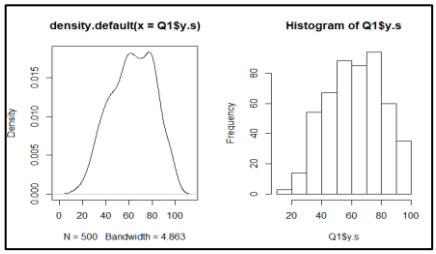
1. QUESTION 1 – BEST DISTRIBUTION

To start this study, we fit different GLM models and compare their AIC values to see which distributions best fit the ***Score_exam*** variable. The density and the histogram are used as guidelines.



Gaussian Distribution

- $E[Y_i|X_i] = \xi_i = X_i^T\beta + \epsilon_i = \beta_0 + \beta_1 x_1 + \beta_2 x_{2,\ldots,} + \beta_{11} x_{11}$
- $Distribution$: $\frac{1}{\sqrt{2\pi\sigma^2}} EXP\{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\}$
- $Distribution\ as\ Exponential\ Family$: $EXP\left\{\frac{(y_i - \frac{\mu_i}{2})^2}{\sigma^2} - \frac{1}{2}(\frac{y_i}{\sigma^2} - log(2\pi\sigma^2))\right\}$
- Parameters: $\theta = \mu$; $b(\theta) = \frac{1}{2}\mu^2$; $a(\phi) = \sigma^2$; $c(y;\phi) = -\frac{x}{2\sigma^2} - \frac{1}{2}\ log\ (2\pi\sigma^2)$
- ***The Link function*** is the Identity Link: $g(u) = \theta = E[\mu] = \mu$; $\mu = \theta$.

Log-Normal Distribution

- $E[\log(Y_i)|X_i] = \xi_i = X_i^T\beta + \epsilon_i = \beta_0 + \beta_1 x_1 + \beta_2 x_{2,\ldots,} + \beta_{11} x_{11}$
- $Distribution$: $\frac{1}{X\sqrt{2\pi\sigma^2}} EXP\left\{-\frac{1}{2\sigma^2}(log\ (x) - \mu_i)^2\right\}$
- $Dist.Exponential\ Family$: $\frac{1}{X\sqrt{2\pi\sigma^2}} EXP\left\{\frac{\mu_i}{\sigma^2} log\ (x) - \frac{\mu_i}{\sigma^2} log\ (x)^2 - \log(\sigma) - \frac{\mu_i}{2\sigma^2}\right\}$
- ***The Link function*** is the Identity Link: $E[log(y)] = \mu = X_i^T\beta$.

Gamma Distribution

- $E[Y_i|X_i] = \xi_i = \alpha/\beta$
- $Distribution$: $f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta y}\ for\ x \geq 0, with\ \alpha > 0\ and\ \beta > 0$
- $Dist.Exponential\ Family$: $EXP\left\{\frac{\frac{\beta}{\alpha}x - \log[\beta]}{-\frac{1}{\alpha}} + (\alpha - 1)\ log(x) - \log[\Gamma(\alpha)])\right\} =$
  $EXP\left\{\frac{\theta x - \log(\theta)}{-\Phi} + \frac{log(\Phi)}{\Phi} + (\frac{1}{\Phi} - 1)log(x) - log\ [\Gamma(\frac{1}{\Phi})])\right\}$
- Parameters: $\theta = \frac{\beta}{\alpha}$; $\beta = \theta\alpha = \frac{\theta}{\Phi}$; $b(\theta) = \log(\theta)$; $a(\phi) = -\phi$; $\phi = \frac{1}{\alpha}$;
  $c(y;\phi) = \frac{log(\Phi)}{\Phi} + (\frac{1}{\Phi} - 1)log(x) - log\ [\Gamma(\frac{1}{\Phi})])$
- ***The Link function*** is the *inverse* Link: $E[u] = \frac{1}{\theta} = \frac{\alpha}{\beta}$; $= \theta = \mu^{-1}$.

1.A. AIC Values

The AIC values indicate that the Normal model is the best as it shows the minimum value. Hence, the models for the following questions use a ***Normal*** distribution. Five different distributions were tried but only the best 3 were explained above.

Normal	Log_Normal	Inv_Normal	Gamma	Poisson
4025.56	4118.163	4217.447	4105.438	4465.539

1.B. Representative code

```
Data<- read.csv("H:/Dataset/OnlineCourse2021.txt",sep="",header=T)
studentnumber = 731529
set.seed(731529)
rownumbers = sample(1:3042,size=500);mydata = data[rownumbers,]
names(mydata) = c('y.p','y.s', 'x1','x2','x3','x4','x5','x6','x7',
                  'x8','x9', 'x10', 'x11')
Q1 = mydata[,-1]

## Density for response variable
plot(density(Q1$y.s));hist(Q1$y.s)
# Possible models
fit.norm <-glm(y.s~.,data = Q1,family=gaussian(link = "identity"))
fit.inv.nor <-glm(y.s~.,data = Q1,family= inverse.gaussian(link =
                                  "1/mu^2"))
fit.Gam <-glm(y.s~.,data = Q1, family = Gamma(link = "inverse"))
fit.poiss <-glm(y.s~.,data=Q1, family = poisson(link = "log"))
fit.log.nor <-glm(log(y.s)~.,data=Q1,family = gaussian(link =
                                "identity"))
## Log Normal AIC Values
log.n = gamlss(formula = formula(y.s~.),sigma.formula = ~1,
                nu.formula = ~1, data=Q1,     tau.formula = ~1,
                family = LOGNO())
#AIC's
log.n$aic ; fit.norm$aic ; fit.log.Gam$aic ; fit.inv.nor$aic ;
fit.poiss$aic
```

2. QUESTION 2 - BEST SEMI PARAMETRIC MODELS

2.A. The following is the list with the three best semi-parametric models out of 15 models that were tried. Please, remark that X1 (4 levels), X2 (2 levels), X3 (5 levels), X8 (2 levels) are categorical variables which require "i-1" dummy variables where "i" is the number of levels present in each variable. Moreover, four different variables were treated as non-parametric functions in the different models as follows: X5 (*model semi1*), X9 (*model semi2*), X10 (*model semi3*), X4 (*model semi4*), and all of them at the same time in *model sem5*.

I) **Semi5.3 {$f(5), f(9), f(10), f(11)$; radial, P= 3}**: $Y_i = \beta_0 + \beta_1 X_{1i} + X_{2i} + X_{3i} + \sum_{k=1}^{k_5} b_1 (X_{5i} - k_{5k})^3{}_+ + X_{6i} + X_{8i} + \sum_{k=1}^{k_9} b_2 (X_{9i} - k_{9k})^3{}_+ + \sum_{k=1}^{k_{10}} b_3 (X_{10i} - k_{10k})^3{}_+ + \sum_{k=1}^{k_{11}} b_3 (X_{11i} - k_{11k})^3{}_+ + \varepsilon_i$

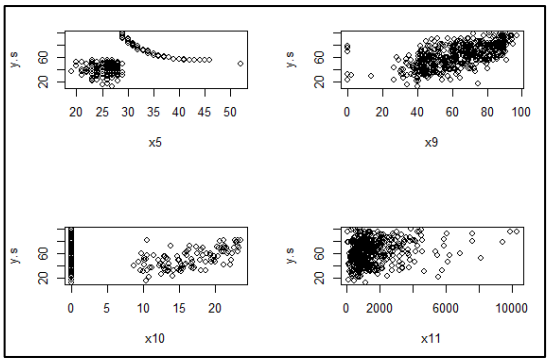
II) **Semi1.3 {$f(5)$; radial, P= 3}**: $Y_i = \beta_0 + \beta_1 X_{1i} + X_{2i} + X_{3i} + \sum_{k=1}^{k_5} b_1 (X_{5i} - k_{5k})^3{}_+ + X_{6i} + X_{8i} + X_{9i} + X_{10} + X_{11i} + \varepsilon_i$

III) **Semi1.1 {$f(5)$; Truncated, P= 1}**: $Y_i = \beta_0 + \beta_1 X_{1i} + X_{2i} + X_{3i} + \sum_{k=1}^{k_5} b_1 (X_{5i} - k_{5k})_+ + X_{6i} + X_{8i} + X_{9i} + X_{10} + X_{11i} + \varepsilon_i$

Where $\epsilon_i \sim N(0, \sigma^2)$

To select the models, first, an AIC variable selection process was performed in the parametric additive model that identified which variables significantly contributed to the model, and it suggested removing X7. Also, the correlation matrix showed no significant correlation amongst the variables, and thereby, no variable was dismissed due to correlation.

	Var	Df	Deviance	AIC
	<none>	-	85498	4023.8
-	x4	1	85844	4023.8
-	x10	1	86049	4025
+	x7	1	85464	4025.6
-	x6	1	86222	4026
-	x11	1	86240	4026.1
-	x8	1	86975	4030.3
-	x2	1	87443	4033
-	x1	3	89196	4038.9
-	x3	4	89652	4039.5
-	x5	1	88813	4040.8
-	x9	1	122600	4202



We constructed fifteen models using the four variables that showed a possible non-parametric trend in the plots against the response variable. Each model consisted of one of the possible non-parametric variables using a truncated polynomial with degrees p = 1, P =2, and a radial basis function with p = 3. Also, a group of models using all variables and the different basis functions were performed as well. The second digit 1,2 and 3 that each model has (*exmaple1.**1***) represent the "p" degree that each non-parametric function is using.

	AIC Values
semi1.1	**3762.965**
semi1.2	3810.861
semi1.3	**3755.896**

	AIC Values
semi2.1	4033.964
semi2.2	4019.484
semi2.3	4020.928

	AIC Values
semi3.1	4046.989
semi3.2	4042.697
semi3.3	4049.93

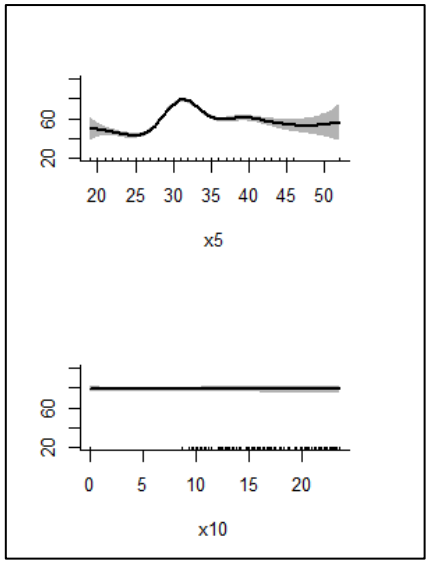
	AIC Values
semi4.1	4049.631
semi4.2	4053.396
semi4.3	4049.631

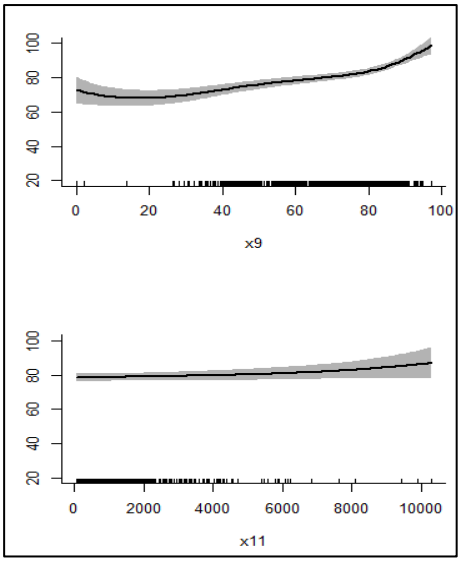
	AIC Values
semi5.1	3931.482
semi5.2	3926.509
semi5.3	**3750.821**

After the AIC analysis the model ($\boldsymbol{Semi5.3}$) which contains four non-parametric function is the best model with AIC =**-175.17169**. This model uses a Gaussian distribution, with a radial polynomial basis function of degree three. The spline was selected after comparing the AIC values from the same model using a linear, quadratic and a radial spline (p=3). Regarding the smoothing parameter $\boldsymbol{\lambda}$, the coefficients $\boldsymbol{b_k}$ are treated as random variables and we assume that the parameters are independent and $\boldsymbol{b_k} \sim \boldsymbol{N(0, \sigma_u^2)}$ $\boldsymbol{for\ all\ K = 1, .., k}$. Therefore, the smoothing parameter is calculates using the variance components of the mixed model representation ($\boldsymbol{\lambda = \sigma_\varepsilon^2 / \sigma_b^2}$), and the parameters are estimated via MLE to make comparisons through the AIC values.

2.B. The plot illustrates that a parametric function would not describe the behavior of the variables x5, and x9 as good as the truncated polynomial function does, because this non-parametric function adapted to different peaks and curvatures that a linear model would not be able to fit, and thereby, those functions illustrated the data in a better way. Furthermore, the functions corresponding to variables ***X10*** and ***X11*** also seem to represent the data in a better way as it depicts almost a horizontal or flat trend that a parametric function or a linear model would not describe accurately. The summary of these estimations is shown below.

```
Summary for non-linear components:
          df     spar knots
f(x5)  6.636     3.61     6
f(x9)  3.677    73.69    34
f(x10) 1.000 31030.00    25
f(x11) 1.433 47530.00    34
```





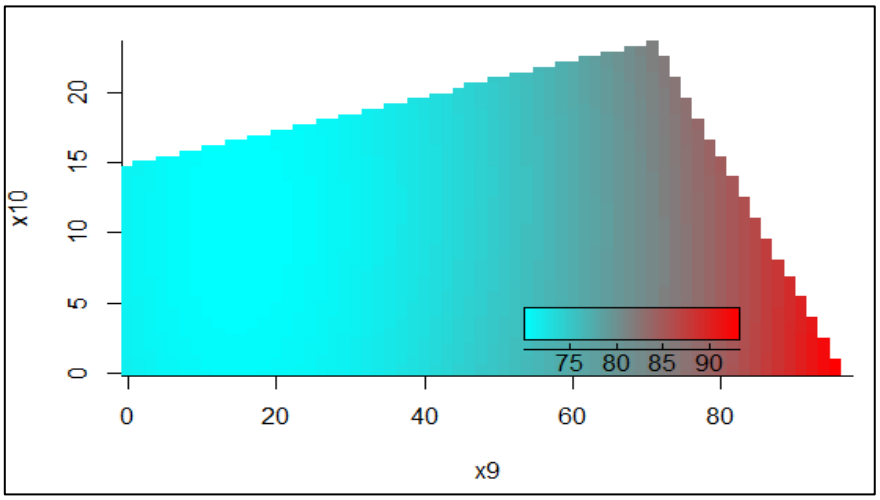
2.C. Representative Code

```
## Data for Q2
Q2 = Q1; str(Q2)
## Variable correlation
cor.var= select(Q2, y.s,x4, x5, x6, x7,x9,x10,x11)#continuous var
cor.var; str(cor.var);cor(cor.var) # No important correlations
## AIC Variable selesction
fit.int <-glm(y.s ~., data = Q2, family = gaussian())
stepboth = stepAIC(fit.int, k =2, direction = "both", scope =
                   list(upper = ~. , lower= ~1))
summary(stepboth)
## Selected Model
fit.select <-glm(y.s ~ x1 + x2 + x3 + x4 + x5 + x6 + x8 + x9 +
                       x10 + x11,data = Q2, family = gaussian())
summary(fit.select)
## Plots
par(mfrow=c(2,2))#Only non-parametric variables
plot( x5, y.s);plot( x9, y.s);plot( x10, y.s);plot( x11, y.s)
## Semiparametric models – Only Best models
semi1.1 = spm( y.s ~ x1 + x2 + x3 + x4 + f(x5,basis='trunc.poly')+
                     x6 + x8 + x9 + x10 + x11, spar.method = "ML")
semi1.3 = spm(y.s ~ x1 + x2 + x3 + x4 + f(x5) + x6 + x8 + x9 +
                    x10 + x11, spar.method = "ML")
```

```
semi5.3 = spm(y.s ~ x1 + x2 + x3 + x4 + f(x5) + x6 + x8 + f(x9) +
                    f(x10) + f(x11), spar.method = "ML")
## AIC Values
AIC.func <- function(x) {
                AIC <- -2*(x$fit$logLik - sum(x$aux$df))
                return(AIC)}

AIC = c( AIC.func(semi1.1), AIC.func(semi1.3), AIC.func(semi5.3))
AIC # Only best models in this report
## Model plot and Summary
summary(semi5.3); y = Q2$y.s); par(mfrow=c(2,2))
plot(semi5.3, ylim=c(20,100)
```

3. QUESTION 3 - BIVARIATE SMOOTH EFFECT

3.A. Bivariate Smooth Effect



3.B. Model $\{f(5), f(9,10), f(11)$; Radial, p=3$\}$: $Y_i = \beta_0 + \beta_1 X_{1i} + X_{2i} + X_{3i} + \sum_{j=1}^{d} g_j(X_{5i}) + X_{6i} + X_{8i} + \sum_{k,l=1}^{c} g_{k,l}(X_{9i}, X_{10i}) + \sum_{j=1}^{d} g_j(X_{11i}) + \varepsilon_i$. Where $\varepsilon_i \sim N(0, \sigma^2)$

The plot shows the behavior of the response variable ***Score_exam*** based on the variables corresponding to ***the weighted score in tutor marked assignments***, and ***the Weighted score in computer marked assignments***. Overall, the two variables depict almost a negative relationship as the Y response produces higher exam scores if X9 increases and X10 decreases. Computer scores close to 0, and tutor scores higher than 70 seem to produce exam results that range between 80 and 90, which could be considered as a high score. Therefore, this indicates that ***the Weighted score in computer-marked assignments*** is usually mistaken as students with low scores show high results when the tutor checks the assignments.

3.C. In this part we pretend to identify whether an interaction between the variable X9 and X10 is significant. For this end we use the order selection test which is a non-parametric test that fits a

model against a range of alternative models, that for this exercise correspond to a range of polynomials up to degree four which are constructed between the variables X9 and X10.

We have a model of the form $Y_i = \mu(.) + \varepsilon_i$ Where $\varepsilon_i \sim N(0, \sigma^2)$, so the null hypothesis is:

H_0: There exist values $(\theta_0, \theta_1, \ldots, \theta_{11}) \in R^2$ such that $\mu_\theta(x) = \theta_0 + \theta_1 x, \ldots, + \theta_{11} x$.
H_α: $\mu(.) \notin \{\mu_\theta(.): \theta = (\theta_0, \theta_1, \ldots, \theta_{11}) R^2\}$.

Model under alternative hypothesis: The LRT_m indicates that there are 12 different models constructed with the polynomials between $x9, x10$.
$\mu(x) = \mu(x) + \sum^4_{j=1} a_j \varphi_j(x9, x10)$.

In the non-parametric hypothesis for this model, the basis functions $\varphi_j(.)$ corresponds to an orthogonal polynomial, and m the degree of the polynomial which in this case we selected it as 4. The null model M_0 is the model without the interactions between x9 and x10. Also, the test statistic is constructed as shown below:

- $T_{n,}os = max_{m \geq 4} \frac{LRT(M_m, M_0}{m=4}$

After calculating the statistic, we obtained ***$T_{n,}os = 39.63$***, which p-value equals to ***3.05e-10***, and thereby there is significant evidence to reject H_0. Hence, an interaction between x9 and x10 significantly influences the response variable and the estimators for $\beta_9 \; and \; \beta_{10}$. In fact, as mentioned in the graph above there seemed to be a relationship between these variables, and when we calculate the AIC values for the parametric additive model and the model with interactions using the ***glm*** function, we get lower scores for the model with the interaction.

- `AIC.Int.null = 4026.838`
- `AIC.Int.mod =4022.347`

3.D. Representative Code

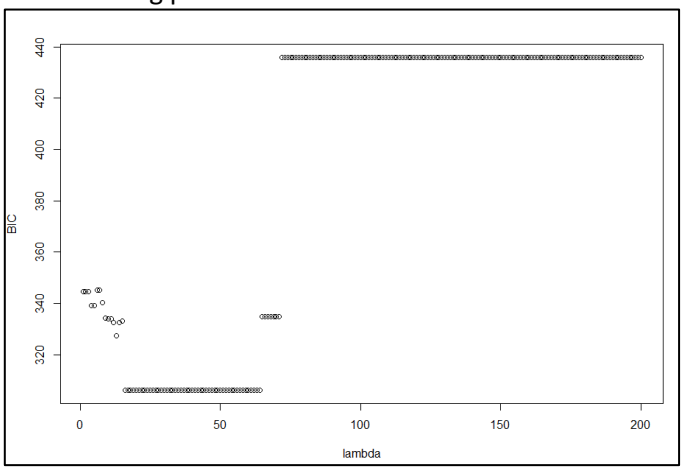
```
## sections A and B
bivar2 = spm(y.s ~ x1 + x2 + x3 + x4 + f(x5) + x6 +x8 +f(x9,x10) +
             f(x11), spar.method = "ML")
plot(bivar2)
## Section C - Parametric model
null.fit <- glm(y~x1+x2+x3+x4+x5+x6+x8+x9o+x10o+x11)
int.fit  <- glm(y~x1+x2+x3+x4+x5+x6+x8+x9o+x10o+x11+x10*x11)
null.fit$aic;int.fit$aic
##non-parametric test
y = y.s; new = poly(cbind(x9,x10),4; colnames(new)
x9o = new[,1]; x10o = new[,5]
model <- null <- lm(y~x1+x2+x3+x4+x5+x6+x8+x9o+x10o+x11)

new <- new[, -c(1,5)]; k <- dim(new)[2]; LogL <- NULL; LRT <- NULL
df <- NULL; m <- NULL; LRT_m <- NULL

for ( i in 1:k){
  model <- update(model,as.formula(paste0(".~. + new[,",i,"]")))
  LogL[i] <- logLik(model)[1]
```

```
  m[i] <- length(model$coefficients)-length(null$coefficients)
  LRT[i] <- 2*(LogL[i]-logLik(null))
  LRT_m[i] <- LRT[i]/m[i]
  }
LRT_m; TnOS <- max(LRT_m); order<- which.max(LRT_m)
##pvalue
m2 <- 1000; cn <- TnOS
pvalue = 1-exp(-sum((1-pchisq((1:m2)*cn, 1:m2))/(1:m2)))
pvalue
```

4 . QUESTION 4

4.A. To find a good lasso fit for this data we considered a range from 0 to 2000 for the regularization parameter $\boldsymbol{\lambda}$, and 200 groups were analyzed as $\boldsymbol{\lambda}$ *was explored by groups of 10.* Therefore, we evaluated the BIC values for 201 models and the results indicated that the model with the best BIC values equaled **306.215** and used a $\boldsymbol{\lambda}$ equal to 50.

The following plot illustrates the aforementioned results.



4.B. First, the penalized model shows to be sparse as most of the parameter estimates were zero, except for those corresponding to the variables X9, X5, and the one for the intercept, as we used the lasso estimator and the bias of the constrained estimators increased. Besides, these results also prove that the estimator approximates to the Least Squares Estimator when the penalty approaches to zero. Therefore, we can say that the lasso estimator is good technique or performed well in this case as most of the coefficients are truly zero.

Variables	No_Penalized	Penalized
(Intercept)	-1.25E+01	-9.25704284
as.factor(x1)Sem2	3.33E-01	0
as.factor(x1)Sem3	1.76E+00	0
as.factor(x1)Sem4	-2.26E-01	0
as.factor(x2)M	5.79E-01	0
as.factor(x3)HE Qualification	7.43E-01	0
as.factor(x3)Lower Than A Level	-1.43E-01	0
as.factor(x3)No Formal quals	-2.21E+02	0
as.factor(x3)Post Graduate Qualification	2.26E+02	0
x4	1.27E-01	0
x5	3.02E-01	0.27791983
x6	-2.86E-01	0
x7	5.64E-03	0
as.factor(x8)Y	-1.35E-01	0
x9	6.19E-02	0.05607789
x10	1.06E-01	0
x11	-7.20E-05	0

4.C. EXPECTATION FOR A STUDENT PASSING THE COURSE

$E[Y_i = 1] = E[logit^{-1}(-9.25 + 0.27X_5 + 0.05X_9 + U_i)]$ *where* $U_i \sim N(0, 0.64)$

Std(U_i) = 0.80

4.D. REPRESENTATIVE CODE

```
Q4<- mydata[,-2];names(Q4);attach(Q4);hist(y.p)
## 4.A penalized L1 estimation
## Starting values
lambda <- seq(0,2000,by=10);dist = binomial(link = logit)
BIC<-rep(Inf,length(lambda))
PQL<-glmmPQL(y.p~1,random = ~1|x1,family=dist,data=Q4)

for(j in 1:length(lambda))
{print(paste("Iteration ", j,sep=""))
  fits <- try(glmmLasso(y.p~ as.factor(x1)+ as.factor(x2)+
        as.factor(x3)+ x4+ x5+ x6+ x7+ as.factor(x8)+x9+x10+x11,rnd
        = list(x1=~1),family = dist, data = Q4,  lambda=lambda[j],
        switch.NR=T,final.re=TRUE),silent=TRUE)
  if(class(fits)!="try-error")
  {
   BIC[j]<-fits$bic
  }
 }
BIC; opt<-which.min(BIC);opt
lasso.fit <- glmmLasso(y.p~ as.factor(x1)+ as.factor(x2)+
             as.factor(x3)+x4+x5+x6+x7+as.factor(x8)+x9+x10+x11,
             rnd = list(x1=~1),family = dist, data = Q4,
             lambda=lambda[opt],switch.NR=F,final.re=TRUE)
summary(lasso.fit); plot(lambda,BIC);lambda[6];BIC[6]

## 4.B Table
lasso.fit2 <- glmmLasso(y.p~ as.factor(x1)+ as.factor(x2)+
              as.factor(x3)+x4+x5+x6+x7+as.factor(x8)+x9+x10+x11,
              rnd = list(x1=~1),family = dist, data = Q4,
              lambda=0,switch.NR=F,final.re=TRUE)

summary(lasso.fit)
Penalized = lasso.fit$coefficients
No_Penalized = lasso.fit2$coefficients

table = data.frame(No_Penalized,Penalized)
table
```