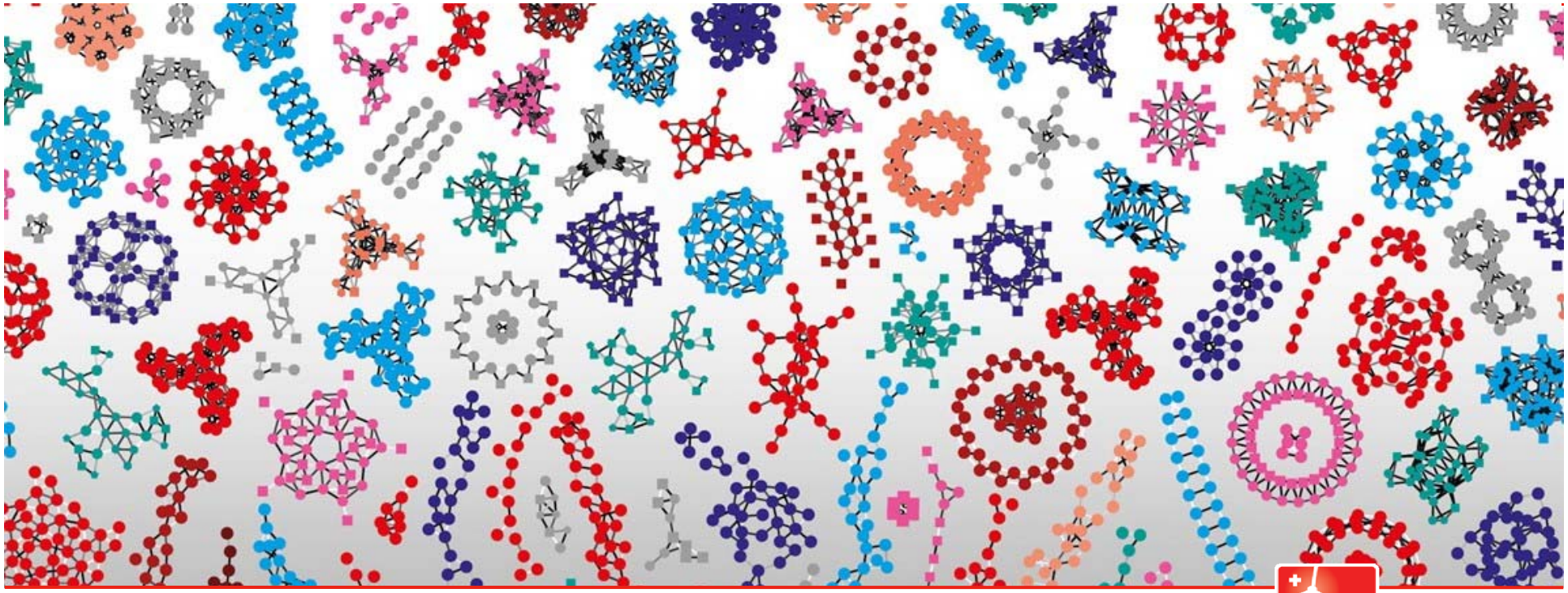


SIB Swiss Institute of Bioinformatics

BCF

Logistic Regression and GLM

Isabelle Dupanloup – 17th of August 2021
Slides modified from Linda Dib and Frédéric Schutz

Warm up

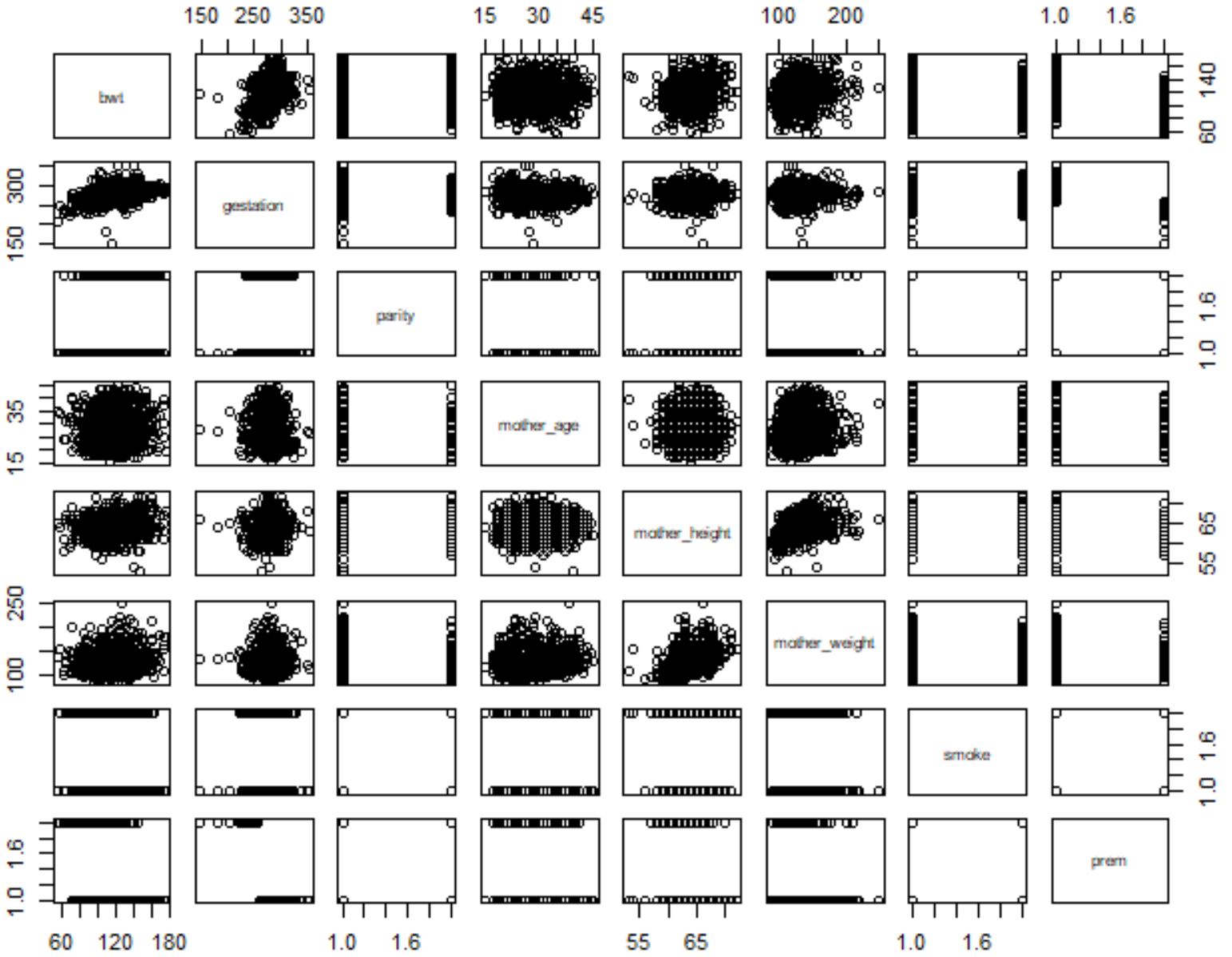
Load and explore the dataset babies.
load("exercises/babies.RData")

The data records the birth weight of 1174 babies along with information on the mother and the pregnancy.

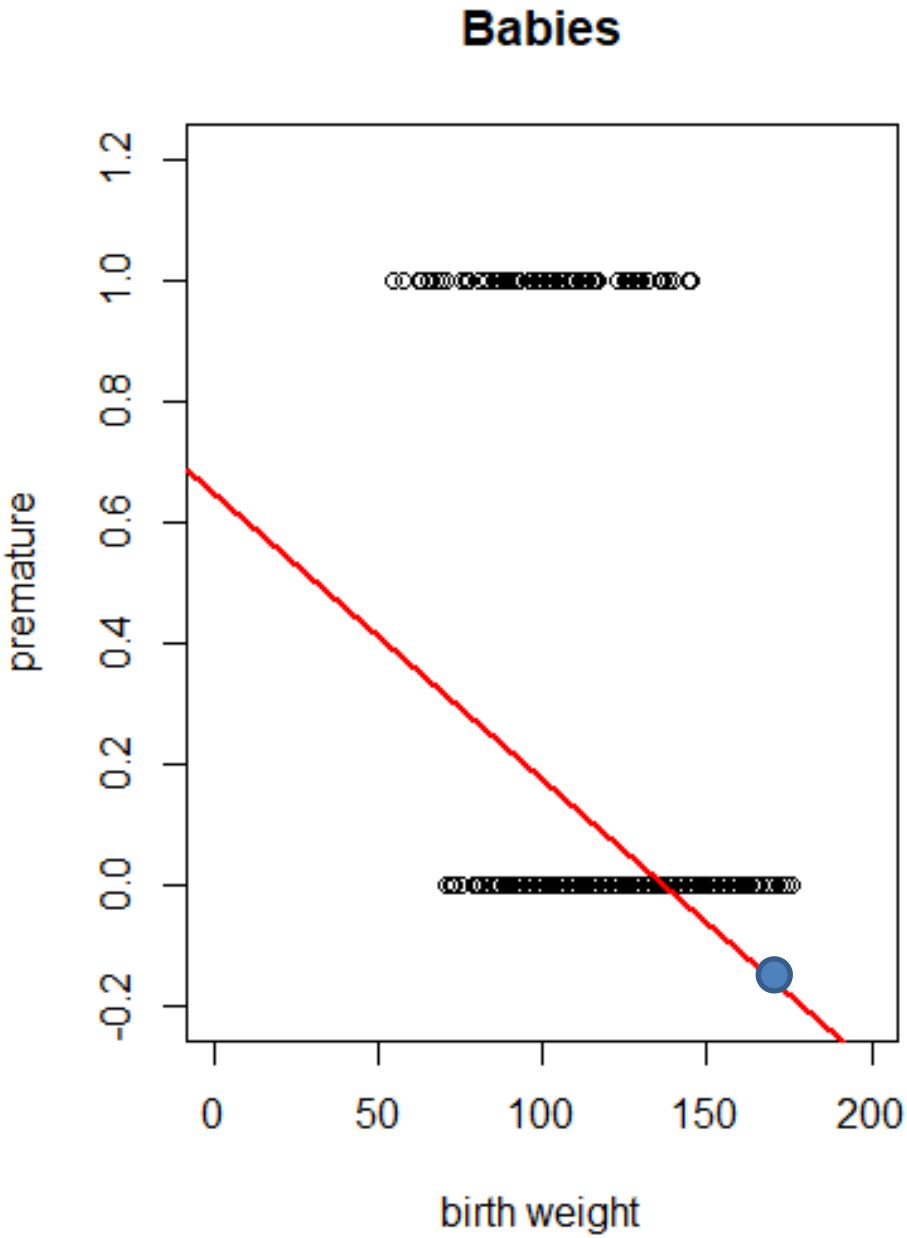
- Perform a graphical exploration of the data
- Which factor can explain prematurity?
- Can we use a linear model? If so, try to make predictions.

```
> summary(babies)
      bwt          gestation         parity      mother_age
 Min.   : 55.0   Min.   :148.0   first    :866   Min.   :15.00
 1st Qu.:108.0   1st Qu.:272.0   not first:308   1st Qu.:23.00
 Median :120.0   Median :280.0                   Median :26.00
 Mean   :119.5   Mean   :279.1                   Mean   :27.23
 3rd Qu.:131.0   3rd Qu.:288.0                   3rd Qu.:31.00
 Max.   :176.0   Max.   :353.0                   Max.   :45.00

 mother_height   mother_weight        smoke      prem
 Min.   :53.00   Min.   : 87.0   non-smoker:715   0:1078
 1st Qu.:62.00   1st Qu.:114.2   smoker    :459   1:  96
 Median :64.00   Median :125.0
 Mean   :64.05   Mean   :128.5
 3rd Qu.:66.00   3rd Qu.:139.0
 Max.   :72.00   Max.   :250.0
```

bwt: birth weight in ounces (1 ounce = 28.35 grams)
gestation: length of pregnancy in days
parity: first/not first
age: mother's age in years
height: mother's height in inches (1 inch = 2.54 cm)
weight: mother's pre-pregnancy weight
smoke: smoking status (smoker or non-smoker)
prem: prematurity indicator, ie, gestation shorter than 37 full weeks


bwt
gestation
parity
mother_age
mother_height
mother_weight
smoke
prem
150 250 350
15 30 45
100 200
1.0 1.6
60 140
150 300
1.0 1.6
15 35
55 65
100 250
1.0 1.6
60 120 180
1.0 1.6
55 65
1.0 1.6

Prediction if birth weight=170 ?



Statistical Models

Are used for explanation and prediction

Statistical models predicts the mean Y for any combination of predictors.

General form: g(Y) = f(X) + error

Y: dependent variable (response variable, outcome)

X: independent variable(s) (grouping variable, predictor)

Types of response and predictors variables

- binary (2 groups)

(e.g. yes/no, passed/failed, male/female, ...)

- categorical (k groups)

(e.g. phenotype, genotype, degree of smoking, ...)

- continuous (i.e. infinite number of groups)

(e.g. age, blood pressure, gene expression value, ...)

Types of variables

Response variable's type determines the regression method to use:

if continuous response	->	Linear regression
if binary response	->	Logistic regression
if count response	->	Poisson regression

Linear models

$$g(Y) = f(X) + \text{error}$$ **General Model**

$$Y = aX + b + \text{error}$$ **Linear Model**

Major assumptions in linear models:

- The *error* term has **zero mean** $(E[\epsilon_i] = 0)$
- The *error* term has **constant variance** $(Var[\epsilon_i] = \sigma_i)$
- The *errors* are **uncorrelated** $(Cov(\epsilon_i, \epsilon_j) = 0)$
- The *errors* are **normally distributed** $(\epsilon_i \sim N(\mu_i, \sigma_i))$

Types of variables

Response variable's type determines the regression method to use:

if continuous response	-> Linear regression
if binary response	**-> Logistic regression**
if count response	-> Poisson regression

What is Logistic Regression?

Form of regression that allows the prediction of discrete variables by a mix of continuous and discrete predictors.

Discrete ~ continuous/discrete

Gender ~ Height

Binary Logistic Regression Model

Y = Binary response, ex. Gender (male=1, female=0)

X = Quantitative predictor, ex. height

π = Proportion of success at any X

Proportion of "success"

Y = Binary response, ex. Gender (male=1, female=0)

In linear regression the model predicts the mean Y for any combination of prediction.

What's the mean of a 0/1 indicator variable?

$$\pi = \overline{y} = \frac{\sum y_i}{n}$$

Goal of logistic regression: Predict the "true" proportion of success, π, at any value of the predictor(s).

Binary Logistic Regression Model

Y = Binary response, ex. Gender (male=1, female=0)

X = Quantitative predictor, ex. Height

π = Proportion of success (1, male, yes, success) at any X

Logit form

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Background

Logit is the logarithm of the odds $\log\left(\frac{\pi}{1-\pi}\right)$

Probability of success

Probability of failure

π = 0.50, then logit = 0

π = 0.70, then logit = 0.84

π = 0.30, then logit = -0.84

Binary Logistic Regression Model

Y = Binary response, ex. Gender (male=1, female=0)

X = Quantitative predictor, ex. Height

π = Proportion of success (1, male, yes, success) at any X

Logit form

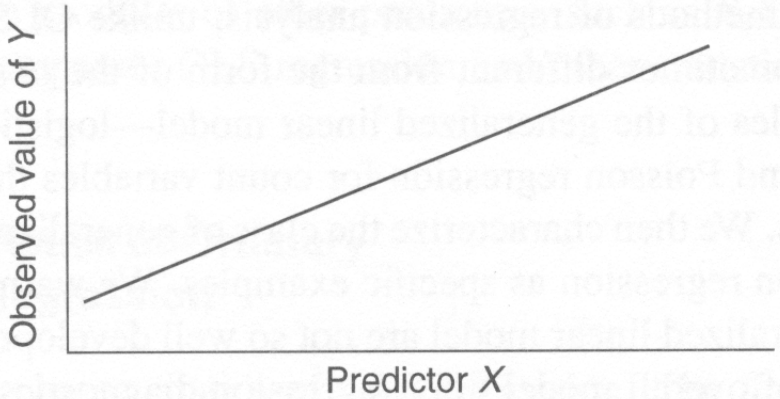
$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Probability form

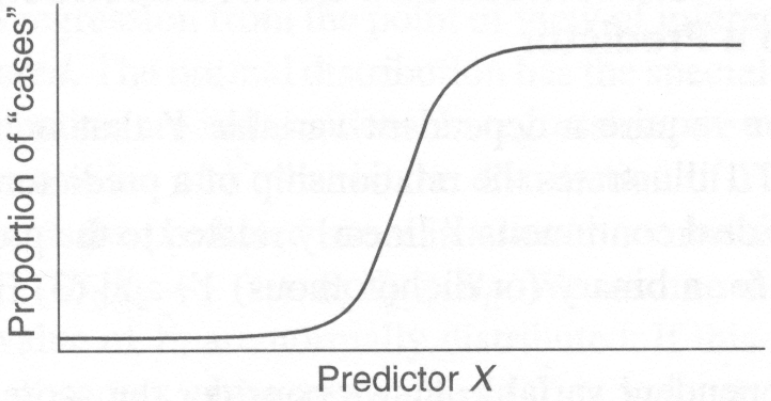
$$\pi = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

The logistic function

(A) For a continuous outcome variable Y, the numerical value of Y at each value of X.

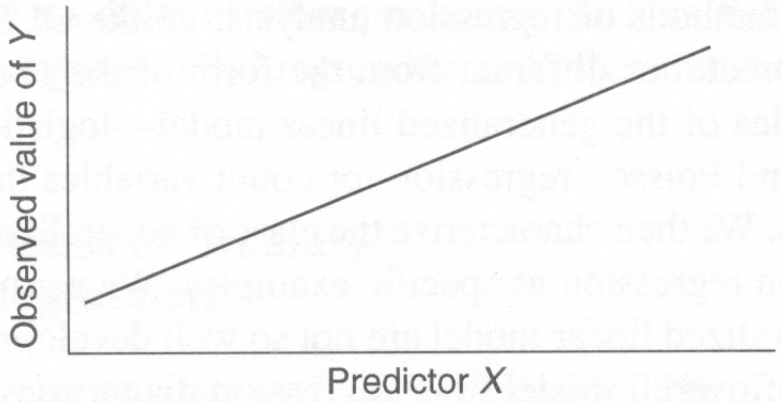


(B) For a binary outcome variable, the proportion of individuals who are "cases" (exhibit a particular outcome property) at each value of X.

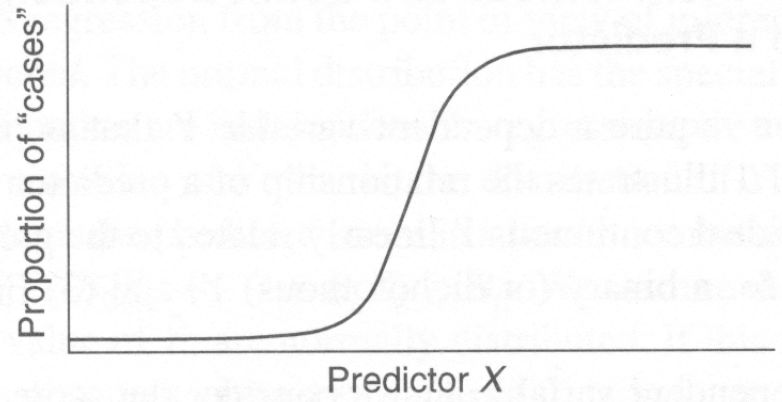


The logistic function

(A) For a continuous outcome variable Y, the numerical value of Y at each value of X.



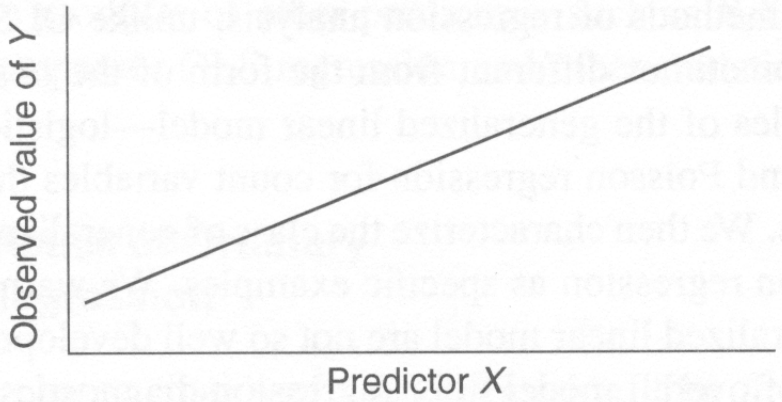
(B) For a binary outcome variable, the proportion of individuals who are "cases" (exhibit a particular outcome property) at each value of X.



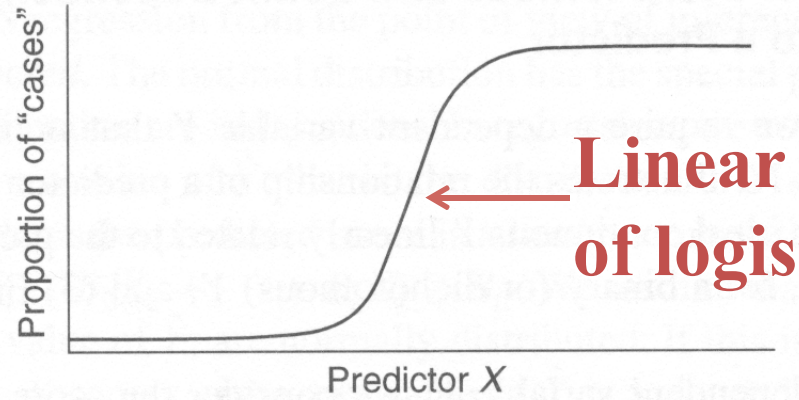
$$\pi = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

The logistic function

(A) For a continuous outcome variable Y, the numerical value of Y at each value of X.



(B) For a binary outcome variable, the proportion of individuals who are "cases" (exhibit a particular outcome property) at each value of X.



Linear part of logistic fit

Change in probability is not constant (linear) with constant changes in X

Assumptions

Linearity in the logit:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Absence of multicollinearity

No outliers

Generalized Linear Models

Ordinary Least Squares regression provides linear models of continuous variables. However, much data of interest to statisticians and researchers are not continuous and so other methods must be used to create useful predictive models.

The glm() command is designed to perform generalized linear models (regressions) on binary outcome data, count data, probability data, proportion data and many other data types.

Generalized Linear Models

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

glm(*formula*, **family=***familytype*(**link=***linkfunction*), **data=)**

Family	Default Link Function
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

Diabetes example

In R

```
# Load the data and remove NAs
> data("PimaIndiansDiabetes2", package = "mlbench")
> PimaIndiansDiabetes2 <- na.omit(PimaIndiansDiabetes2)

# Run model
> logitmodel_R <- glm( diabetes ~ glucose, data =
PimaIndiansDiabetes2, family = binomial)
> summary(logitmodel_R)
```

In R

```
Call:
glm(formula = diabetes ~ glucose, family = binomial, data =
PimaIndiansDiabetes2)


Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1728  -0.7475  -0.4789   0.7153   2.3860

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.095521   0.629787  -9.679   <2e-16 ***
glucose      0.042421   0.004761   8.911   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 498.10  on 391  degrees of freedom
Residual deviance: 386.67  on 390  degrees of freedom
AIC: 390.67


Number of Fisher Scoring iterations: 4
```

```
> summary( lm( Height ~ Age, data = class) )

Call:
lm(formula = Height ~ Age)

Residuals:
      Min        1Q    Median        3Q       Max
-12.59000  -3.57300  -0.07867   3.49000  15.57133

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   64.069     16.565   3.868  0.00124 **
Age            7.079      1.237   5.724 2.48e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.832 on 17 degrees of freedom
Multiple R-squared: 0.6584,     Adjusted R-squared: 0.6383
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

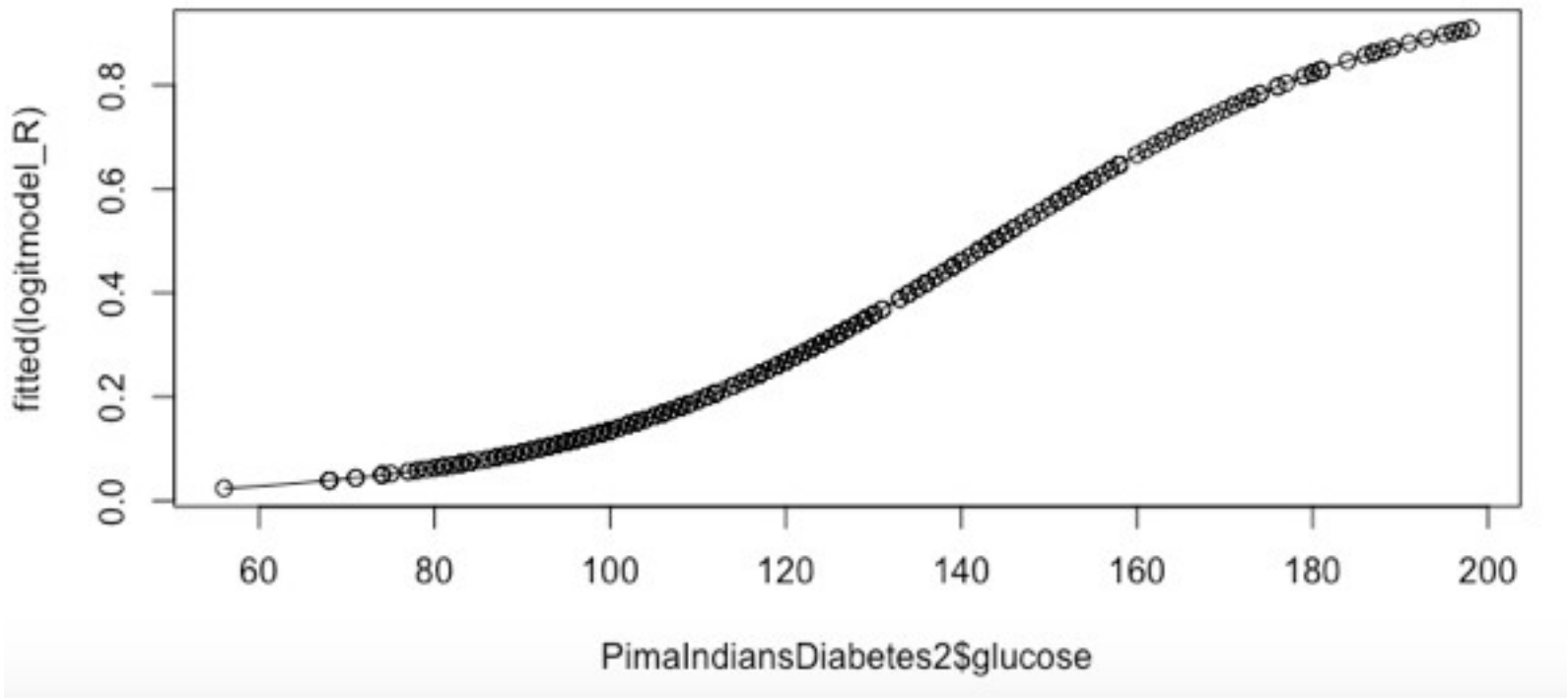
```
Call:
glm(formula = diabetes ~ glucose, family = binomial, data
= PimaIndiansDiabetes2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.095521   0.629787  -9.679   <2e-16 ***
glucose      0.042421   0.004761   8.911   <2e-16 ***
---
```

$$\pi = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Proportion of diabetic patients **at the estimate glucose level**

$$\hat{\pi} = \frac{e^{-6.09+0.04\,g}}{1 + e^{-6.09+0.04\,g}}$$

```
> plot(fitted(logitmodel_R)~ PimaIndiansDiabetes2$glucose)

> curve(exp(-6.0955+0.0424*x)/(1+exp(-6.0955+0.0424*x)),
add=TRUE)
```



Logistic regression, odds and odds ratios

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

The logistic model assumes a linear relationship between the *predictors* and *log(odds).*

$$odds = \frac{\pi}{1-\pi} = e^{\beta_0 + \beta_1 X}$$

$$odds = \frac{\pi}{1-\pi} \Leftrightarrow \pi = \frac{odds}{1+odds}$$

Example: TMS for Migraines

Transcranial Magnetic Stimulation vs. Placebo

Pain Free?	TMS	Placebo
YES	39	22
NO	61	78
Total	**100**	**100**

$$odds_{TMS} = \frac{39/100}{61/100} = \frac{39}{61} = 0.639 \qquad \hat{\pi} = \frac{0.639}{1+0.639} = 0.39$$

$$odds_{Placebo} = \frac{22}{78} = 0.282 \qquad \hat{\pi}_{Placebo} = 0.22$$

$$Odds\ ratio = \frac{0.639}{0.282} = 2.27$$

Odds are 2.27 times higher of getting relief using TMS than placebo

Logistic regression, odds and odds ratios

Odds for X: $$odds = e^{\beta_0 + \beta_1 X}$$

Odds for X+1: $$odds = e^{\beta_0 + \beta_1 (X+1)}$$

Odds ratio (odds for X / odds for X+1):

$$\frac{e^{\beta_0 + \beta_1 (X+1)}}{e^{\beta_0 + \beta_1 X}} = e^{\beta_0 + \beta_1 (X+1) - (\beta_0 + \beta_1 X)} = e^{\beta_1}$$

In R

```
Call:
glm(formula = diabetes ~ glucose, family = binomial, data =
PimaIndiansDiabetes2)
```

Note: $e^{0.042421} = 1.043334$ = odds ratio

```
Deviance Residuals:
    Min       1Q   Median                   Max
-2.1728  -0.7475  -0.4789   0                2.3860

Coefficients:
             Estimate                  z value Pr(>|z|)
(Intercept) -6.095521                  -9.679   <2e-16 ***
glucose      0.042421                   8.911   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 498.10  on 391  degrees of freedom
Residual deviance: 386.67  on 390  degrees of freedom
AIC: 390.67


Number of Fisher Scoring iterations: 4
```

Multiple logistic regression

Multiple Logistic Regression

Extension to more than one predictor variable (either numeric or dummy variables).

With *k* predictors, the model is written:

$$\pi = \frac{e^{\beta_0 + \beta_1 x_1 + .. + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + .. + \beta_k x_k}}$$

Adjusted Odds ratio for raising x_i by 1 unit, holding all other predictors constant:

$$OR_i = e^{\beta_i}$$

Challenge

Using the babies dataset

- Fit a logistic regression to find parameters explaining the probability of prematurity ?
- What is the effect of birth weight on the probability of prematurity ?
- What about parity ?

Solution

```
> model2 <- glm(prem ~ bwt, family=binomial)
> summary(model2)

Call:
glm(formula = prem ~ bwt, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2879  -0.3985  -0.2784  -0.1810   3.0710

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.017338   0.717952   6.988 2.78e-12 ***
bwt         -0.067061   0.006808  -9.851  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 664.66  on 1173  degrees of freedom
Residual deviance: 545.31  on 1172  degrees of freedom
AIC: 549.31

Number of Fisher Scoring iterations: 6
```

Solution

```
> model3 <- glm(prem ~ bwt + parity, family = binomial)
> summary(model3)

Call:
glm(formula = prem ~ bwt + parity, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3375  -0.4074  -0.2758  -0.1795   3.0340

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     5.129006   0.722464   7.099 1.25e-12 ***
bwt            -0.067046   0.006806  -9.850  < 2e-16 ***
paritynot first -0.465924   0.281371  -1.656   0.0977 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 664.66  on 1173  degrees of freedom
Residual deviance: 542.39  on 1171  degrees of freedom
AIC: 548.39

Number of Fisher Scoring iterations: 6
```

Solution

```
> model4 <- glm(prem ~ bwt*smoke+parity, family=binomial)
> summary(model4)

Call:
glm(formula = prem ~ bwt * smoke + parity, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4798  -0.3998  -0.2784  -0.1682   2.9571

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.839354   0.978834   4.944 7.65e-07 ***
bwt              -0.062082   0.008741  -7.103 1.22e-12 ***
smokesmoker       2.247047   1.609071   1.396   0.1626
paritynot first -0.470085   0.283836  -1.656   0.0977 .
bwt:smokesmoker -0.028043   0.015781  -1.777   0.0756 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 664.66  on 1173  degrees of freedom
Residual deviance: 532.93  on 1169  degrees of freedom
AIC: 542.93

Number of Fisher Scoring iterations: 6
```

Solution

- bwt is the only significant factor
- increasing the birth weight has the effect of decreasing the probability of prematurity

Challenge: baby food

The data for this exercise study infant respiratory disease, namely the proportions of children developing bronchitis or pneumonia in their first year of life by type of feeding, and sex. Data may be found in Payne (1987) and Faraway (2006)

```
library(faraway)
data(babyfood)
```

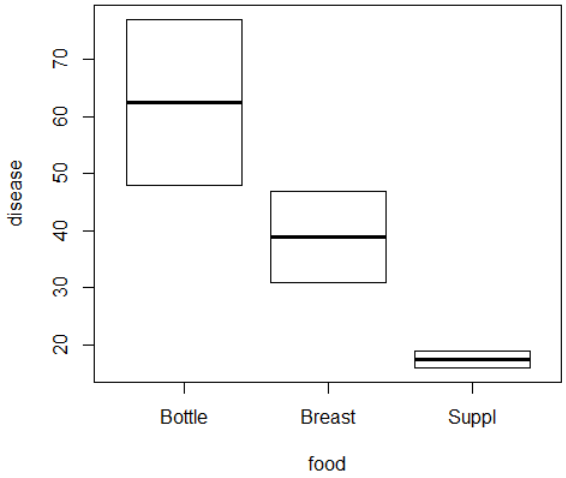
1. Explore the data
2. Fit a logistic regression to explain the probability of disease by sex and food.

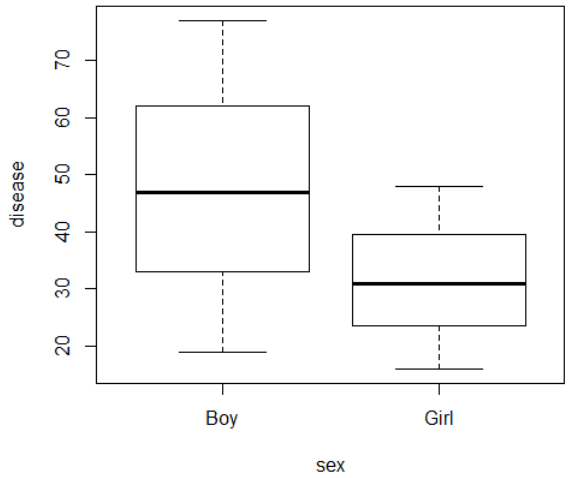
Challenge: baby food: solution

1. summary(babyfood)
boxplot(disease ~ food, babyfood)
boxplot(disease ~ sex, babyfood)

	disease	nondisease	sex	food
1	77	381	Boy	Bottle
2	19	128	Boy	Suppl
3	47	447	Boy	Breast
4	48	336	Girl	Bottle
5	16	111	Girl	Suppl
6	31	433	Girl	Breast

```
> summary(babyfood)
    disease        nondisease      sex        food
 Min.   :16.00   Min.   :111.0   Boy :3   Bottle:2
 1st Qu.:22.00   1st Qu.:180.0   Girl:3   Breast:2
 Median :39.00   Median :358.5            Suppl :2
 Mean   :39.67   Mean   :306.0
 3rd Qu.:47.75   3rd Qu.:420.0
 Max.   :77.00   Max.   :447.0
```





Challenge: baby food: solution

2. mdl <- glm(cbind(disease, nondisease) ~ sex + food, family = binomial, babyfood)
summary(mdl)

```
Call:
glm(formula = cbind(disease, nondisease) ~ sex + food, family = binomial,
    data = babyfood)

Deviance Residuals:
      1       2       3       4       5       6
 0.1096  -0.5052  0.1922  -0.1342  0.5896  -0.2284

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6127     0.1124 -14.347  < 2e-16 ***
sexGirl      -0.3126     0.1410  -2.216   0.0267 *
foodBreast   -0.6693     0.1530  -4.374 1.22e-05 ***
foodSuppl    -0.1725     0.2056  -0.839   0.4013
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26.37529  on 5  degrees of freedom
Residual deviance:  0.72192  on 2  degrees of freedom
AIC: 40.24

Number of Fisher Scoring iterations: 4
```

Logistic regression

is a special case of

General Linear Model

General Linear Model

- GLM is a generalization of linear model
- LM and Logistic regression are special cases of GLM

General Linear Model

$$\underbrace{E(Y)}_{Link\ function} = \underbrace{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}_{Systematic} \underbrace{+ e}_{Random}$$

Three components

Y: dependent variable (response variable, outcome)

X_i: independent variable(s) (grouping variable, predictor)

Random (stochastic) component

In GLM the random component is not restricted to a normal or Gaussian distribution

Random component defines the exponential distribution (Gaussian, Poisson, binomial, gamma, and inverse Gaussian distributions) from which the responses are assumed to be drawn.

$$\underbrace{E(Y)}_{Link\ function} = \underbrace{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}_{Systematic} \underbrace{+ e}_{Random}$$

Random (stochastic) component

- Continuous outcomes - Random component has *Normal* distribution – model = **Linear models**
- Binary outcomes (i.e. success or failure) - Random component has *Binomial* distribution – model = **Logistic Regression**
- Count data (i.e. number of events in fixed duration of time) - Random component has *Poisson* distribution – model = **Poisson Regression**
  - generalization of Poisson regression (variance ≠ mean) – model = **Negative Binomial Regression**

Systematic component

In GLM the systematic component defines the linear combinations of predictors

$$\underbrace{E(Y)}_{Link\ function} = \underbrace{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}_{Systematic} \underbrace{+ e}_{Random}$$

Link function

Transforms the mean of the response variable such that it has a linear relationship with the independent variables covariates

$$\underbrace{E(Y)}_{Link\ function} = \underbrace{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}_{Systematic} \underbrace{+ e}_{Random}$$

Linear models

$$E(Y) = Y$$

Logistic regression

$$E(Y) = \log\left(\frac{\pi}{1-\pi}\right)$$

Poisson regression

$$E(Y) = \log(\lambda)$$

GLM and diagnostic

- Deviance residuals vs fitted (linear and constant variance)
- Residuals vs covariates (multicolinearity)
- Uniform residuals checks (quantile residuals)
- Detection of influential observations

Deviance

- In standard linear models, we estimate the parameters by minimizing the sum of the squared residuals. Equivalent to finding parameters that maximize the likelihood.
- Deviance is a measure of goodness of fit of a generalized linear model. Estimation is equivalent to finding parameter values that minimize the deviance.
- 2 forms of deviance
  - ❑ Null deviance: how well the response variable is predicted by a model that includes only the intercept (grand mean)
  - ❑ Residual deviance: how deviance is reduced by including the independent variables

Akaike Information Criterion (AIC)

- allows to assess the quality of a model through comparison of related models
- based on the Deviance, but penalizes more complicated model (much like adjusted R-squared, it's intent is to prevent including irrelevant predictors)
- unlike adjusted R-squared, the number itself is not meaningful: always select the model that has the smallest AIC !

Residuals

- Residuals = difference between the data and the model
- GLM : no assumption on constant variance or normality
- Residuals can be used to spot influencial observations
- Standardized residuals should be close to a normal distribution with same variance -> use quantile residuals (~N(0;1)) for logistic regression !

- Points with large residuals should be checked !!!

Hat function

High leverage ('influential') points are far from the center, and have potentially greater influence

One way to identify these points is through the *hat values* (obtained from the *hat matrix H*):

h_{ij}: contribution of the ith observation to the jth fitted value

h_i: contribution of the ith observation to the fitted values

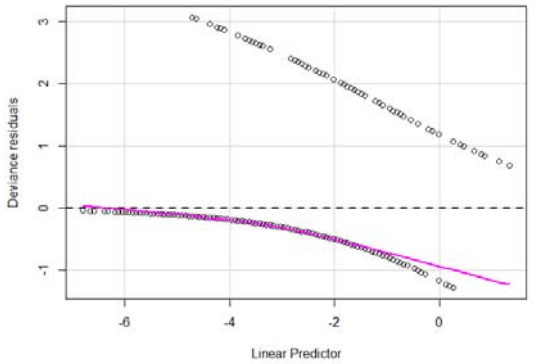
Challenge

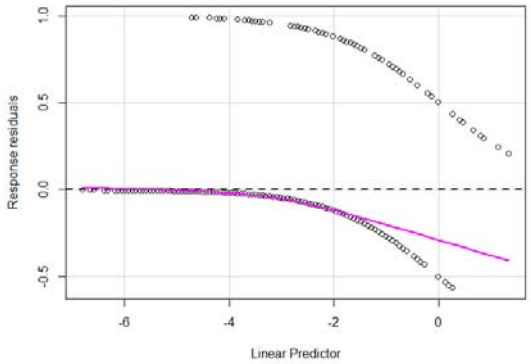
In the baby (premature) dataset, is the logistic regression you fitted appropriate for the data?

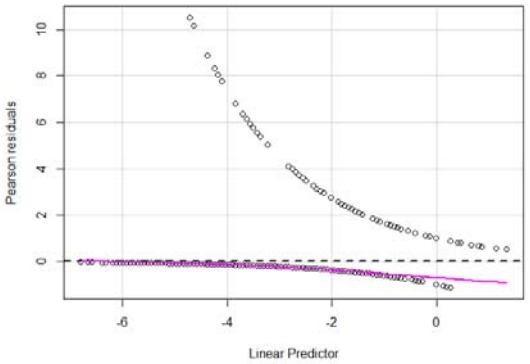
- Check the deviance residuals using the residualPlot function in the car library
- Construct the quantile residuals using the qresiduals function in the statmod library
- Analyze the deviance
- Look for potential influencial and outlying observations

Residuals

```
> library(car)
> residualPlot(model2, type = "deviance")
> residualPlot(model2, type = "response")
> residualPlot(model2, type = "pearson")
```







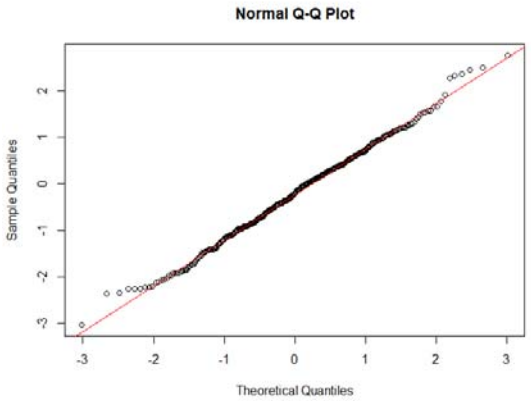
Residuals

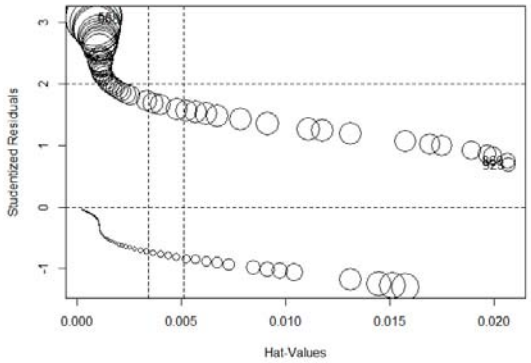
```
> library(statmod)
> model2.residuals <- qresiduals(model2)
> qqnorm(qnorm(model2.residuals))
> qqline(qnorm(model2.residuals), col="red")
```

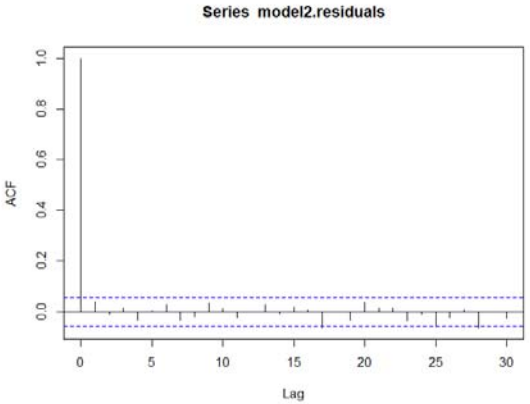
```
> model.null <- glm(prem ~ 1, family = binomial)
> anova(model.null, model2, test = "Chisq")
Analysis of Deviance Table

Model 1: prem ~ 1
Model 2: prem ~ bwt
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1173     664.66
2      1172     545.31  1   119.36 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> influencePlot(model2)
      StudRes          Hat       CookD
55  3.0854913 0.0008055876 0.044647682
860 0.7534952 0.0206397301 0.003483327
923 0.6894463 0.0206805355 0.002854404
968 3.0632647 0.0008248950 0.042754055
```

```
> influencePlot(model2)
> acf(model2.residuals)
```







Poisson Regression
for
count data

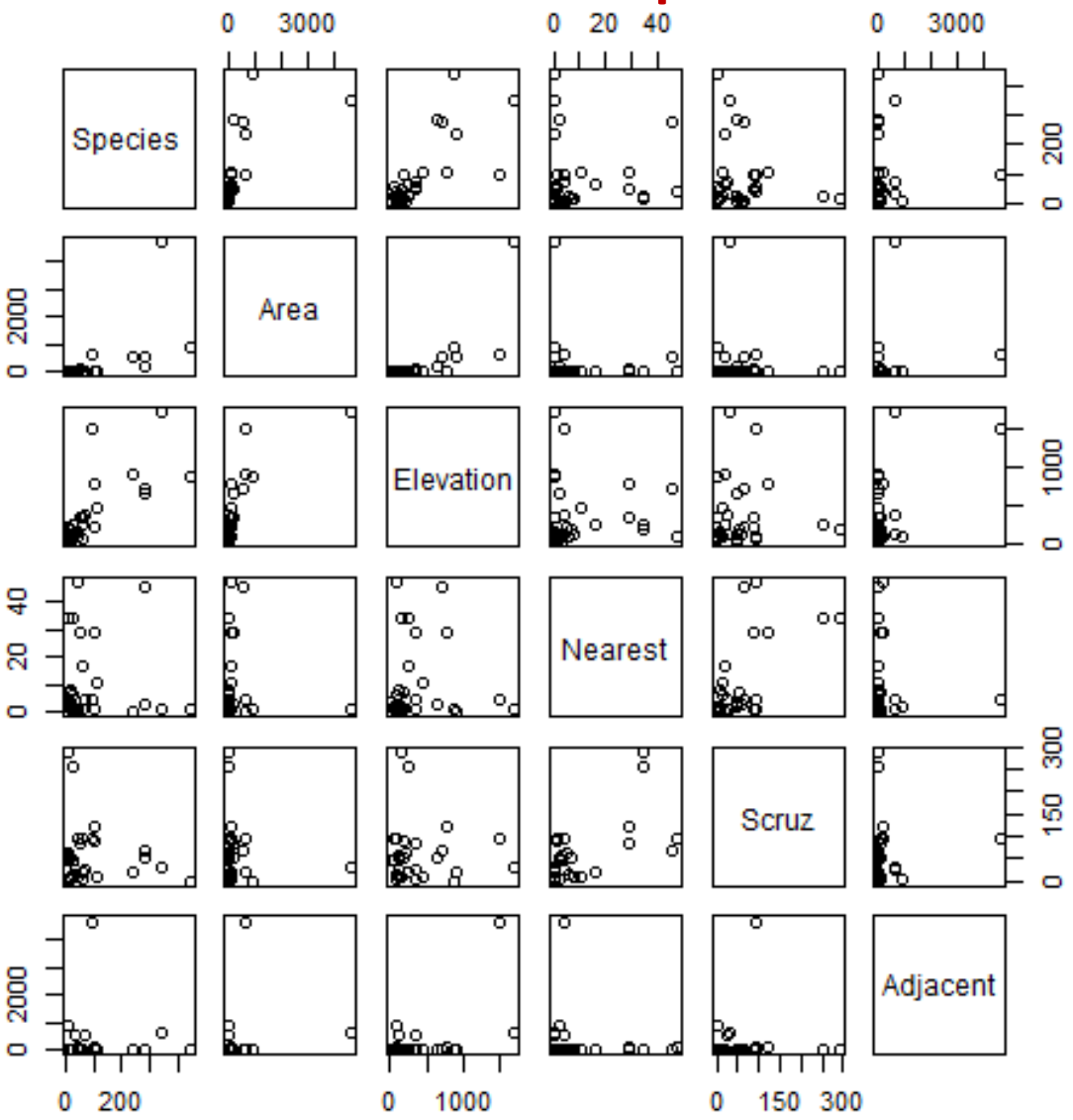
Warmup

Explore the dataset gala in library faraway. Remove the variable "endemics" which we will not use here.

```
> library(faraway)
> data(gala)
> gala <- gala[,-2]
```

- Study the relationship between the number of plant species and several geographical variables of interest.

Warmup


Species
Area
Elevation
Nearest
Scruz
Adjacent
0 3000
0 20 40
0 3000
0 2000
0 200
0 1000
0 20 40
0 150 300
0 2000
0 200
0 1000
0 150 300

Poisson regression

- Generally used to model Count data
- Distribution: Poisson
  (Restriction: mean = variance : E(Y)=V(Y)=λ
- Link Function: log link:

$$\ln(\lambda) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

$$\Rightarrow \lambda(X_1, \ldots, X_k) = e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}$$

Tests are conducted as in Logistic regression

When the mean and variance are not equal (over-dispersion), often replace the Poisson Distribution replaced with Negative Binomial Distribution

Poisson regression - assumptions

- Poisson Response: the response variable is a count per unit of time or space, described by a Poisson distribution.
- Independence: the observations must be independent of one another.
- Mean=Variance: by definition, the mean of a Poisson random variable must be equal to its variance.
- Linearity: he log of the mean rate, log(λ), must be a linear function of x.

Challenge

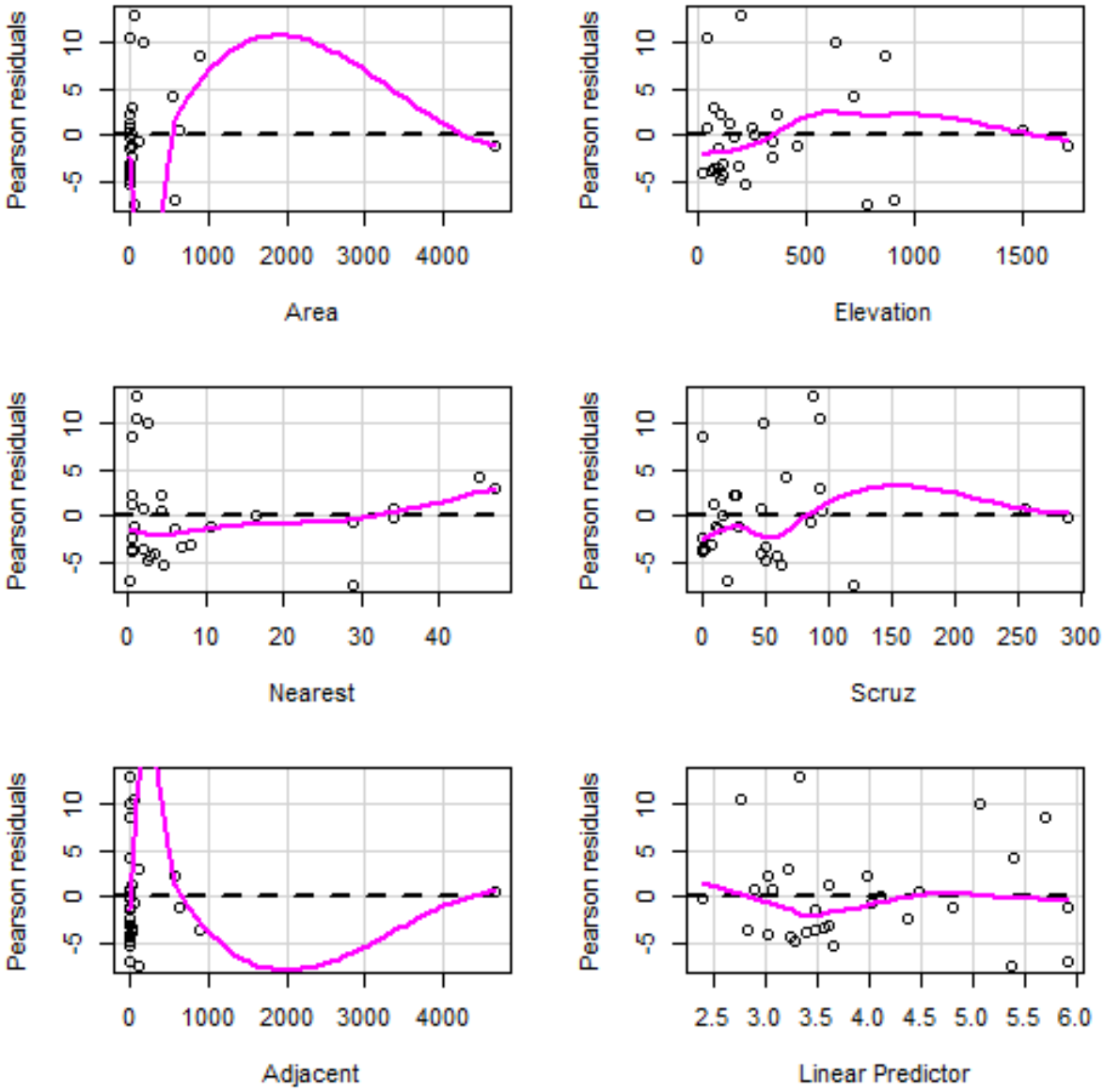
Using the glm function with family=poisson,

- Fit a poisson model to the galapagos data.
- Which variables are significant ?
- Check the deviance of the model

Solution

```
> poisson.glm <- glm(Species ~., data=gala, family=poisson)
> summary(poisson.glm)

Call:
glm(formula = Species ~ ., family = poisson, data = gala)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2752  -4.4966  -0.9443   1.9168  10.1849

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 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: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68

Number of Fisher Scoring iterations: 5
```

Solution


Pearson residuals
-5 0 5 10
0 1000 2000 3000 4000
Area
Pearson residuals
-5 0 5 10
0 500 1000 1500
Elevation
Pearson residuals
-5 0 5 10
0 10 20 30 40
Nearest
Pearson residuals
-5 0 5 10
0 50 100 150 200 250 300
Scruz
Pearson residuals
-5 0 5 10
0 1000 2000 3000 4000
Adjacent
Pearson residuals
-5 0 5 10
2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0
Linear Predictor

Summary

- When the response cannot be explained linearly by the predictors, transform the mean of the response so that is is linear

$$Y = \beta_0 + \beta_1 X$$

Logistic Regression:

- Response is a binary data or a probability of success
- Logit form:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Poisson regression

- Use for count data
- Poisson assumes mean=variance
- Logit form:

$$\ln(\lambda) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$