loan \cdot Address$$

B.2

Considering parametric models with main effects only, all subsets of glm.full including the intercept gives the following 99% Naïve and PostAIC confidence intervals:

	Estimate	Naive 0.5%	Naive 99.5%	Naive CI width	PostAIC 0.5%	PostAIC 99.5%	PostAIC CI width
(Intercept)	-3.074	-4.504	-1.681	2.823	-5.039	-1.111	3.928
Income	0.069	0.035	0.105	0.07	0.017	0.122	0.105
EducationM	0.687	0.387	0.989	0.602	0.163	1.21	1.047
EducationN	0.771	0.305	1.236	0.931	-0.028	1.571	1.599
EducationP	0.864	-0.876	2.602	3.478	-1.732	3.46	5.192
EducationS	-0.59	-1.527	0.214	1.741	-1.906	0.726	2.632
Children	-0.173	-0.342	-0.006	0.336	-0.425	0.079	0.504
Employment	-0.006	-0.009	-0.003	0.006	-0.012	-0.0000027	0.01199731
PhoneY	-0.336	-0.686	0.017	0.703	-0.917	0.245	1.162
Term36	-0.547	-0.987	-0.113	0.874	-1.244	0.15	1.394

The Naïve confidence intervals were calculated using the *confint* function. It uses the following formula (since the variance of the population is unknown it uses the t-distribution instead of the normal).

$$CI_n(\mu) = \left[\hat{\mu}_{\hat{S}} - t_{\alpha/2}\,\hat{\sigma}_{\hat{S}}/\sqrt{n}, \hat{\mu}_{\hat{S}} + t_{\alpha/2}\,\hat{\sigma}_{\hat{S}}/\sqrt{n}\right]$$

However, these confidence intervals don't consider the uncertainty due to AIC selection procedure. To account for this uncertainty the Post-AIC confidence intervals are used. Using the PostAIC function different quantiles are used to construct the confidence intervals, based on the conditioned selection region. As expected, the Post-AIC confidence intervals are clearly wider.

Question C

C.1

Based on the histogram for the response variable Y (figure 2) and that it is discrete, a Generalized Linear Models with a Poisson distribution is selected. In table 1 the coefficients of different regression models (Ridge regression, Lasso estimation and Elastic Nets) are presented, with the corresponding penalty estimation (λ) using a 7-fold cross-validation.

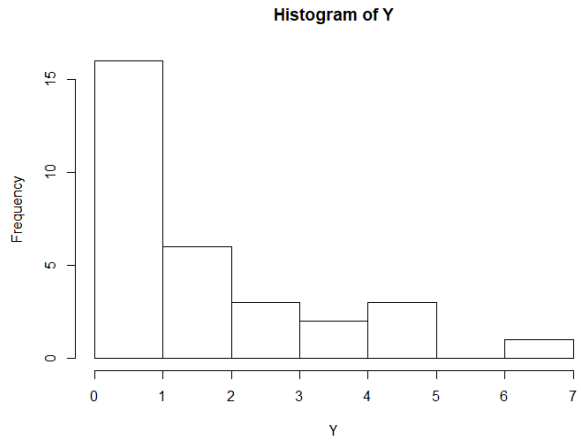

Figure 4. Histogram for Y (bike station 29)

Table 1. Estimated coefficients and penalty estimators

	Ridge (alpha = 0) (λ = 12.21)	**Lasso (alpha = 1) (λ = 0.4187)**	**Elastic net (alpha = 0.5) (λ = 0.7284)**	**Elastic net (alpha = 0.8) (λ = 0.5484)**
(Intercept)	0.6264558	-0.150209967	0.317467616	0.070842894
V1	.	.	.	.
V2	.	.	.	.
V3	.	.	.	.
V4	.	.	.	.
V5	.	.	.	.
V6	.	.	.	.
V7	.	.	.	.
V8	.	.	.	.
V9	.	.	.	.
V10	.	.	.	.
V11	.	.	.	.
V12	.	.	.	.
V13	.	.	.	.
V14	.	0.037354185	0.005118487	0.027180607
V15	.	.	.	.
V16	.	0.028915806	.	0.005737992
V17	.	.	.	.
V18	.	.	.	.
V19	.	0.140324967	0.035463322	0.099110599
V20	.	0.026800491	.	0.003721014
V21	.	.	.	.
V22	.	.	.	.
V23	.	.	.	.
V24	.	.	.	.
V25	.	.	.	.
V26	.	0.096645302	0.097686498	0.101703659
V27	.	.	.	.
V28	.	.	.	.

	Ridge (alpha = 0) (λ = 12.21)	**Lasso (alpha = 1) (λ = 0.4187)**	**Elastic net (alpha = 0.5) (λ = 0.7284)**	**Elastic net (alpha = 0.8) (λ = 0.5484)**
V30	.	0.05912991	0.020690979	0.04941603
V31	.	.	.	.
V32	.	.	.	.
V33	.	.	.	.
V34	.	.	.	.
V35	.	.	.	.
V36	.	.	.	.
V37	.	.	.	.
V38	.	.	.	.
V39	.	.	.	.
V40	.	.	.	.
V41	.	0.009671175	.	.
V42	.	.	.	.
V43	.	.	.	.
V44	.	.	.	.
V45	.	.	.	.
V46	.	.	.	.
V47	.	.	.	.
V48	.	.	.	.
V49	.	.	.	.
V50	.	.	.	.

C.2

The penalty estimator (λ) is calculated by a 7-fold cross-validation. Splitting the data in k = 7 pieces, using k – 1 of those to build the model and validating on the kth piece, via predictive likelihood. Validating on each of the k pieces, and then averaging the k different deviances [6]. In general, the cross-validation function is defined as:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{\{y_i - \hat{f}(x_i)\}^2}{(1-H_{ii})^2}$$

Where $\hat{f}(x_i)$ is the estimated value at x_i using all of the data, and H is the hat matrix such that $\hat{f} = HY = X(X^TX + \lambda D)^{-1}X^TY$. The cv.glmnet function minimizes the deviance to obtain the value of λ (lambda.min) that gives minimum mean cross-validated deviance residual. For more details of the cv.glmnet calculation, please refer to [7]

Denoting $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$ and $\|\beta\|_2 = \sum_{j=1}^{p}\beta_j^2$ with $\beta = (\beta_1, \dots, \beta_p)^T$. Then the penalized estimator is:

$$(\hat{\beta}_0, \hat{\beta}^T)^T = argmin_{\beta_0,\beta}\left\{\frac{1}{n}\sum_{i=1}^{n} logf(Y_i; x_i, \beta_0, \beta, \phi) + \lambda P_\alpha(\beta)\right\}$$

Where $P_\alpha(\beta) = \alpha\|\beta\|_1 + \frac{1}{2}(1-\alpha)\|\beta\|_2$

Glmnet uses an outer Newton loop, and an inner weighted least-squares loop optimize this criterion [8]. For the GLM with a Poisson distribution regression model (canonical link function)

$$f(y_i, \theta_i) = exp\{y_i\theta_i - exp(\theta_i) - log(y!)\}$$

$$E[Y_i] = e^{\beta_0 + x_i^T\beta} = \xi_i, \qquad for\ i = 1, \dots, n\ independent\ observations$$

$$x_i = (x_{i1}, \dots, x_{i50})^T$$

$$\Rightarrow y_i = e^{\beta_0 + x_i^T\beta} + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

With probability function for Y_i

$$P(Y_i = y) = \frac{e^{-\xi_i}\xi_i^y}{y!}, with\ y = 0,1,2,3, \dots$$

For the following methods, the α value is modified, obtaining a different λ from CV and a different penalized model.

(a) Ridge regression. α = 0, $\lambda_{min} = 12.21$. Only the intercept is kept.

(b) Lasso estimation. α = 1, $\lambda_{min} = 0.4187$. The intercept, V14, V16, V19, V20, V26, V30 and V41 are kept.

(c) Elastic net, with α = 0.5, $\lambda_{min} = 0.7284$. The intercept, V14, V19, V26 and V30 are kept.

(d) Elastic net, with α = 0.8, $\lambda_{min} = 0.5484$. The intercept, V14, V16, V19, V20, V26 and V30 are kept.

Is important to highlight that the penalization constraint $P_\alpha(\beta)$, since the naïve elastic net estimator presents a double amount of shrinkage [9], the ℓ_2 penalization presents a ½.

It is shown that the lasso estimator (under some conditions and assumptions) is consistent in the sense that with large probability it estimates the true non-zero parameters as non-zero. However, the penalization has a shrinkage effect, hence the estimated values obtained via a penalization approach are smaller in magnitude than when using non-penalized estimators (such as maximum likelihood, least squares, etc) [10]. The lasso tends to give higher sparsity, in this example it presents more variables than the ridge and elastic net formulations.

The ridge formulation on the other hand, presents a much higher penalization estimate λ. It tends to perform well in cases of multicollinearity. It trades variance for bias, driving the overall size of the weight values down and constraining the variance of the model. However, in this example, trying to reduce the variance, the penalization estimate was increased and all variables were driven to zero, having as a final result just the intercept, with the highest value compare to the other methods.

Finally, the elastic net is a combination of the other two. It is less aggressive in the elimination of features and presents smaller weight values overall. As the α is closer to 0 the λ_{min} is higher, this is reflected in the coefficients of the variables. The highest coefficients are obtained with α = 1, and the lower with α = 0.

References

[6] Simon N, Friedman J, Hastie T and Tibshirani R (2011). *Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent.* Journal of Statistical Software. Volume 39, Issue 5.

[7] Hastie T (2019). *Github repository*. URL: https://github.com/cran/glmnet/tree/master/R

[8] Hastie T and Qian J (2016). *Glmnet Vignette*. Stanford, URL: https://web.stanford.edu/~hastie/glmnet/glmnet_beta.html#poi

[9] Hastie T and Zou H (2004). *Regularization and variable selection via the elastic net.* J.R. Statist. Soc. B 67, Part 2, pp 301-320.

[10] Claeskens G (2018). Statistical Modelling. Acco, Leuven.

Appendix A

```
# ------------ Question A ------------------
#install.packages("SemiPar")

library(SemiPar)
library(MASS)

DataQA = read.table("covidDATA.txt")
modulo11=function(x) {x- floor(x/11)*11}
studentnumber=773435
mycol=modulo11(studentnumber)
mydataA=DataQA[,c((1+mycol*4):(4+mycol*4))]

hist(mydataA$NEW_IN.3)
hist(log(mydataA$NEW_IN.3+1))

mydataA$DATE.3 <- as.Date(mydataA$DATE.3, '%Y-%m-%d')
mydataA['Days'] <- seq(from=1, to=56, by=1)
mydataA['log_new_in'] <- log(mydataA$NEW_IN.3+1)

attach(mydataA)
par(mfrow=c(1,1))
plot(DATE.3,log_new_in)

#Lm is not a good fit
model0 <- spm(log_new_in~Days, spar.method="ML")
summary(model0)

# Penalized regression splines with truncated polynomial basis of degree 1
model1 <- spm(log_new_in~ f(Days, basis="trunc.poly", degree=1))
summary(model1)
model1$fit$coef
model1$info$pen$knots

# Penalized regression splines with truncated polynomial basis of degree 2
model2 <- spm(log_new_in~ f(Days, basis="trunc.poly", degree=2))
summary(model2)
model2$fit$coef
model2$info$pen$knots

# Penalized regression splines with radial basis of degree 3
model3 <- spm(log_new_in~ f(Days))
summary(model3)
model3$fit$coef
model3$info$pen$knots

####### Part A.1 #######
par(mfrow=c(1,1))
plot(model1, se=F, ylim=c(min(log_new_in), max(log_new_in)), main= "Models estimators", ylab =
"Number of new intakes",  xlab = "Days")
```

```
lines(model2, se=F, col="blue")
lines(model3, se=F, col="green")
legend("topright", legend=c("model 1", "model 2", "model 3"), col=c("black", "blue", "green"), lty =
1)
points(Days,log_new_in)

####### Part A.2 #######

predict1 <- predict(model1, newdata = mydataA, se=TRUE)
predict2 <- predict(model2, newdata = mydataA, se=TRUE)
predict3 <- predict(model3, newdata = mydataA, se=TRUE)

#Pointwise confidence interval using a t-distribution since the population variance is unknown
n=length(Days)
par(mfrow=c(1,1))
plot(model1, se=F, ylim=c(min(log_new_in), max(log_new_in)), main= "Model 1 estimator and con-
fidence interval", ylab = "Number of new intakes",  xlab = "Days")
lines(unlist(Days),predict1$fit+qt(0.99,df=n-1)*predict1$se,col="blue", lty=2)
lines(unlist(Days),predict1$fit-qt(0.99,df=n-1)*predict1$se,col="blue", lty=2)
points(Days,log_new_in)

par(mfrow=c(1,1))
plot(model2, se=F, ylim=c(min(log_new_in), max(log_new_in)), main= "Model 2 estimator and con-
fidence interval", ylab = "Number of new intakes",  xlab = "Days")
lines(unlist(Days),predict2$fit+qt(0.99,df=n-1)*predict2$se,col="blue", lty=2)
lines(unlist(Days),predict2$fit-qt(0.99,df=n-1)*predict2$se,col="blue", lty=2)
points(Days,log_new_in)

par(mfrow=c(1,1))
plot(model3, se=F, ylim=c(min(log_new_in), max(log_new_in)), main= "Model 3 estimator and con-
fidence interval", ylab = "Number of new intakes",  xlab = "Days")
lines(unlist(Days),predict3$fit+qt(0.99,df=n-1)*predict3$se,col="blue", lty=2)
lines(unlist(Days),predict3$fit-qt(0.99,df=n-1)*predict3$se,col="blue", lty=2)
points(Days,log_new_in)

detach(mydataA)
```

Appendix B

```
# ------------ Question B ------------------
#install.packages("glmnet")

library(glmnet)
library(MASS)

DataQB = read.table("BankDefaultData.txt",header=T)
studentnumber = 773435
set.seed(studentnumber)
rownumbers = sample(1:6436,size=1000)
mydataB = DataQB[rownumbers,]

attach(mydataB)

####### Part B.1 #######
mydataB$Term <- as.factor(Term)

plot(Income, Default)
plot(Employment, Default)
plot(Loan, Default)

# Full model 1
glm.full=glm(Default~.,data=mydataB, family=binomial)
summary(glm.full)

# Full model 2
glm.full2=glm(Default~.^2,data=mydataB, family=binomial)
summary(glm.full2)

# StepAIC Model 1
bm_aic1<- stepAIC(glm.full, k=2, direction='both', scope=list(upper=~.^2, lower=~1))
summary(bm_aic1)

# StepAIC Model 2
bm_aic2<- stepAIC(glm.full, k=2, direction='backward', scope=list(upper=~., lower=~1))
summary(bm_aic2)

# StepAIC Model 3
bm_aic3<- stepAIC(glm.full2, k=2, direction='backward', scope=list(upper=~.^2, lower=~1))
summary(bm_aic3)

all_models <- list(glm.full, glm.full2, bm_aic1, bm_aic2, bm_aic3)

tab = matrix(0, ncol=1, nrow=length(all_models))
for( i in 1:length(all_models)){
  tab[i,] = AIC(all_models[[i]], k = 2)
}

tab = cbind(round(tab, digits=2), c("glm.full","glm.full2", "Model 1", "Model 2", "Model 3"))
```

```
colnames(tab) = c("AIC", "model")
order_models <- tab[order(tab[,1]),]
order_models


####### Part B.2 #######
source("PostAICupdate.R")

# All subsets of glm.full including the intercept
# Postaic intervals
Postaic0.99 <- PostAIC(y=Default, mydataB[,-which(names(mydataB) == "Default")], model.set =
"partsubsets", quant=0.99, family = binomial, common=c(), intercept = T, linearcomb = F )
Postaic0.99$`PostAIC intervals`

# Naive intervals (bm_aic2 is the same model as the one obtained by PostAIC, if it wasn't the case
should fit the glm model with the corresponding features and levels)
summary(bm_aic2)
summary <- cbind(bm_aic2$coefficients, confint(bm_aic2),confint(bm_aic2)[,2]-con-
fint(bm_aic2)[,1], Postaic0.99$'PostAIC intervals', Postaic0.99$'PostAIC intervals'[,2]-
Postaic0.99$'PostAIC intervals'[,1])

colnames(summary) <- c("Estimate", "Naive 0.5%", "Naive 99.5%", "Naive CI width", "PostAIC
0.5%", "PostAIC 99.5%", "PostAIC CI width")
summary

detach(mydataB)
```

Appendix C

```
# ------------ Question C ------------------
library(MASS)
library(glmnet)

DataQC = read.table("bikestations.txt")
digitsum = function(x) sum(floor(x/10^(0:(nchar(x)-1)))%%10)
studentnumber=773435
mysum = digitsum(studentnumber)
Y = DataQC[,mysum]
X = DataQC[,-mysum]
hist(Y)

####### Part C.1 #######
x<- as.matrix(X)
fit_glmnet <- glmnet(x = x, y = Y, family= "poisson")

#### (a) Ridge regression
alpha1 <- 0
cv_fit1 <- cv.glmnet(x, Y, alpha =alpha1, family= "poisson", nfold=7)
plot(cv_fit1)
cv_fit1$lambda.min #Lambda = 12.21
coef(fit_glmnet, s=cv_fit1$lambda.min)

#### (b) Lasso estimation
alpha2 <- 1
cv_fit2 <- cv.glmnet(x, Y, alpha =alpha2, family= "poisson", nfold=7)
plot(cv_fit2)
cv_fit2$lambda.min # Lambda = 0.4187
coef(fit_glmnet, s=cv_fit2$lambda.min)

#### (c) Elastic net with alpha = 0.5
alpha3 <- 0.5
cv_fit3 <- cv.glmnet(x, Y, alpha =alpha3, family= "poisson", nfold=7)
plot(cv_fit3)
cv_fit3$lambda.min # Lambda = 0.7284
coef(fit_glmnet, s=cv_fit3$lambda.min)

#### (d) Elastic net with alpha = 0.8
alpha4 <- 0.8
cv_fit4 <- cv.glmnet(x, Y, alpha =alpha4, family= "poisson", nfold=7)
plot(cv_fit4)
cv_fit4$lambda.min # Lambda = 0.5484
coef(fit_glmnet, s=cv_fit4$lambda.min)

coef_summary <- cbind(coef(fit_glmnet, s=cv_fit1$lambda.min), coef(fit_glmnet,
s=cv_fit2$lambda.min), coef(fit_glmnet, s=cv_fit3$lambda.min), coef(fit_glmnet,
s=cv_fit4$lambda.min))
colnames(coef_summary) <- c("Ridge (alpha=0)", "Lasso (alpha=1)", "Elastic net (alpha=0.5)",
"Elastic net (alpha=0.8)")
coef_summary
```