Task 1: The EUxPancreas Cohort Study

The EUxPancreas study investigates regional and demographic patterns in pancreatic cancer incidence across Europe. The dataset includes demographic, lifestyle, and regional predictors for new cancer cases, with the goal of identifying key risk factors.

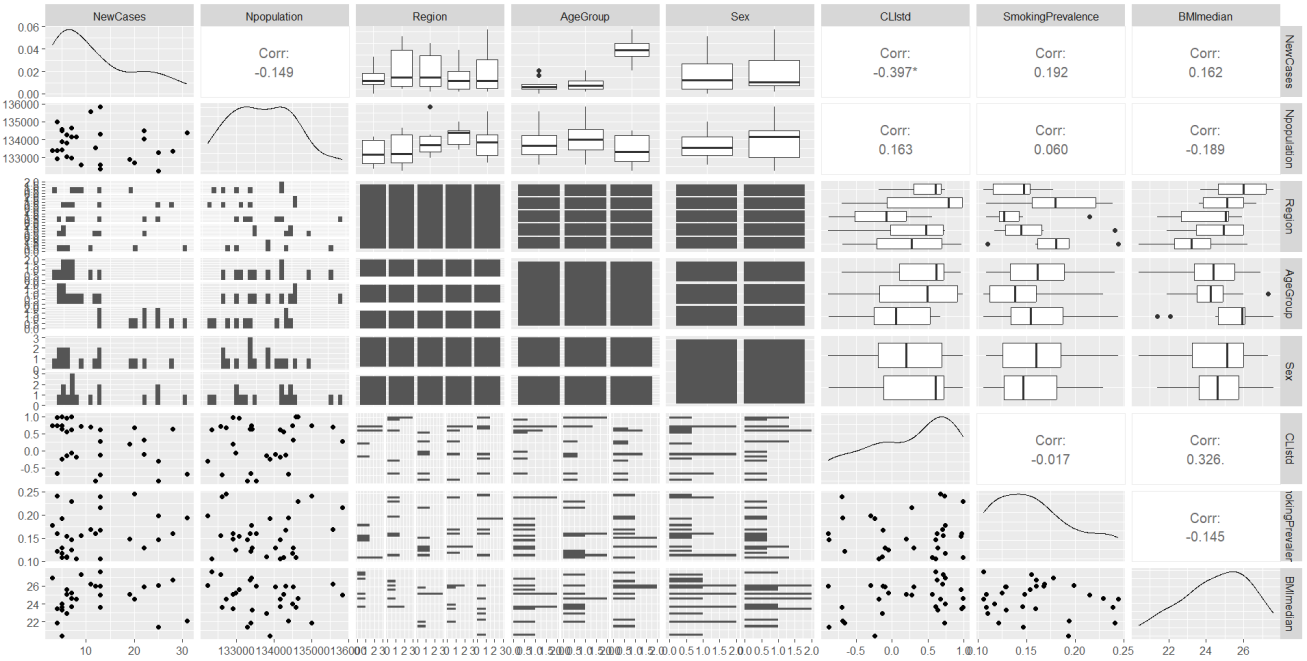
Explore the dataset:

```
> head(data)
  X NewCases Npopulation  Region AgeGroup  Sex CLIstd SmokingPrevalence BMImedian
1 1        7      134167 Region1    20-39 Male  0.608           0.14618      23.7
2 2        6      133057 Region2    20-39 Male  0.961           0.15393      25.6
3 3        7      132978 Region3    20-39 Male -0.055           0.12506      25.2
4 4        5      133420 Region4    20-39 Male  0.637           0.12901      23.3
5 5       11      135585 Region5    20-39 Male  0.705           0.16821      26.2
6 6        8      134167 Region1    40-59 Male -0.182           0.10592      24.5
```

```
> summary(data)
       X            NewCases      Npopulation        Region
 Min.   : 1.00   Min.   : 3.00   Min.   :132219   Length:30
 1st Qu.: 8.25   1st Qu.: 5.25   1st Qu.:132998   Class :character
 Median :15.50   Median : 8.50   Median :133660   Mode  :character
 Mean   :15.50   Mean   :11.93   Mean   :133744
 3rd Qu.:22.75   3rd Qu.:17.50   3rd Qu.:134366
 Max.   :30.00   Max.   :31.00   Max.   :135836
   AgeGroup             Sex                CLIstd        SmokingPrevalence
 Length:30          Length:30          Min.   :-0.8910   Min.   :0.1058
 Class :character   Class :character   1st Qu.:-0.1705   1st Qu.:0.1259
 Mode  :character   Mode  :character   Median : 0.4350   Median :0.1545
                                       Mean   : 0.2368   Mean   :0.1596
                                       3rd Qu.: 0.7133   3rd Qu.:0.1891
                                       Max.   : 0.9970   Max.   :0.2447

   BMImedian
 Min.   :20.40
 1st Qu.:23.52
 Median :24.80
 Mean   :24.55
 3rd Qu.:25.98
 Max.   :27.60
```

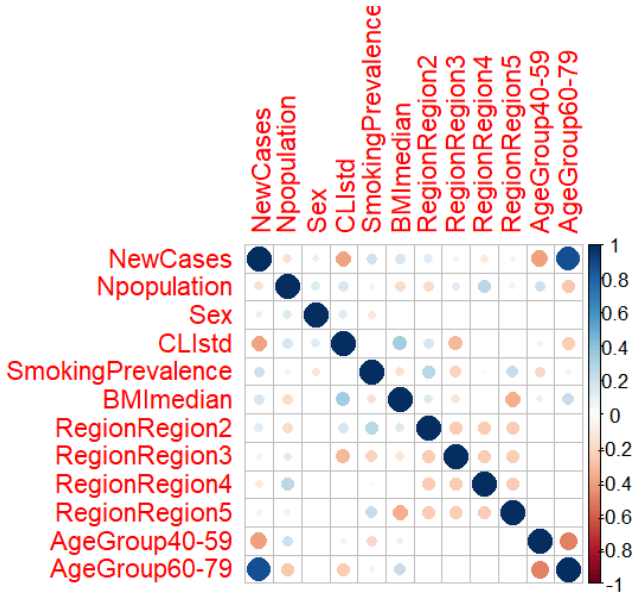
```
> colSums(is.na(data)) #Check for missing data
                X          NewCases       Npopulation            Region
                0                 0                 0                 0
         AgeGroup               Sex           CLIstd SmokingPrevalence
                0                 0                 0                 0
        BMImedian
                0
```

The dataset contained 30 observations across 8 variables with no missing data.



In the data set there are some nonnumerical values. We convert AgeGroup and Region to categorical values and sex to numerical (0 and 1) and plot them in a corrplot for better visibility

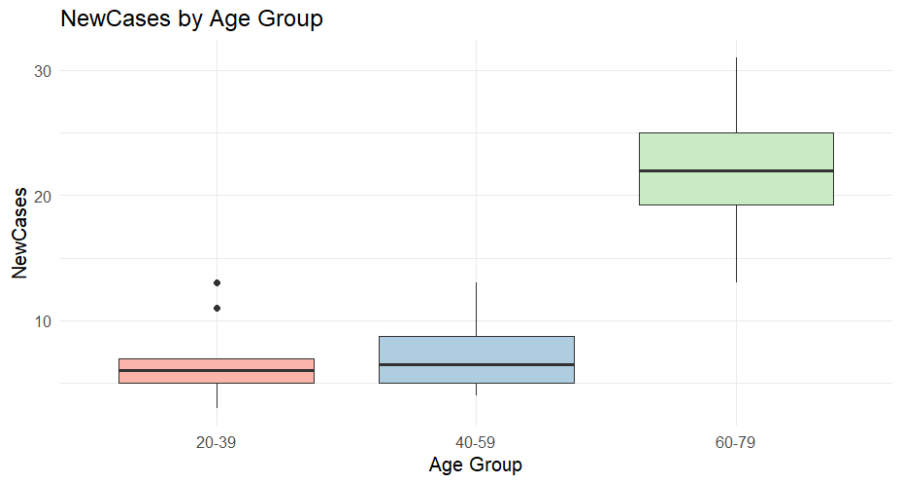
```r
#Ensure the columns are factors
data$Region <- as.factor(data$Region)
data$AgeGroup <- as.factor(data$AgeGroup)


# Convert Sex to binary
data$Sex <- ifelse(data$Sex == "Male", 1, 0)
```



The correlation plot suggests that AgeGroup has a strong positive association with NewCases, while CLIstd (composite lifestyle index) is negatively correlated, suggesting protective effects of healthier lifestyles

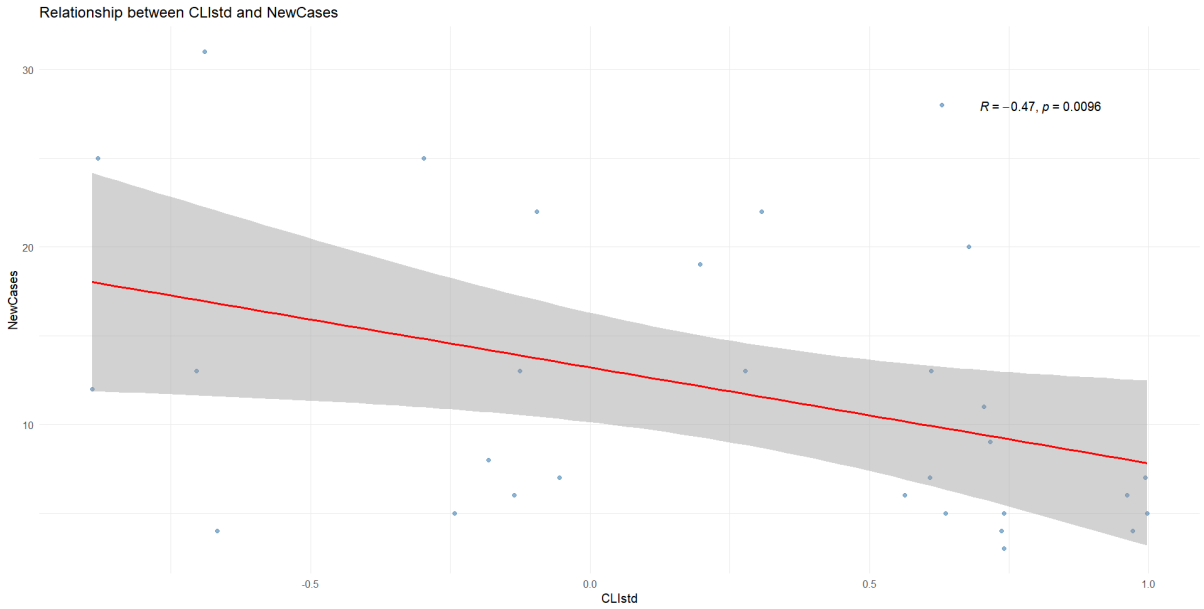
We can therefore plot NewCases by Age Group in a boxplot.

```
> ggplot(data, aes(x = AgeGroup, y = NewCases, fill = AgeGroup)) +
+   geom_boxplot() +
+   theme_minimal() +
+   labs(title = "NewCases by Age Group",
+        x = "Age Group",
+        y = "NewCases") +
+   scale_fill_brewer(palette = "Pastel1") +
+   theme(legend.position = "none")
```



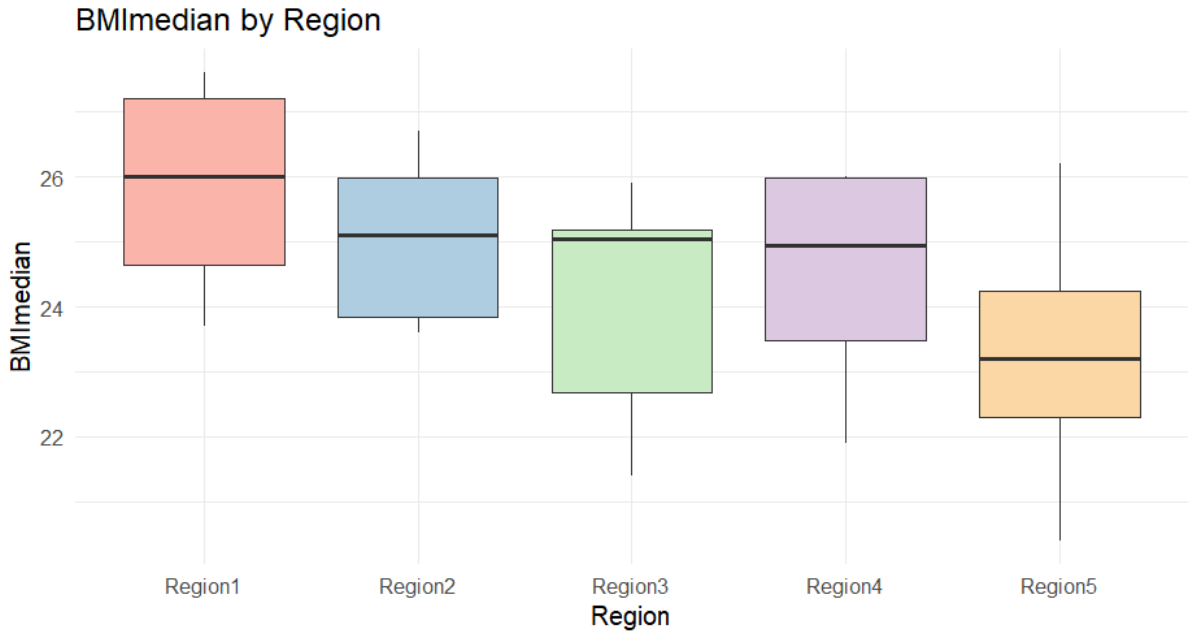
We can clearly see that older people have higher cancer prevalence.

We can also see that the continuous variable CLIstd seems to have a negative correlation with New Cases

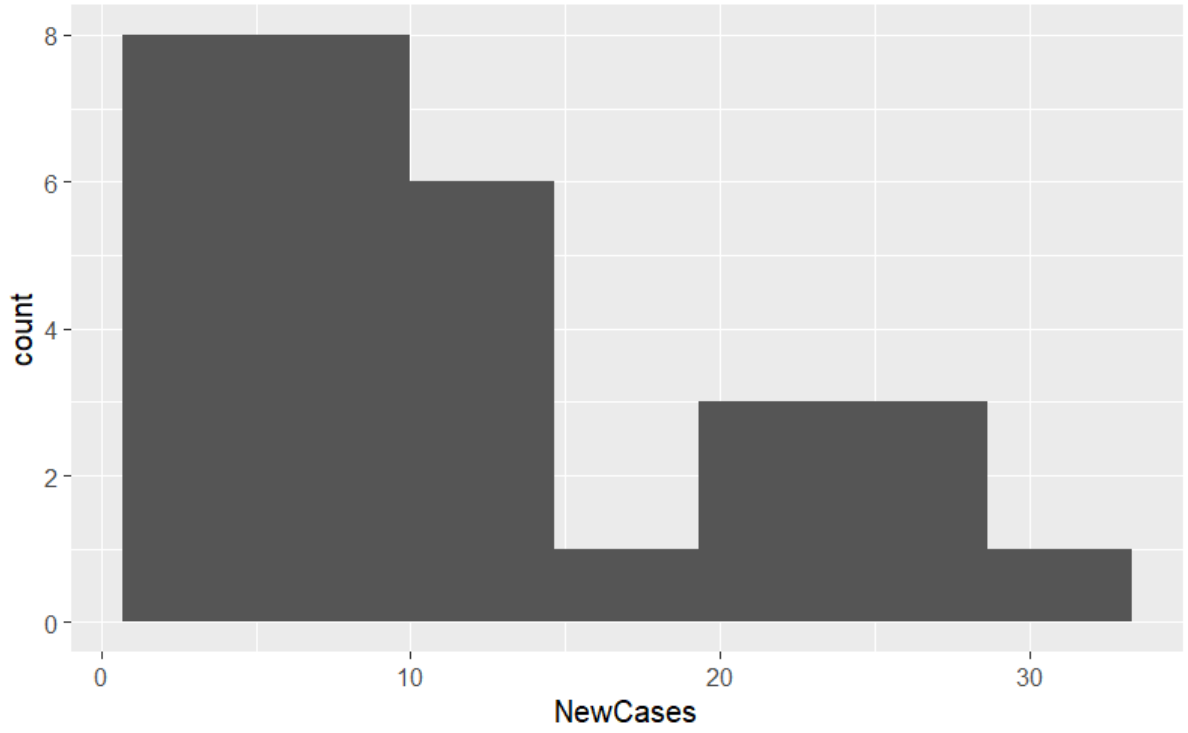


A better lifestyle indicates less cancer prevalence.

We also notice that there is not that strong of a relation between variables. However there is a significant negative intervariable correlation between Region and BMI



NewCases are right-skewed, supporting Poisson modeling.



Exploratory data analysis revealed that the number of new pancreatic cancer cases (NewCases) follows a count distribution, making the Poisson family a suitable modeling choice. Higher age groups showed significantly higher incidence rates, while healthier lifestyle scores (CLIstd) were negatively associated with new cases. Regional differences were observed in median BMI. These findings guided the subsequent model development, focusing on Poisson regression with population offsets and selected interaction terms.

Develop a model to examine the trends in new cases of pancreatic cancer across the recorded population variables:

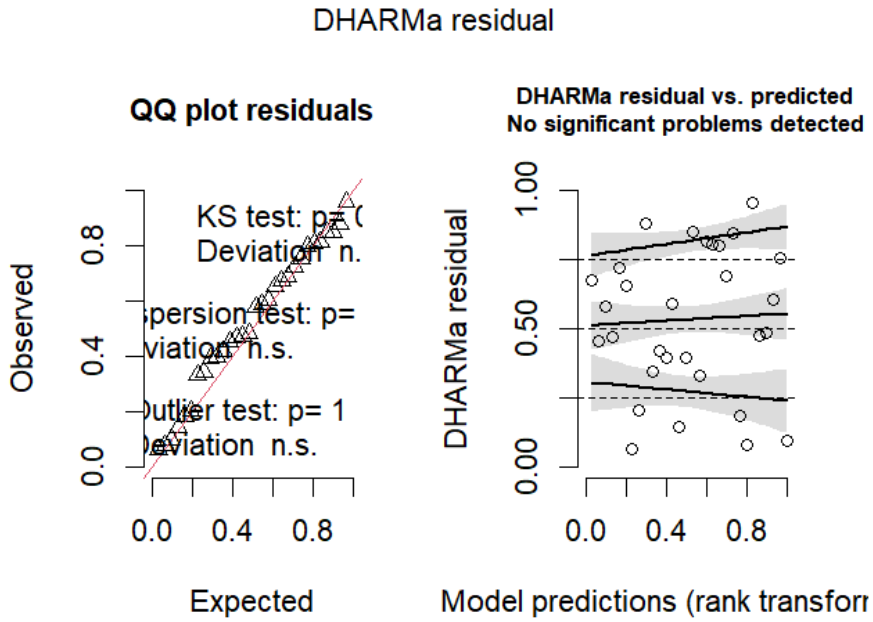
Since the outcome variable (NewCases) represents count data and is approximately Poisson-distributed, we used a Poisson GLM with a log link. The population size was included as an offset to account for varying subgroup sizes. For the first model we included all the predictors.

In Model 1 all predictors were included. AgeGroup and CLIstd were significant ($p < 0.05$), while BMI, region and SmokingPrevalence were not

```
Call:
glm(formula = NewCases ~ Sex + AgeGroup + Region + CLIstd + BMImedian +
    SmokingPrevalence, family = poisson(link = "log"), data = data_glm,
    offset = log(Npopulation))

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -12.24995    1.10368 -11.099  < 2e-16 ***
Sex                 0.12479    0.10854   1.150  0.25028
AgeGroup40-59       0.09439    0.17133   0.551  0.58168
AgeGroup60-79       1.02073    0.15183   6.723 1.78e-11 ***
RegionRegion2       0.21867    0.19230   1.137  0.25549
RegionRegion3       0.18079    0.19382   0.933  0.35092
RegionRegion4      -0.04542    0.18872  -0.241  0.80981
RegionRegion5       0.23155    0.22066   1.049  0.29400
CLIstd             -0.32063    0.11642  -2.754  0.00589 **
BMImedian           0.07719    0.04176   1.849  0.06453 .
SmokingPrevalence   2.33964    1.64550   1.422  0.15507
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 152.871  on 29  degrees of freedom
Residual deviance:  20.459  on 19  degrees of freedom
AIC: 165.89

Number of Fisher Scoring iterations: 4

> BIC(model1)
[1] 181.3069
> rsq(model1, adj=TRUE)
[1] 0.7722274
```

DHARMa residual diagnostics showed no overdispersion and a uniform distribution of simulated residuals, indicating a good model fit.

```
> testDispersion(model1)

        DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated

data:  simulationOutput
dispersion = 0.86085, p-value = 0.728
alternative hypothesis: two.sided
```



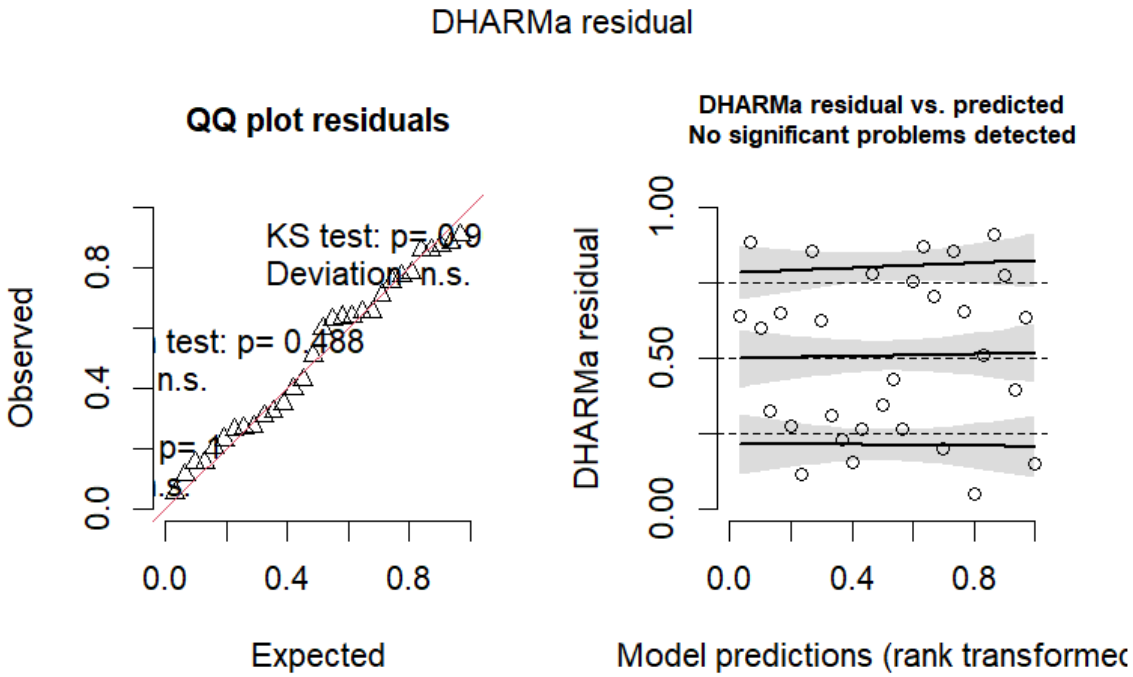
We moved on by trying CLIstd*Region as an interaction term.

In Model 2, only AgeGroup were significant ($p < 0.05$), while BMI, sex, region, CLIstd, SmokingPrevalence and the interaction term were not significant. The performance was clearly worse than Model 1.

```
Call:
glm(formula = NewCases ~ Sex + AgeGroup + CLIstd * Region + BMImedian +
    SmokingPrevalence, family = poisson(link = "log"), data = data_glm,
    offset = log(Npopulation))

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -12.83625    1.94048  -6.615 3.72e-11 ***
Sex                    0.12488    0.11798   1.058    0.290
AgeGroup40-59          0.04086    0.18625   0.219    0.826
AgeGroup60-79          0.96615    0.18734   5.157 2.51e-07 ***
CLIstd                -0.78950    0.49850  -1.584    0.113
RegionRegion2         -0.02977    0.29447  -0.101    0.919
RegionRegion3          0.03496    0.25662   0.136    0.892
RegionRegion4         -0.23240    0.26257  -0.885    0.376
RegionRegion5          0.07642    0.28062   0.272    0.785
BMImedian              0.10476    0.07635   1.372    0.170
SmokingPrevalence      3.12162    1.91066   1.634    0.102
CLIstd:RegionRegion2   0.59176    0.54062   1.095    0.274
CLIstd:RegionRegion3   0.45336    0.49503   0.916    0.360
CLIstd:RegionRegion4   0.49343    0.63513   0.777    0.437
CLIstd:RegionRegion5   0.30808    0.46539   0.662    0.508
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 152.871  on 29  degrees of freedom
Residual deviance:  18.715  on 15  degrees of freedom
AIC: 172.15

Number of Fisher Scoring iterations: 4
```

```
> testDispersion(model2)

        DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated

data:  simulationOutput
dispersion = 0.72836, p-value = 0.488
alternative hypothesis: two.sided
```

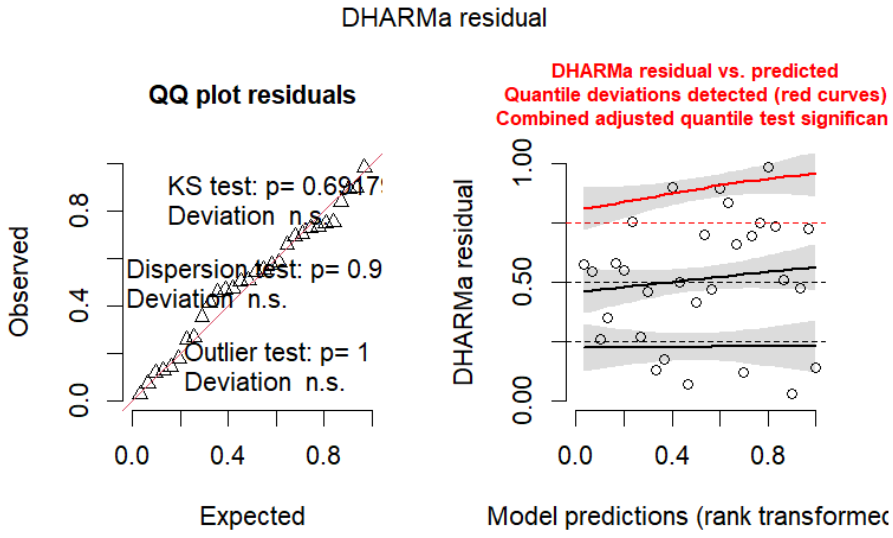
DHARMa residual diagnostics showed no overdispersion and a uniform distribution of simulated residuals, indicating a good model fit.



We try using CLIstd*Sex as an interaction term and we drop region as a predictor.
In Model 3, AgeGroup, BMI and SmokingPrevalence were significant ($p < 0.05$), while, sex, CLIstd, and the interaction term were not significant.

```
Call:
glm(formula = NewCases ~ AgeGroup + CLIstd * Sex + BMImedian +
    SmokingPrevalence, family = poisson(link = "log"), data = data_glm,
    offset = log(Npopulation))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -12.22217    0.96727 -12.636  < 2e-16 ***
AgeGroup40-59      0.10061    0.17204   0.585   0.5587
AgeGroup60-79      0.95984    0.16431   5.842 5.16e-09 ***
CLIstd            -0.20450    0.14733  -1.388   0.1651
Sex                0.17298    0.11164   1.549   0.1213
BMImedian          0.07565    0.03726   2.030   0.0423 *
SmokingPrevalence  3.19854    1.35674   2.358   0.0184 *
CLIstd:Sex        -0.25151    0.22087  -1.139   0.2548
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 152.871  on 29  degrees of freedom
Residual deviance:  22.637  on 22  degrees of freedom
AIC: 162.07

Number of Fisher Scoring iterations: 4
```

```
> testDispersion(model3)

        DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated

data:  simulationOutput
dispersion = 0.95334, p-value = 0.936
alternative hypothesis: two.sided
```



The residuals seem normally distributed but we get some quantile deviation indicating a worse fit.

We continue our model fitting with CLIstd*BMI as an interaction term.
In Model 4, AgeGroup and SmokingPrevalence were significant ($p < 0.05$), while sex, BMI, region, CLIstd, and the interaction term were not significant.

```
Call:
glm(formula = NewCases ~ AgeGroup + CLIstd * BMImedian + Sex + 
    SmokingPrevalence, family = poisson(link = "log"), data = data_glm, 
    offset = log(Npopulation))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -12.23217    0.97542 -12.540  < 2e-16 ***
AgeGroup40-59      0.12090    0.17311   0.698  0.48491
AgeGroup60-79      1.01068    0.15146   6.673 2.51e-11 ***
CLIstd            -1.91563    1.34035  -1.429  0.15295
BMImedian          0.07047    0.03605   1.955  0.05059 .
Sex                0.12376    0.10969   1.128  0.25921
SmokingPrevalence  3.80180    1.40104   2.714  0.00666 **
CLIstd:BMImedian   0.06538    0.05462   1.197  0.23128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 152.871  on 29  degrees of freedom
Residual deviance:  22.497  on 22  degrees of freedom
AIC: 161.93

Number of Fisher Scoring iterations: 4
```

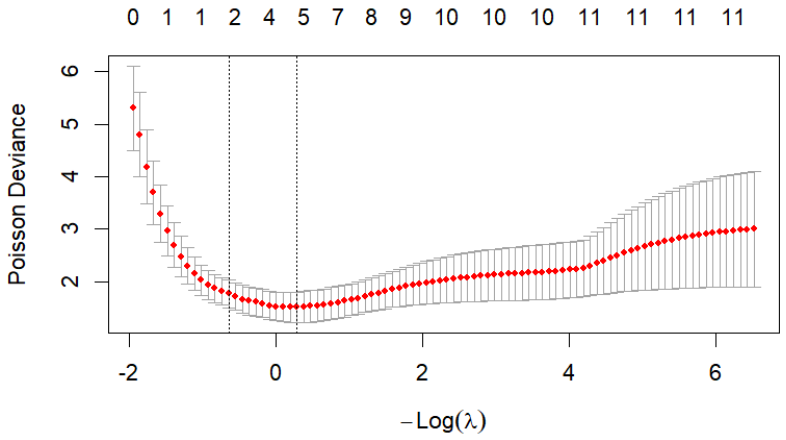
	Model	AIC	BIC	Rsq_adj
1	model1	165.8937	181.3069	0.7722274
2	model2	172.1497	193.1677	0.7580223
3	model3	162.0713	173.2809	0.7805745
4	model4	161.9316	173.1412	0.7893618

AIC evaluates how well your model fits the data while penalizing complexity (i.e. the number of parameters).
BIC - Similar to AIC but penalizes complexity more strongly,
R^2 - The proportion of variation in the response variable that your model explains.

We see that Model 3 and Model 4 have similar performance.

We applied LASSO regularization to identify the most relevant predictors and reduce potential multicollinearity between correlated variables

```
> cv_lasso <- cv.glmnet(X, y,
+                       family = "poisson",
+                       offset = offset_var,
+                       alpha = 1)  # alpha = 1 -> LASSO
> plot(cv_lasso)
> cv_lasso$lambda.min     # lambda that minimizes cross-validation error
[1] 0.7503966
> coef(cv_lasso, s = "lambda.min")
12 x 1 sparse Matrix of class "dgCMatrix"
                     lambda.min
(Intercept)       -9.882868809
AgeGroup40-59      .
AgeGroup60-79      0.967443219
RegionRegion2      0.070845699
RegionRegion3      .
RegionRegion4     -0.004433446
RegionRegion5      .
CLIstd            -0.153208338
BMImedian          .
Sex                .
SmokingPrevalence  0.824450956
CLIstd:BMImedian   .
```

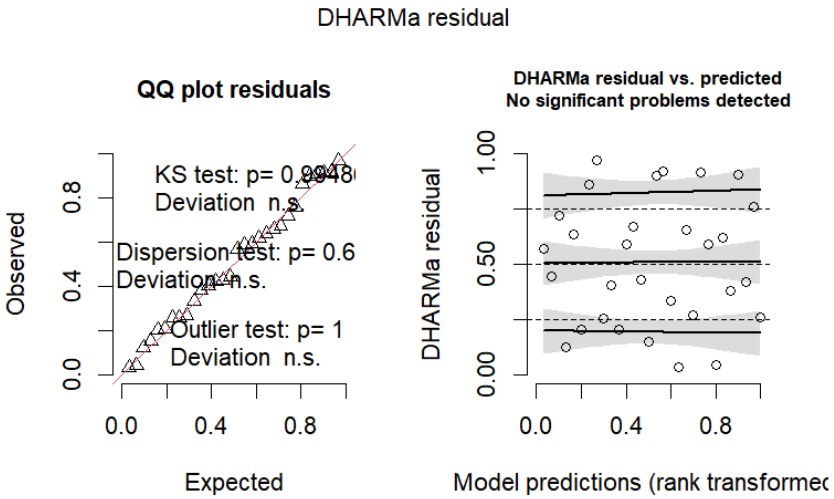


Moving right -> stronger penalty (simpler model, fewer nonzero coefficients).
y-axis -The cross-validated error. It shows how well the model performs on unseen data for each λ.

```
Call:
glm(formula = NewCases ~ AgeGroup + CLIstd + SmokingPrevalence +
    Region, family = poisson(link = "log"), data = data_glm,
    offset = log(Npopulation))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -10.22060    0.29818 -34.277  < 2e-16 ***
AgeGroup40-59      0.11090    0.17258   0.643   0.5205
AgeGroup60-79      1.11941    0.14376   7.787 6.88e-15 ***
CLIstd            -0.22617    0.10083  -2.243   0.0249 *
SmokingPrevalence  1.90042    1.61120   1.180   0.2382
RegionRegion2      0.22945    0.18856   1.217   0.2237
RegionRegion3      0.08436    0.18807   0.449   0.6538
RegionRegion4     -0.07578    0.18747  -0.404   0.6861
RegionRegion5      0.07318    0.19940   0.367   0.7136
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 152.871  on 29  degrees of freedom
Residual deviance:  24.758  on 21  degrees of freedom
AIC: 166.19

Number of Fisher Scoring iterations: 4
```

LASSO selected AgeGroup and CLIstd as main predictors, supporting previous GLM findings

```
> BIC(model5)
[1] 178.8029
> rsq(model5, adj=TRUE)
[1] 0.8035233
> testDispersion(model5)

        DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated

data:  simulationOutput
dispersion = 0.82572, p-value = 0.616
alternative hypothesis: two.sided
```



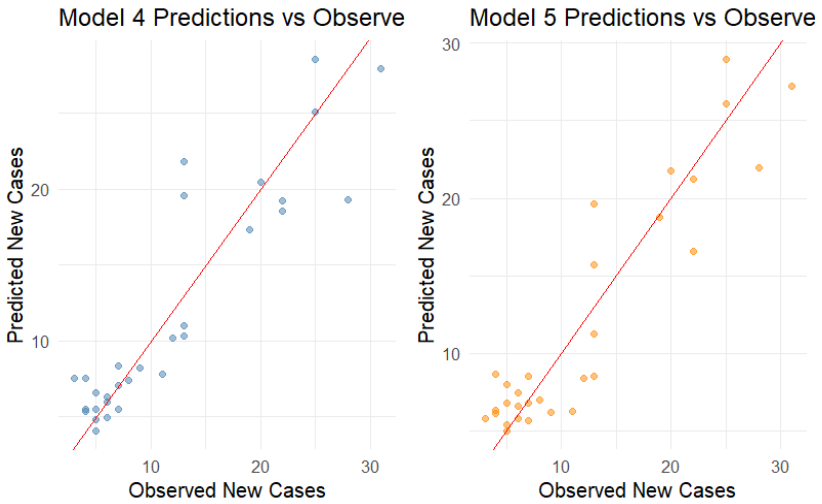
In the LASSO model, the interaction term (CLIstd:BMImedian) was penalized to zero, meaning the algorithm considered it non-essential for improving predictive accuracy under

cross-validation. However, in model 4, including this interaction improved model fit (e.g., lower BIC and higher pseudo-R^2). This difference arises because LASSO prioritizes model simplicity and generalization, whereas standard regression prioritizes in-sample explanatory power.

If we try an elastic net we can see that the model actually keeps the interaction term, however it is practically zero.

```
> #We try using Elastic net
> set.seed(123)
> cv_elastic <- cv.glmnet(
+   X, y,
+   family = "poisson",
+   offset = offset_var,
+   alpha = 0.5,  # Elastic Net
+   nfolds = 10
+ )
> # Coefficients
> coef(cv_elastic, s = "lambda.min")
12 x 1 sparse Matrix of class "dgCMatrix"
                     lambda.min
(Intercept)       -9.868209685
AgeGroup40-59      .
AgeGroup60-79      0.927054773
RegionRegion2      0.076877188
RegionRegion3      .
RegionRegion4     -0.016654776
RegionRegion5      .
CLIstd            -0.139097022
BMImedian          .
Sex                .
SmokingPrevalence  0.894556968
CLIstd:BMImedian  -0.001008695
```

Finally we use the two best models (4 and 5) to see how well they can predict new cases in pancreatic cancer.



Overall, this analysis highlights the strong influence of age and lifestyle on pancreatic cancer incidence and suggests that promoting healthier lifestyles could have a measurable impact on reducing cancer risk. However, given the cross-sectional design, causality cannot be established, and future research should extend these findings using longitudinal data and additional behavioral or genetic covariates.