Generalized Linear Models (GLM)

Linear models (lm) have the standard assumptions of independent and identically distributed normal random variables. Often response variables break these assumptions, and generalized linear models (GLMs) are excellent at dealing with them.

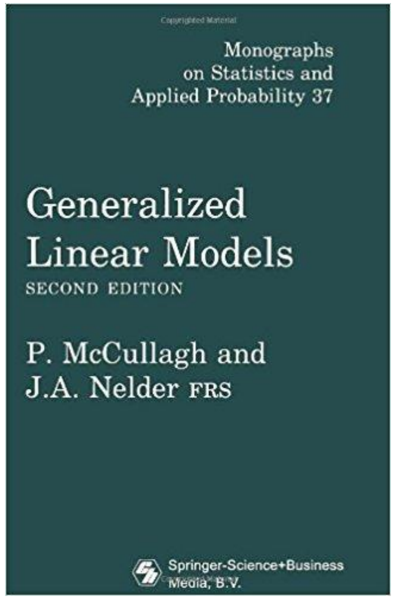
GLM is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

The use of GLMs is recommended either when:

- the variance is not constant, and/or
- the errors are not normally distributed

Specifically, GLMs should be used when the response variable is:

- count data expressed as proportions
- count data that are not proportions, integers
- binary response variables



Copyrighted Material

Monographs
on Statistics and
Applied Probability 37

Generalized
Linear Models
SECOND EDITION

P. McCullagh and
J.A. Nelder FRS

Springer-Science+Business
Media, B.V.

Geographical data is typically non-normally distributed

- In the earlier chapter on linear regression, we focused primarily on the classic setting where the response y is continuous and typically assumed to have a normal distribution, at least approximately.
- However, in most of the geographical data analysis examples, the data to be modeled are clearly non-normal.
- For instance, we may have a binary response (e.g. presence or absence). Binary variable is clearly non-normal.
- Often the response of interest is a count, e.g. the number of plants in a 1-m2 plot. The Poisson distribution is often used to model count data and a Poisson regression can be used to relate count responses to predictors.

Components of the GLM

Error structure – refers to the probability distribution of the response variable (Y); e.g. normal distribution for Y in the linear regression, binomial distribution for Y in the binary logistic regression and Poisson distribution for Y in the Poisson regression with count data. Also called a noise model or error model.

Linear predictor - specifies the explanatory variables (X1, X2, ... Xk) in the model, more specifically their linear combination in creating the so called linear predictor; e.g., $\beta0 + \beta1x1 + \beta2x2$ as we have seen in a linear regression, or as we will see in a logistic regression in this lesson.

Link lunction - specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables; e.g. identity link for GLM with normal distribution, logit for logistic regression or log for GLM with Poisson distribution.

Error structure and GLM

A GLM allows the specification of a variety of different error distributions, two most common ones with geographical data:

- Poisson errors, useful with count data, integers (0,1,2…n);
- Binomial errors, useful with data on proportions, (0 and 1);

The Poisson distribution may be useful to model events such as

- number of species, trees, boulders or solifluctions in an interval of time or space (e.g. how many individual trees per 1-m2 plot)

The binomial distribution may be useful to model events such as

- presence-absence, active-inactive or death-alive in an interval of time or space (e.g. probability of tree presence per 1-m2 plot)

Linear predictor

- The data Y1, Y2, ..., Yn are independently distributed, i.e., cases are independent.
- The dependent variable Yi does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal,...)
- GLM does NOT assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the transformed response in terms of the link function and the explanatory variables

Link function and GLM

- The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.
- The link function provides the relationship between the linear predictor and the mean of the distribution function.
- Typical link functions in relation to error distribution of the data:

Error distribution	**Link function**	**Data**	**Anova-test**
Normal	*Identity*	*Gaussian*	*F or Chisq*
Poisson	*Log*	*Count data*	*Chisq*
Binomial	*Logit*	*Binary (0/1)*	*Chisq*

Non-normal errors and GLM

Up to this point, we have dealt with the statistical analysis of data with normal errors. In practice, however, many kinds of data have non-normal errors, for example:

- errors that are strongly skewed;
- errors that are kurtotic;
- errors that are strictly bounded (as in proportions);
- errors that cannot lead to negative fitted values (as in counts).

In the past, the only tools available to deal with these problems were transformation of the response variable or the adoption of non-parametric methods.

Family argument

- The error structure is defined by means of the family argument, used as part of the model formula. Examples are *glm(y ~ x, family* = *poisson)* which means that the response variable y has Poisson errors, and *glm(y ~ x, family = binomial)* which means that the response is binary, and the model has binomial errors.
- As with previous models, the explanatory variable *x* can be continuous (leading to a regression analysis) or categorical (leading to an ANOVA-like procedure called analysis of deviance)
- The general linear model (lm in R) may be viewed as a special case of the generalized linear model with identity link and responses normally distributed.

Summary of advantages of GLMs over traditional regression

- No need to transform the response Y to have a normal distribution.
- The choice of link is separate from the choice of random component thus we have more flexibility in modeling.
- The models are fitted via Maximum Likelihood estimation; thus optimal properties of the estimators.
- Most of the inference tools and model checking for linear regression models apply for the GLMs too.
- There is often one procedure in a software package to capture all the models listed above, e.g. glm() in R, etc... with options to vary the three components.

Practical benefits of GLMs

- GLMs constitute a more flexible family of regression techniques than traditional multivariate modelling tools.
- GLMs do not force data into unnatural scales; they allow for non-linearity and non-constant variance structures in the data.
- The GLM approach enables scientists to use a wide range of environmental data types, such as discrete, categorical, ordinal and continuous data, under a single theoretical and computational framework.
- GLM techniques combined with a geographic information system can play an important role in analysing and modelling spatial data sets.

GLM functions

Generalized linear model in R – main functions (almost like in lm!)

- ?glm
- glm() for fitting a generalized linear model
- summary() for extracting model results (estimated coefficients and their p-values)
- anova() significance of the model terms, model inter-comparison
- I(x^2) define second order polynomial term
- resid() extract model residuals
- fitted() extract fitted values from the model object
- predict.glm() use the fitted model to predict Yi at Xi

glm() function

Generalized linear models are fit using the glm( ) function. The form of the glm function is:

glm(formula, family=familytype)

more precisely:

```
gdd_glm <- glm(w ~ x + z, family="gaussian") # normally-distributed data
spr_glm <- glm(y ~ x + z, family="poisson") # count data
empher_glm <- glm(v ~ x + z, family="binomial") # binomial data
```

Prediction in glm

predict.glm

Obtains predictions and optionally estimates standard errors of those predictions from a fitted generalized linear model object. Note, *type*-argument!

The type of prediction required. The default is on the scale of the linear predictors; the alternative "response" is on the scale of the response variable. In most of the cases use *type="response"*

```
prediction <- predict.glm(glm_model, new_data, type="response")
```

GLM examples

Normally distributed data

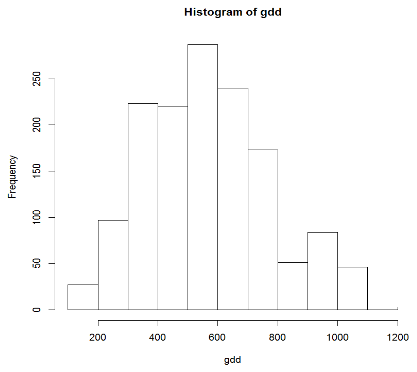
```
gdd_glm <- glm(gdd ~ xcoord + ycoord, family="gaussian")
summary (gdd_glm)
anova (gdd_glm, test="F")
```

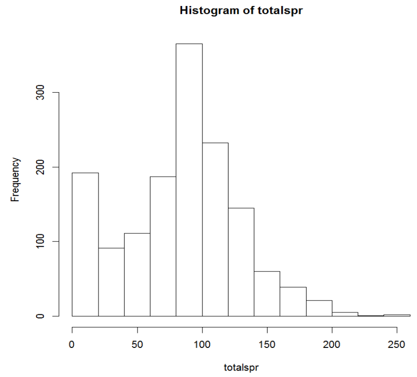
Poisson data (0,1,2,3…n) integers

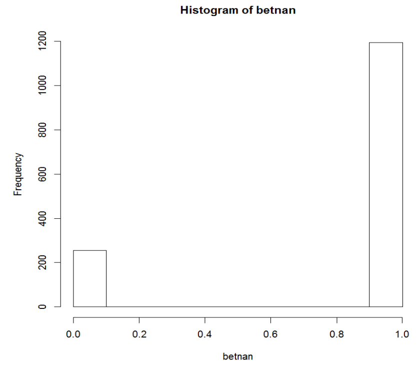
```
totalspr_glm <- glm(totalspr ~ xcoord + ycoord, family="poisson")
summary (totalspr_glm)
anova (totalspr_glm, test=“Chisq")
```

Binomial data (0,1)

```
betnan_glm <- glm(betnan ~ xcoord + ycoord, family="binomial")
summary (betnan_glm)
anova(betnan_glm, test="Chisq")
```







GLM and explained deviance

Linear models come with an R-squared value that measures the proportion of variation that the model accounts for. The R-squared is provided with summary(model) in R.

For generalized linear models (GLMs), the equivalent is the amount of deviance accounted for (D-squared), but this value is not provided with the summary (model).

Explained deviance (D-squared) = (Null deviance – Residual deviance) / Null deviance

Both Null deviance and Residual deviance are provided by summary(model) and anova(model)

Explained deviance: example

- Deviance is a goodness-of-fit statistic for a statistical model.
- It is a generalization of the idea of using the sum of squares of residuals in ordinary least squares to cases where model-fitting is achieved by maximum likelihood.
- Deviance plays an important role in GLMs
- Null deviance = 1352.2
- Residual deviance = 1320.5
- D-squared = 0.0234
- In the glm-model example two variables explained only 2.3% of the deviance.

```
> anova(betnan_glm, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: betnan

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    1450     1352.2
xcoord  1  30.8400      1449     1321.3 2.802e-08 ***
ycoord  1   0.8338      1448     1320.5    0.3612
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> (1352.2 - 1320.5)/1352.2
[1] 0.02344328
```

Poisson GLM, example: model summary

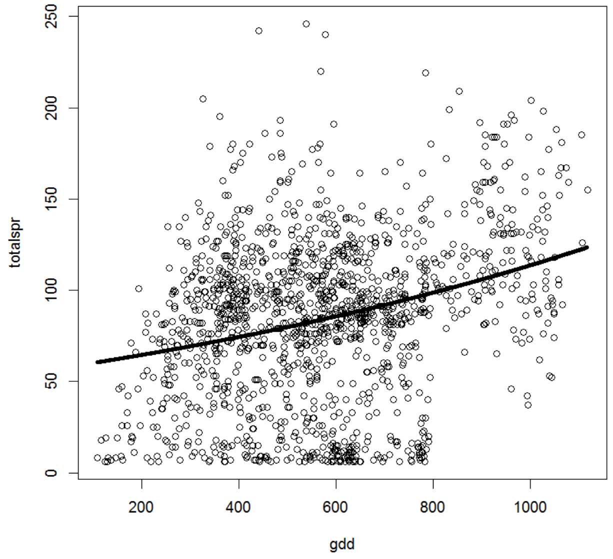
- Poisson regression is useful when predicting an outcome variable representing counts from a set of continuous predictor variables.
- How much does gdd explain the variation of totalspr in NW Finland?
- gdd = growing degree days
- totalspr = species richness per 1 km2
- Z-value is the test-statistic for the Wald-test that the parameter is 0

```
> totalspr_glm <- glm(totalspr ~ gdd, family="poisson")
> summary(totalspr_glm)

Call:
glm(formula = totalspr ~ gdd, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-12.223   -3.066    0.237    2.876   15.034

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 4.026e+00  8.560e-03  470.39   <2e-16 ***
gdd         7.066e-04  1.348e-05   52.42   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 39447  on 1450  degrees of freedom
Residual deviance: 36743  on 1449  degrees of freedom
AIC: 45509

Number of Fisher Scoring iterations: 5
```

Poisson GLM, example: prediction and plot with a trend line

- predict(object, newdata = NULL, type = c("link", "response", "terms"), se.fit = FALSE, dispersion = NULL, terms = NULL, na.action = na.pass, ...)
- type="response" gives the predictions on the scale of the response, in Poisson regression as continuous value >= 0

```
> totalspr_glm <- glm(totalspr ~ gdd, family="poisson")
> gddv <- seq (min(gdd),max(gdd),1)
> yv <- predict (totalspr_glm, list(gdd=gddv),type="response")
>  plot(gdd,totalspr)
>  lines(gddv,yv)
```



Binomial GLM, example: model summary

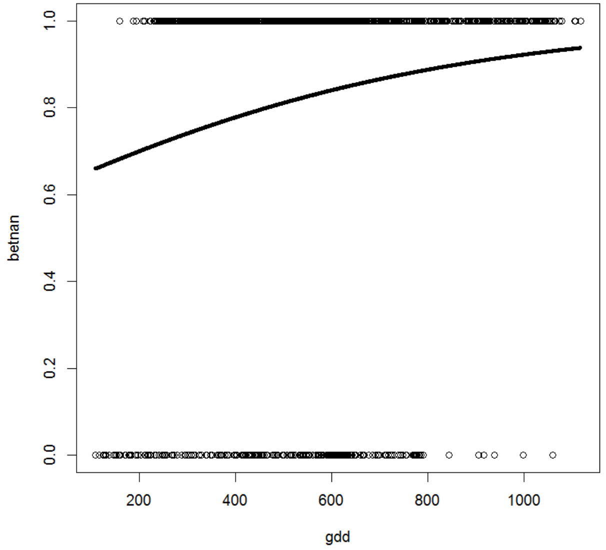
- Binomial regression is suitable when predicting an outcome variable representing presence-absence from a set of continuous predictor variables.
- How much does gdd explain the variation of betnan in NW Finland?
- gdd = growing degree days
- betnan = Betula nana per 1 km2
- Z-value is the test-statistic for the Wald-test that the parameter is 0

```
> betnan_glm <- glm(betnan ~ gdd, family="binomial")
> summary(betnan_glm)

Call:
glm(formula = betnan ~ gdd, family = "binomial")

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.3107   0.4340   0.5816   0.6774   0.8742  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) 0.441005   0.200421   2.200   0.0278 *  
gdd         0.002034   0.000365   5.572 2.51e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1352.1  on 1450  degrees of freedom
Residual deviance: 1318.8  on 1449  degrees of freedom
AIC: 1322.8

Number of Fisher Scoring iterations: 4
```

Binomial GLM, example: prediction and plot with a trend line

- predict(object, newdata = NULL, type = c("link", "response", "terms"), se.fit = FALSE, dispersion = NULL, terms = NULL, na.action = na.pass, ...)
- type="response" gives the predictions on the scale of the response, in binomial regression as a probability value between 0 and 1

```
> betnan_glm <- glm(betnan ~ gdd, family="binomial")
> gddv <- seq (min(gdd),max(gdd),1)
> yv <- predict (betnan_glm, list(gdd=gddv),type="response")
>  plot(gdd,betnan)
>  lines(gddv,yv, lwd=4)
```



Weak positive response shape: betnan probability of presence increase with increasing gdd-value

Binomial GLM: exact prediction of probability values based on model coefficients

```
Coefficients:
             Estimate   Std. Error   z value  Pr(>|z|)
(Intercept)  0.441005   0.200421     2.200    0.0278 *
gdd          0.002034   0.000365     5.572    2.51e-08 ***
```

What is the probability (P) of betnan presence at 200 gdd?

P = 1 / (1 + exp(-Z))

P is the probability that observation belongs to group 1 and Z = b0 + b1*X1 + ... + bp*Xp

exp = exponent function, i.e. x^2.718282 (exp-function is the inverse of log-function)

Z = 0.441005 + (0.002034)*gdd

P = 1 / (1 + exp(-Z)) at the gdd-value 200

Z = 0.441005 + (0.002034)*200 = 0.847805

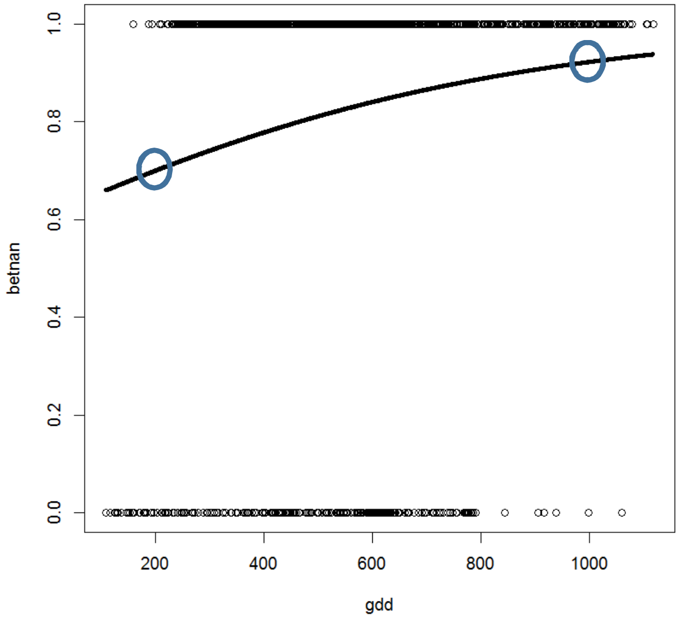
P = 1 / (1 + exp(-(0.847805))) = **0.7001065**

and the gdd-value 1000

P = 1 / (1 + exp(-Z)) at the gdd-value 1000

Z = 0.441005 + (0.002034)*1000 = 2.475005

P = 1 / (1 + exp(-(2.475005))) = **0.9223709**



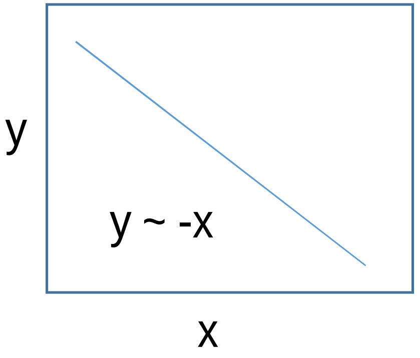
Response shape

Determination of variable-y response on studied variable (gradient) represents one of the basic tasks in modelling.

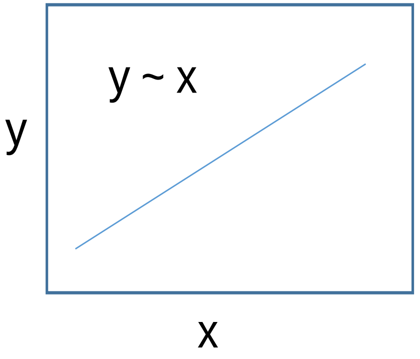
Response curve allows estimation of variable-y optimum along the environmental gradient.

Most of widely used statistical methods assume that response on gradient have symmetrical bell shape of Gaussian curve, even if number of studies showed that this type of response occurs in real data quite rarely.

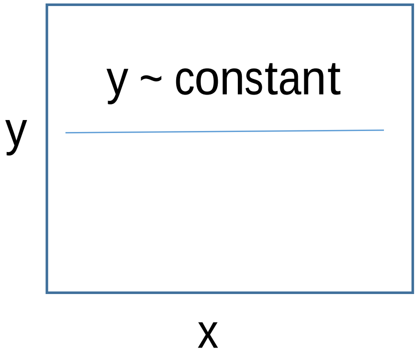
Negative, linear



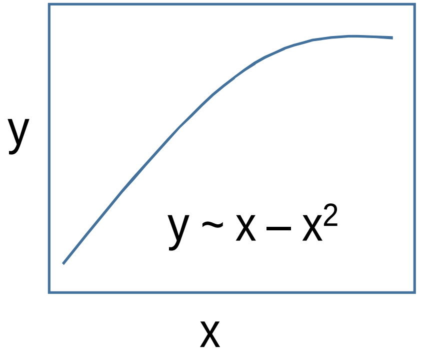
Positive, linear



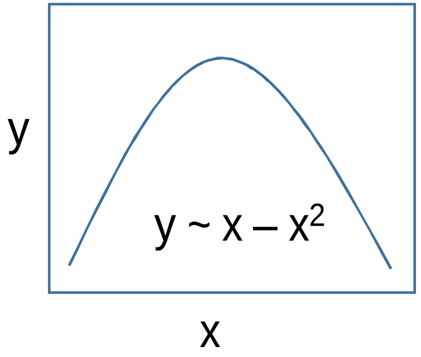
No effect



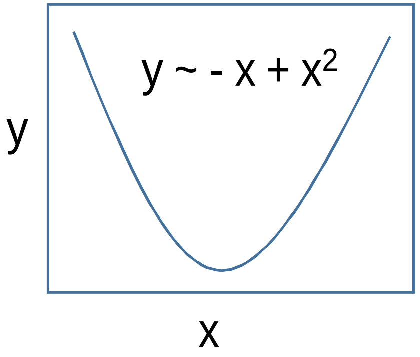
Positive, saturating



Hump shape or
bell shape



U-shape



AUC-value: accuracy of binomial models

The area under the receiver operating characteristic (ROC) curve, known as the AUC, is currently a standard method to assess the accuracy of binomial models.

AUC avoids the supposed subjectivity in the threshold selection process, when continuous probability derived scores are converted to a binary presence–absence variable, by summarizing overall model performance over all possible thresholds.

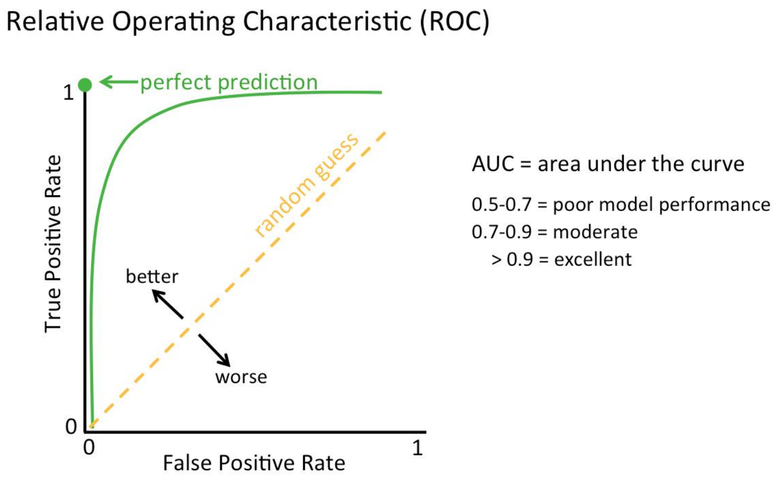
The calculation of the area under this curve (the AUC score) provides a single-number discrimination measure across all possible ranges of thresholds.

Sensitivity: the correctly predicted positive fraction,

Specificity: the correctly predicted negative fraction,

Commission errors: the falsely predicted positive fraction,

Omission errors: the falsely predicted negative fraction.


Relative Operating Characteristic (ROC)

AUC = area under the curve

0.5-0.7 = poor model performance
0.7-0.9 = moderate
> 0.9 = excellent

AUC-value: what is the novelty?

Accuracy of bionamial models deals with ones and zeros, meaning that the class label is right or wrong. But many models are able to quantify their uncertainty about the answer by outputting a probability value. To compute accuracy from probabilities a threshold is needed to decide when zero turns into one. The most natural threshold is 0.5 (default option e.g. in SPSS).

Let’s suppose we have a model, which is able to get all the answers right, but it outputs 0.7 for negative examples and 0.9 for positive examples. Clearly, a threshold of 0.5 won’t get us far here. But 0.8 would be just perfect.

That’s the whole point of using AUC - it considers all possible thresholds. Various thresholds result in different true positive/false positive rates. As the threshold is decreased, more true positives are produced, but also more false positives.

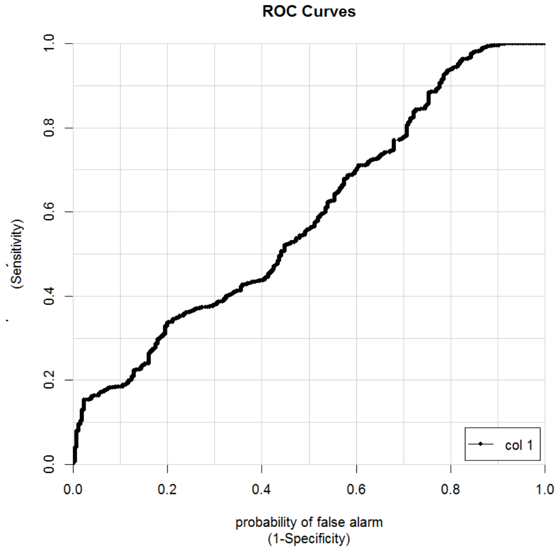
Interpretation of AUC: the probability that a randomly selected presence-observation has a higher probability value than a randomly chosen absence-observation.

AUC-value calculation in R

colAUC-function: calculates Area Under the ROC Curve (AUC) for every column of a matrix. Also, can be used to plot the ROC curves.

Example:

```
> require(caTools)
> betnan_glm <- glm(betnan ~ gdd, family="binomial")
> prediction_betnan <- predict.glm(betnan_glm, my.data,type="response")
> colAUC(prediction_betnan, betnan,plotROC=TRUE)
             [,1]
0 vs. 1 0.5869835
```



Interpretation of the model based on AUC:

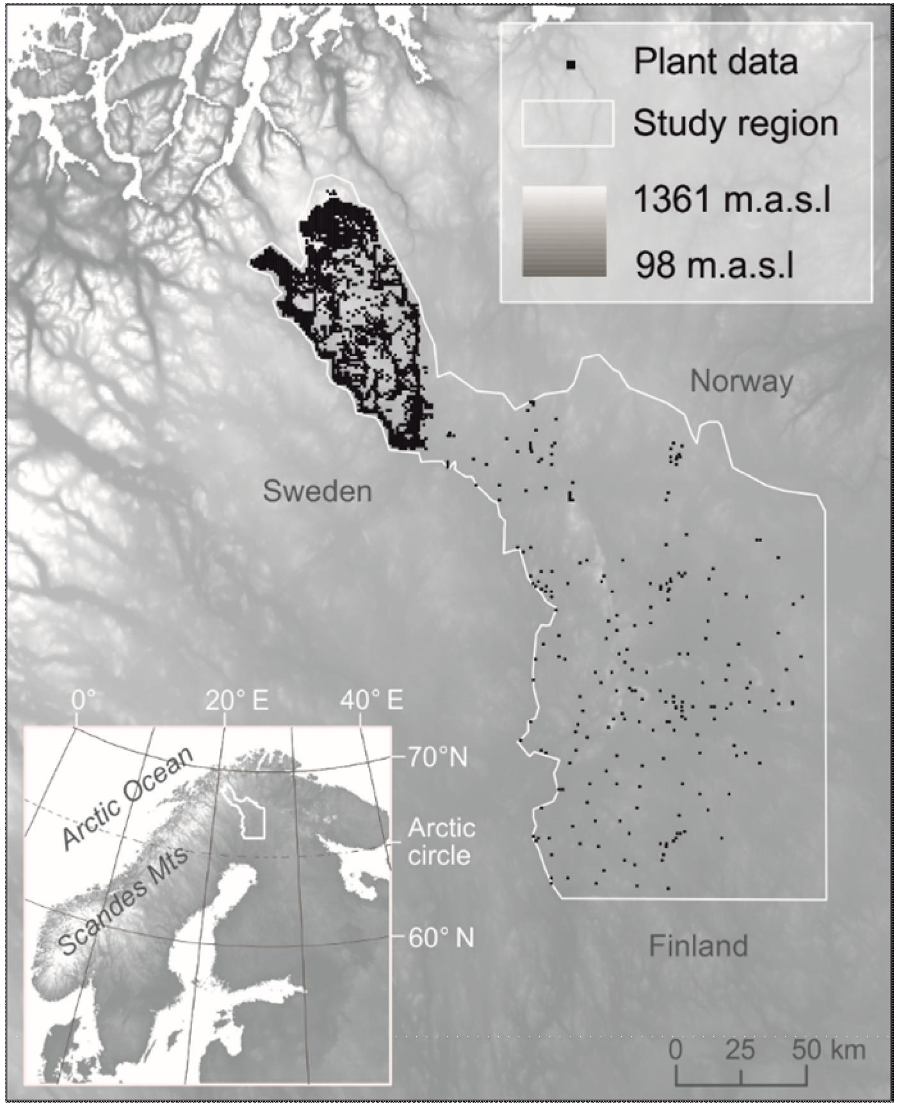
- the accuracy of the model is poor, betnan occurrence can not be satisfactory modelled based on gdd-variable only

Practical – modelling species richness and distributions in Northwestern Lapland with GLM

Data: ”*NW_Lapland_data.csv*”

- Occurrence data at a 1km x 1km resolution
- Variables:
  - xcoord = east coordinate
  - ycoord = north coordinate
  - totalspr = total vascular plant species richness
  - rarespr = rare vascular plant species richness
  - calc = index of calcareous rock type
  - fdd = freezing degree days; overwintering temp conditions
  - gdd = growing degree days; growing season temp conditions
  - wab = water balance; moisture conditions
  - altitude = elevation of the cell (m asl)
  - relalt = relative altitude of the cell in meters
  - *betnan*
  - *dryoct*
  - *empher*
  - *gersyl*
  - *linbor*
  - *phycae*
  - *rangla*
  - *vacmyr*
  - *vacvit*

occurrence of vascular plants (0/1) within a given cell



Questions

1. Based on our data (NW_Lapland_data.csv), what are the single most influential variables (gdd, fdd, wab, cal, altitude, relalt) explaining the variation in totspr in NW Finland? Are the effects (i) linear or curvilinear, (ii) positive or negative, (iii) what is the explained deviance of the models? P<0.05. Think carefully: family-argument and anova-test.
2. Based on our data (NW_Lapland_data.csv), what are the single most influential linear terms of variables (gdd, fdd, wab, cal, altitude, relalt) explaining the occurrence of gersyl, dryoct and empher in NN Finland? Are the effects (i) positive or negative; (ii) are there any differences between the modelling accuracy of the three species based on AUC-value? P<0.05, Chisq-test based on GLM. Think carefully: family-argument and anova-test.
3. What is the probability of gersyl presence at the altitudes of 100 m, 500 m and 1000 m based on a linear term?
4. Plot the response shapes of altitude and three plant species (gersyl, dryoct and empher) based on linear and quadratic terms in one scatterplot per species using different colors for linear and curvilinear responses. Describe the nature of the response shapes verbally. Are there any differences in the responses based on linear terms only and linear and quadratic terms? Note, in total three scatterplots, two response curves per scatterplot.