Assignment - introduction to data science

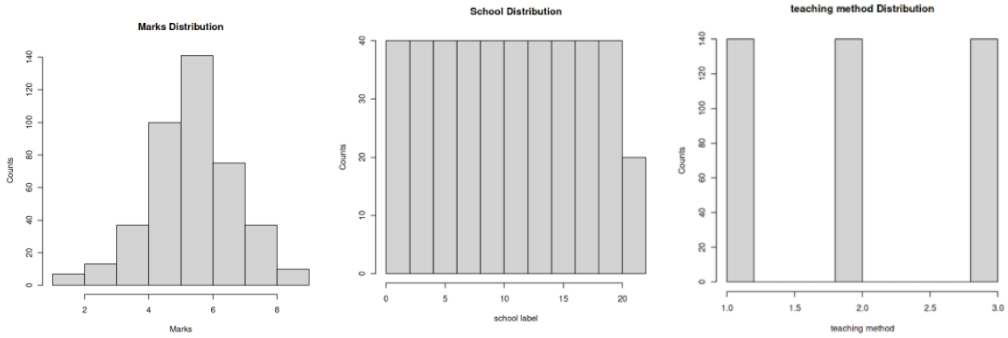
Exercise 1

Point a)

Does the teaching method (as a fixed effect) causally influence students' performance scores, and how much of the variation in performance is attributable to differences between schools (as a random effect)?

Point b)

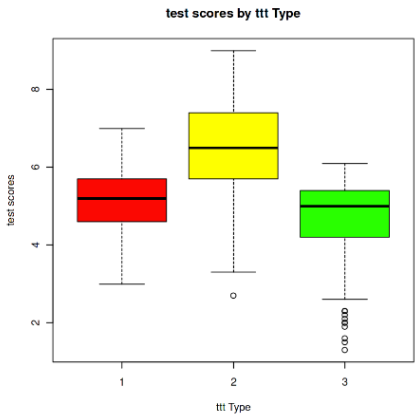
The research question raises from the observation of the histograms of our dataset columns:

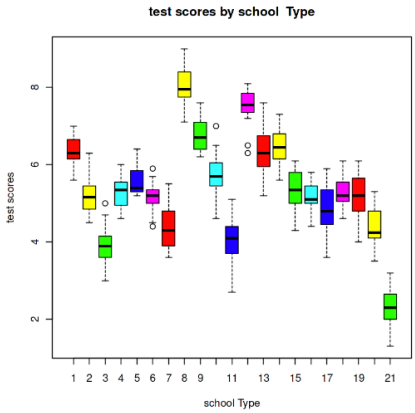


We can see that counts of schools and teaching method are more or less unform, whereas marks are not, hence it makes sense to make inference on the marks column (i.e. test scores) to know why there is variation, hence the choice of marks as a target. Moreover, we came up with such research question also by observing boxplots explaining how test scores are distributed based on teaching method and by having explored that each school adopts a unique teaching method (below the summary table).

school	unique_methods	methods	single_method
1	1	1	TRUE
2	1	1	TRUE
3	1	1	TRUE
4	1	1	TRUE
5	1	1	TRUE
6	1	1	TRUE
7	1	1	TRUE
8	1	2	TRUE
9	1	2	TRUE
10	1	2	TRUE
11	1	2	TRUE
12	1	2	TRUE
13	1	2	TRUE
14	1	2	TRUE
15	1	3	TRUE
16	1	3	TRUE
17	1	3	TRUE
18	1	3	TRUE
19	1	3	TRUE
20	1	3	TRUE
21	1	3	TRUE

We observed that since there are only 3 classes of teaching methods and there are 21 different schools, treating the school type as random effect avoids estimating so many parameters. We also assume that there’s no importance order neither in teaching methods nor in schools.





The left boxplot indicates that there are 3 teaching methods and for each we have the bold horizontal bar representing the median value. Generally all samples are distributed between the 1st and 3rd quartile, but we can see some outliers in the 3rd teaching method. Same for school types, generally all samples are in between such quartiles.

Super important, we have to transform the "school" and "ttt" variables into categorical because there's no school more important than any other, same for teaching method.

c) Formal Analysis

i)

We are treating the teaching method as a fixed effect and the school type as a random effect, hence we have a mixed effects model whose equation is:

$$Y = X\beta + Z\gamma + \varepsilon$$

Where:

- Y is the observed values array of the tests
- β is the array of coefficients of the underlying linear model
- X are the covariates for the fixed effects ie, teaching method
- Z is the sparse design matrix for random effects ("school" effect)
- $\gamma \sim N(0, \sigma^2 random)$ and $Z * \gamma = u_{school}$ represents the random effect contribution of schools (modeled as deviations from the overall mean).
- ε is the Residual variation.

In our case we have:

$$test_i = \beta_0 + \beta_1 ttt2_i + \beta_2 ttt3_i + u_{school} + \varepsilon_i$$

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: test ~ ttt + (1 | school)
   Data: data

REML criterion at convergence: 689.6

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.83181 -0.61447 -0.07317  0.66463  2.57447

Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept) 1.1533   1.0739
 Residual             0.2425   0.4924
Number of obs: 420, groups:  school, 21

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)   5.1214     0.4080 18.0000  12.552 2.44e-10 ***
ttt2          1.3007     0.5770 18.0000   2.254   0.0369 *
ttt3         -0.4729     0.5770 18.0000  -0.819   0.4232
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
     (Intr) ttt2
ttt2 -0.707
ttt3 -0.707  0.500
```

where β_0 is the intercept(baseline for ttt1)=5.1214.

All the beta coefficients can be found under the column "Estimate".

ii)
Hereafter the null and alternative hypothesis for the fixed effects:
H0:$\beta_1 = \beta_2 = 0$

H1: $at\ least\ 1\ \beta\ is\ \neq 0$
For all teaching methods.
Meaning that we want to test if the teaching method has influence on the tests;
Instead, the null and alternative hypothesis for random effects are:
H0: $\sigma^2_{school} = 0$

H1: $\sigma^2_{school} > 0$
iii)
If the variance for the random effect is much greater than zero and significantly contributes to the model fit, it supports the idea that schools influence performance scores.

A anova: 1 × 6

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>
ttt	2.457069	1.228535	2	18	5.066384	0.01796969

In the above table we can see the result of the ANOVA for our model.
We are testing the effect of the fixed effect variable "ttt" in order to test whether it significantly affects the performance scores while accounting for the random effects of schools. Below the explanation of the columns for the ANOVA screenshot:

- Sum Sq: it represents the total variation in the outcome (test scores) attributed to the fixed effect of the teaching method.
- Mean Sq: average variation per degree of freedom for our fixed effect
- NumDF and DenDF (numerator degrees of freedom) corresponds to the number of parameters being tested for the fixed effect (here, 2 for the levels of ttt). DenDF (denominator degrees of freedom, computed numerically) reflects the effective sample size available for estimating residual variance, accounting for the random effects and structure of the model. Both contribute in computation of F-statistics.
- F-value: quantifies how much larger the variation explained by the teaching method is compared to random error or residual variation. The larger the F-value, the more evidence against the null hypothesis.
- P-value: probability of observing a more extreme event with respect to the observed value (F-value in the above table) given the null hypothesis is true. We can see that with a significance level of 5% we can reject the null hypothesis stating that the teaching method makes no difference in the tests.

p value for LRT testing significance of random effects

```
model_null=lm(test ~ ttt , data = data)
pchisq(as.numeric(2 * (logLik(model, REML = FALSE) - logLik(model_null))), 1, lower.tail = FALSE)
```

4.35052745145919e-130

As we can see from the Likelihood Ratio Test (LRT) results above, the p-value is very small, which suggests that the random effect of schools is statistically significant at all significance levels. This means that there is significant variation in student performance between schools.

p value for LRT testing significance of fixed effects

```
reduced_model <- lmer(test ~ (1 | school), data = data, REML = FALSE)
pchisq(as.numeric(2 * (logLik(model) - logLik(reduced_model))), df = 1, lower.tail = FALSE)
```

0.00232197295383208

The fixed effect of the reaching method is also statistically significant with an alpha 1%, meaning that the test scores are influenced by the teaching method.

```
#icc
var_school = 1.1533
var_residual = 0.2425
# Calculating the Intraclass Correlation Coefficient (ICC)
icc = var_school / (var_school + var_residual)
icc
```

0.826264507809142

From the ICC computation we can see that almost 82 percent of total variability in our tests are explained by school differences.

d) Answer the initial research problem.

We can conclude that we have statistical evidence that based on the p-value of around 0.018 (overall effect of teaching method):

- Statistical evidence supports the claim that the teaching method influences test scores.
- The magnitude of the effect depends on the specific teaching method:
    - The baseline teaching method (`ttt1`) is associated with a mean test score of 5.1214, which is significantly different from zero ($p < 0.001$).
    - Teaching Method 2 shows a significant positive effect (+1.30 points).
    - Teaching Method 3 does not show a statistically significant effect compared to the baseline.

Regarding the random effect of school type, there is substantial variability in test scores attributable to differences between schools, as evidenced by a variance component of 1.1533 and an ICC of 82.6%. This indicates that schools play a dominant role in explaining the variability in student performance. Furthermore, the significant p-value from the likelihood ratio test suggests strong evidence that these school-level differences are not random but meaningful.

e) Discussion

In our case, we did not initially expect that school type would play such a dominant role in explaining the variability in student performance. This suggests that a substantial proportion of the total variability in test scores is attributable to differences between schools. While we verified that our dataset is balanced in terms of observations per school type, the unexpected dominance could lie in unmeasured school-level factors such as differences in language of instruction, class schedule management, teaching resources, or other institutional characteristics.

Exercise 2

Point a)

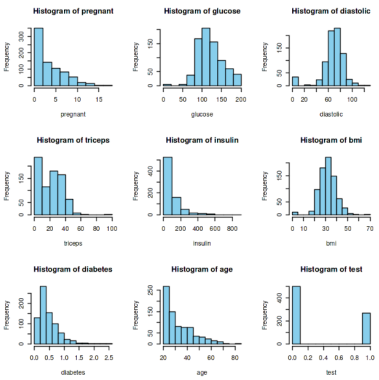
What are the key variables associated with an increased risk of diabetes in adult females from the Pima Indian population, and how much of the variance is explained by those variables?

Point b) EDA

```
summary(data)
```

```
    pregnant         glucose        diastolic         triceps     
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00  
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00  
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00  
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54  
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00  
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00  
    insulin           bmi          diabetes          age       
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00  
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00  
 Median : 30.5   Median :32.00   Median :0.3725   Median :29.00  
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24  
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00  
 Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00  
      test       
 Min.   :0.000  
 1st Qu.:0.000  
 Median :0.000  
 Mean   :0.349  
 3rd Qu.:1.000  
 Max.   :1.000  
```

In the above summary we quantities describing features distributions such as min, max, mean etc... We checked also that the dataset has no missing values, and we also checked for values that make no sense. We noticed that "bmi","insulin","triceps","diastolic","glucose" have zero values, which are biologically impossible for an alive person. Since the rows containing such values are almost half of the total, instead of deleting the rows, we imputed nonsense values with the mean of each covariate.



In the above picture we can see the histograms of covariates and for the target variable.

This visualization helps us in understanding which features are informative. We can say that all of them are, since there aren't covariates with a value being equal for all the rows (ie no histogram with just 1 column). We can also observe that the target "test" in not balanced, in favor the zero value.

```
table(data$test)
prop.table(table(data$test))#DATASET UNBALANCED IN FAVOR OF CLASS 0
```

```
  0   1
500 268

        0         1
0.6510417 0.3489583
```

In the above image we can quantify the counts of class zero are 500 (65% of total observations) vs 268 of class 1 (35% of total observations). This can affect our final statistical model.

	pregnant	glucose	diastolic	triceps	insulin	bmi	diabetes	age	test
pregnant	1.00000000	0.1279115	0.208522309	0.08298907	0.05602701	0.02156505	-0.033522673	0.54434123	0.2218982
glucose	0.12791147	1.0000000	0.218366918	0.19299109	0.42015709	0.23094124	0.137059710	0.26653352	0.4929277
diastolic	0.20852231	0.2183669	1.000000000	0.19281584	0.07251688	0.28126771	-0.002763364	0.32459494	0.1660737
triceps	0.08298907	0.1929911	0.192815844	1.00000000	0.15813897	0.54239773	0.100966445	0.12787247	0.2152992
insulin	0.05602701	0.4201571	0.072516882	0.15813897	1.00000000	0.16658610	0.098633942	0.13673386	0.2144110
bmi	0.02156505	0.2309412	0.281267706	0.54239773	0.16658610	1.00000000	0.153399971	0.02551918	0.3119244
diabetes	-0.03352267	0.1370597	-0.002763364	0.10096644	0.09863394	0.15339997	1.000000000	0.03356131	0.1738441
age	0.54434123	0.2665335	0.324594939	0.12787247	0.13673386	0.02551918	0.033561312	1.00000000	0.2383560
test	0.22189815	0.4929277	0.166073669	0.21529921	0.21441095	0.31192439	0.173844066	0.23835598	1.0000000

The above correlation matrix shows that the features are not highly correlated with one another, which is advantageous for a linear model, as it suggests minimal multicollinearity and that each feature contributes its own unique information to the prediction of the target. The most correlated covariates are "age" and "pregnant" with a value of about 0.54. (We also included the target variable "test" in the above matrix)

c) Formal Analysis

i)

We tried to fit a GLM without smoothed covariate, and we applied model selection via stepAIC() function; the best model we got has AIC =726.2. We then tried to apply smoothing to all covariate, using a "GAM".

```r
full_smooth_model <- gam(test ~ s(pregnant) + s(glucose) + s(diastolic) + s(triceps) + s(insulin) + s(bmi) + s(diabetes) +  s(age),
                         family = binomial, data = data)

# Sommario del modello
summary(full_smooth_model)
```

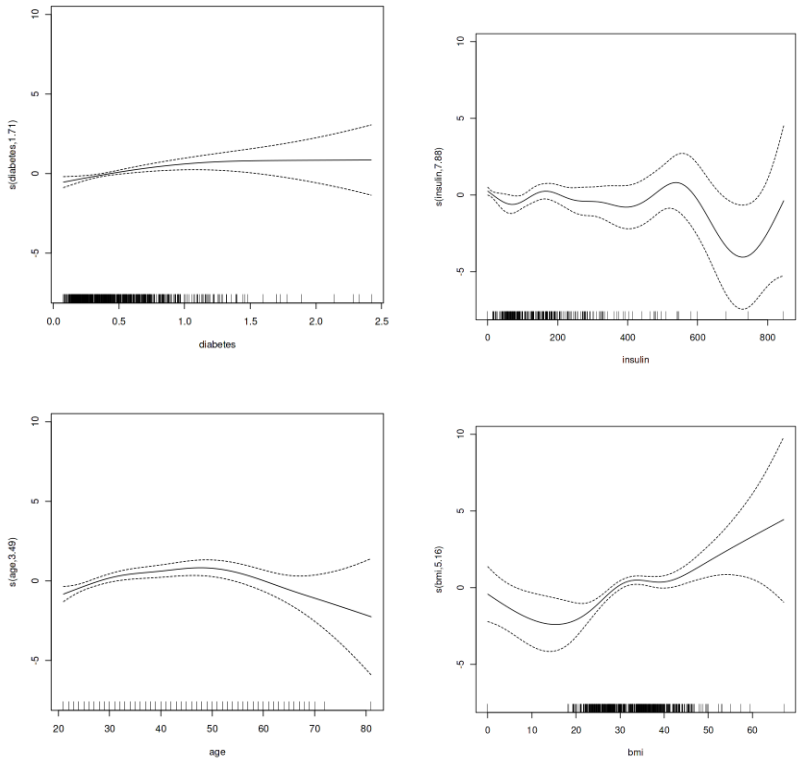
```
Family: binomial
Link function: logit

Formula:
test ~ s(pregnant) + s(glucose) + s(diastolic) + s(triceps) +
    s(insulin) + s(bmi) + s(diabetes) + s(age)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0629     0.1173  -9.058   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
               edf Ref.df Chi.sq  p-value
s(pregnant)  1.000  1.000  2.040 0.153268
s(glucose)   1.000  1.000 85.685  < 2e-16 ***
s(diastolic) 1.000  1.000  2.170 0.140759
s(triceps)   1.000  1.000  0.073 0.787053
s(insulin)   7.366  8.333 16.462 0.053711 .
s(bmi)       3.845  4.830 32.889 5.68e-06 ***
s(diabetes)  1.797  2.259 12.820 0.002351 **
s(age)       3.634  4.497 22.131 0.000424 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.385   Deviance explained = 35.1%
UBRE = -0.10441  Scale est. = 1         n = 768
```

We fitted the generalized additive model additive model with family binomial because we have a binary target and we want to model probability of and individual to fall into class zero or 1 (ie manifesting diabetes symptoms or not). With this model all the coefficients undergo a smoothing function. We observe that the covariates that are treated linearly (ie without smoothing) are “pregnant”, “diastolic” and “triceps” since their edf (effective degrees of freedom) is 1; the other covariates must be treated as non linear.

ii) Linear coefficients interpretation:

- intercept: used to compute the baseline probability of test=1. So we have $P(test = 1) = \frac{e^{-1.0629}}{1+e^{-1.0629}} = 0.3477$ meaning that the baseline probability outcome is 34.77% when all predictors are zero.
- Pregnant: for every 1 unit of increase of “pregnant” the log odds of the outcome changes by the Pregnant coefficient;
- Diastolic : for every 1 unit of increase of “Diastolic” the log odds of the outcome changes by the Diastolic coefficient;
- Triceps: for every 1 unit of increase of “triceps” the log odds of the outcome changes by the triceps coefficient;
- glucose: for every 1 unit of increase of “glucose” the log odds of the outcome changes by the glucose coefficient;
-

Hereafter the plots for the non-linear coefficients:



Our smooth effects plots illustrate the relationships between the predictors and the response variable while accounting for the overall model structure. The predictor **"diabetes"** shows a nearly linear, weak increasing trend, hence, as the diabetes value increases, the log-odds of the outcome (test = 1) gradually increase. **"Insulin"** displays a complex, non-linear relationship, with notable fluctuations, suggesting multiple regions of change; the ranges where log-odds decrease are right after 0 and around 600–700. This non-linearity suggests that the relationship between insulin and the outcome cannot be captured by a simple linear term. The confidence intervals widen significantly at the high end of the insulin range (above 600), indicating greater uncertainty due to sparse data.**"Age"** has a mild, curved trend, peaking around 50 and declining slightly thereafter, thus having a region of increase (before peak) and decrease (after peak) on the log-odds of the outcome . **"BMI"** shows a clear non-linear increasing trend, indicating a stronger association with the log-odds of the response as BMI increases.
In general the steeper the curve, the highest the impact on the log-odds of the target.

However we can see that the only significant covariates are the intercept, “glucose”, “insulin”, “bmi”, “diabetes” and “age”. We greedily select the best model by removing covariates with less significance and checking that AIC did not increase to minimize the Akaike information criterion.

iii)model selection

```
model <- gam(test ~  s(glucose) + s(diastolic) + s(insulin) + s(bmi) + s(diabetes) +  s(age), family = binomial, data = data)
summary(model)
```

```
Family: binomial
Link function: logit

Formula:
test ~ s(glucose) + s(diastolic) + s(insulin) + s(bmi) + s(diabetes) +
    s(age)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0602     0.1166  -9.095   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
             edf Ref.df Chi.sq  p-value
s(glucose)   1.000  1.000 85.376  < 2e-16 ***
s(diastolic) 1.000  1.000  2.156  0.14210
s(insulin)   7.267  8.261 16.184  0.05378 .
s(bmi)       3.822  4.802 37.087 4.68e-07 ***
s(diabetes)  1.833  2.307 12.697  0.00267 **
s(age)       3.624  4.474 39.884  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.386   Deviance explained = 34.9%
UBRE = -0.10699  Scale est. = 1         n = 768
```

```
AIC(model)
```

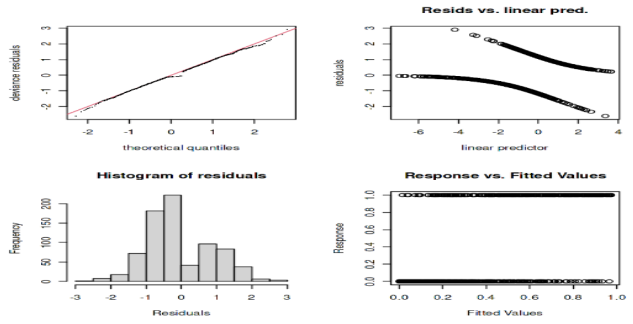
```
685.830247871666
```

Above the summary of the obtained model. We have a slightly lower AIC (which is around 685.83) compared to the full model, which was 687.815.

iv)

```
gam.check(model)
```

```
Method: UBRE   Optimizer: outer newton
full convergence after 8 iterations.
Gradient range [-2.121081e-07,1.623851e-08]
(score -0.1069919 & scale 1).
Hessian positive definite, eigenvalue range [1.339675e-07,0.002815946].
Model rank =  55 / 55

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

             k'  edf k-index p-value
s(glucose)    9.00 1.00    0.97    0.23
s(diastolic) 9.00 1.00    1.05    0.94
s(insulin)    9.00 7.27    1.00    0.49
s(bmi)        9.00 3.82    1.01    0.65
s(diabetes)   9.00 1.83    1.00    0.57
s(age)        9.00 3.62    0.99    0.51
```



We ran a gam.check() on the new model to validate the assumptions.

- The **QQ-plot** of the deviance residuals versus theoretical quantiles shows the residuals closely overlapping the red reference line, suggesting that the residuals are appropriately modeled with no severe deviations.
- In the **Residuals vs. Linear Predictor** plot, the two smooth curves reflect the binary nature of the response variable (0 or 1).This pattern is expected in logistic regression and does not indicate a problem with the model fit. Why do we have those two lines of points? Because we predict a probability for a variable taking values 0 or 1. If the target value is 0, then we always predict more, and residuals must be negative (points on curve lower) and if the true value is 1, then we underestimate, and residuals have to be positive (points on the upper curve). We see that the lower line of points in centered mostly around residual axis=0, meaning that the model can better predict negative, then positive class.

- The **histogram of residuals** appears reasonably symmetric and is approximately centered around zero, with no evident skewness or extreme outliers.
- The **Response vs. Fitted Values** plot shows that when the fitted value is close to 0, the response is also predominantly 0 (evidenced by the higher density of points near (0,0)). Similarly, when the fitted value is close to 1, the response is predominantly 1 (denser points near (1,1)). The two horizontal lines at 0 and 1 reflect the binary nature of the response variable. This pattern is expected in a binary dataset and suggests that the model generally distinguishes between the two levels of the response variable. In an ideal model, all points would cluster tightly around (0,0) and (1,1), indicating perfect classification.

Since we observe EDF > 1 for most predictors in a GAM, it suggests that the model is capturing non-linear relationships, and therefore, the linearity assumption of a traditional linear model is not appropriate here. The assumption of residual normality holds since the QQ plot resembles a line like y=x. Finally The histogram of residuals shows a symmetric and reasonable distribution.

d) Answer the initial research problem.

The answer to our research question is : the key variables associated with an increased risk of diabetes in adult females from the Pima Indian population are glucose, diastolic, insulin, bmi, diabetes and age. Together, these variables explain 34.9% of the deviance in the target , indicating a moderate explanatory power, with room for improvement by including additional predictors or interactions.

e) Discussion

We know that our dataset in unbalanced and this can affect how model explanatory power (possibly biasing our model). The feature selection step could have been carried out in a more refined way.
Also the imputation strategy of the mean should have been discussed with a domain expert to know if it's the right thing to do.