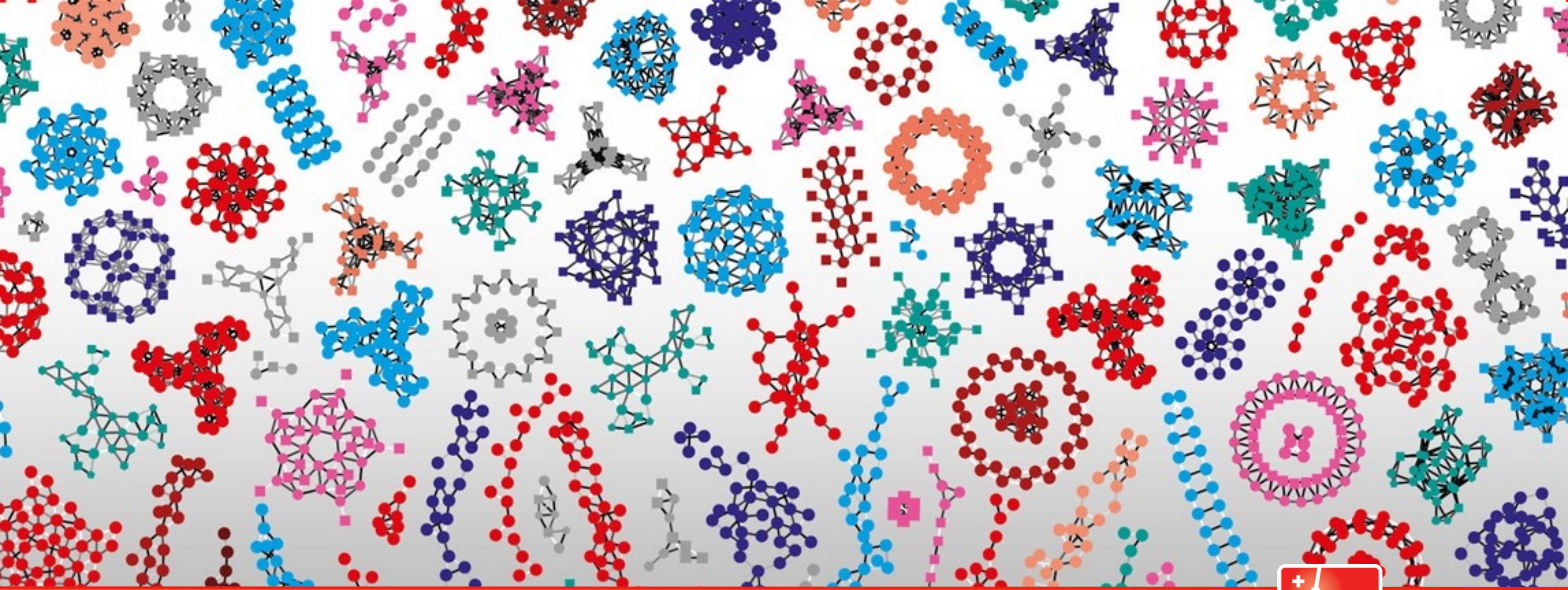


Logistic Regression and GLM

Rachel Marcone (Jeitziner) and Mauro Delorenzi

Slides credit also to Linda Dib, Frédéric Schütz, Isabelle Dupanloup, …

www.sib.swiss

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)

With a stochastic process at some level

Y: dependent variable (response variable, observed outcome)

X: independent/ explanatory 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 that is best adapted:

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

Example: Gender ~ Height

Binary Logistic Regression Model

Y = Binary response, ex. Gender (male=1, female=0)
X = Quantitative predictor, ex. height

π = Proportion / **Probability** of »event 1» at any X

Given π we assume a stochastic process to determine the events observed (numbers of females and males)

Here a **binomial** distribution B(n,p) with

n = number of observations at this X,
p = prob. of event 1 and

Proportion of “success”

In linear regression the model predicts the mean Y for any combination of prediction (the E [Y | X]) resp. E [P(Y=1) | X]) .

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

The ***Proportion*** *of “cases 1” among n observations.*

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

Goal of logistic regression: Predict the ***“true” probability*** *of success, π, at any value of the predictor(s).*

Logistic regression, odds and odds ratios

Relation probability – odds

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

π in [0 , 1] , odds in (- ∞, + ∞),

π = 0.5 odds = 1

π = 0.9 odds = 9

π = 0.1 odds = 1/9 = 0.111

Logistic curve

Logit is the **logarithm of the odds**

Probability of success

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

Probability of failure

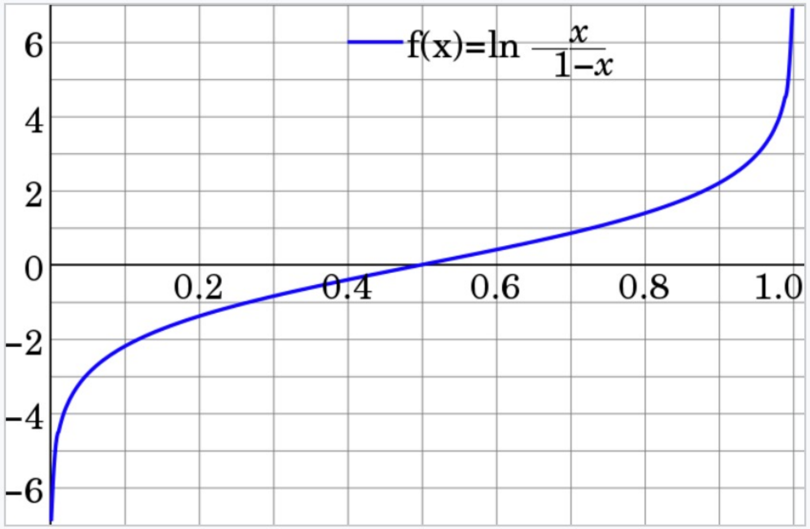
π = 0.50, then logit = 0

π = 0.70, then logit = 0.84

π = 0.30, then logit = -0.84

π -> 1, then logit -> inf

π -> 0, then logit -> - inf



Plot of logit(*p*) in the domain of 0 to 1, where the base of logarithm is *e*.

https://en.wikipedia.org/wiki/Logit

Binary Logistic Regression Model

π = Proportion of success , at any X

Logit : **log**arithm of the odds (log = ln)

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

The predictors acts at the level of the log odds

The probability p than is derived from the log odds.

Binary Logistic Regression Model

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

The logit is called a **link function**,

it links the level of the observed events (**response level**)
to
the level at which the predictors effects are acting (**link level**)

Binary Logistic Regression Model

Y = Binary response

X = Quantitative predictor

π = Proportion of success

Logit form

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

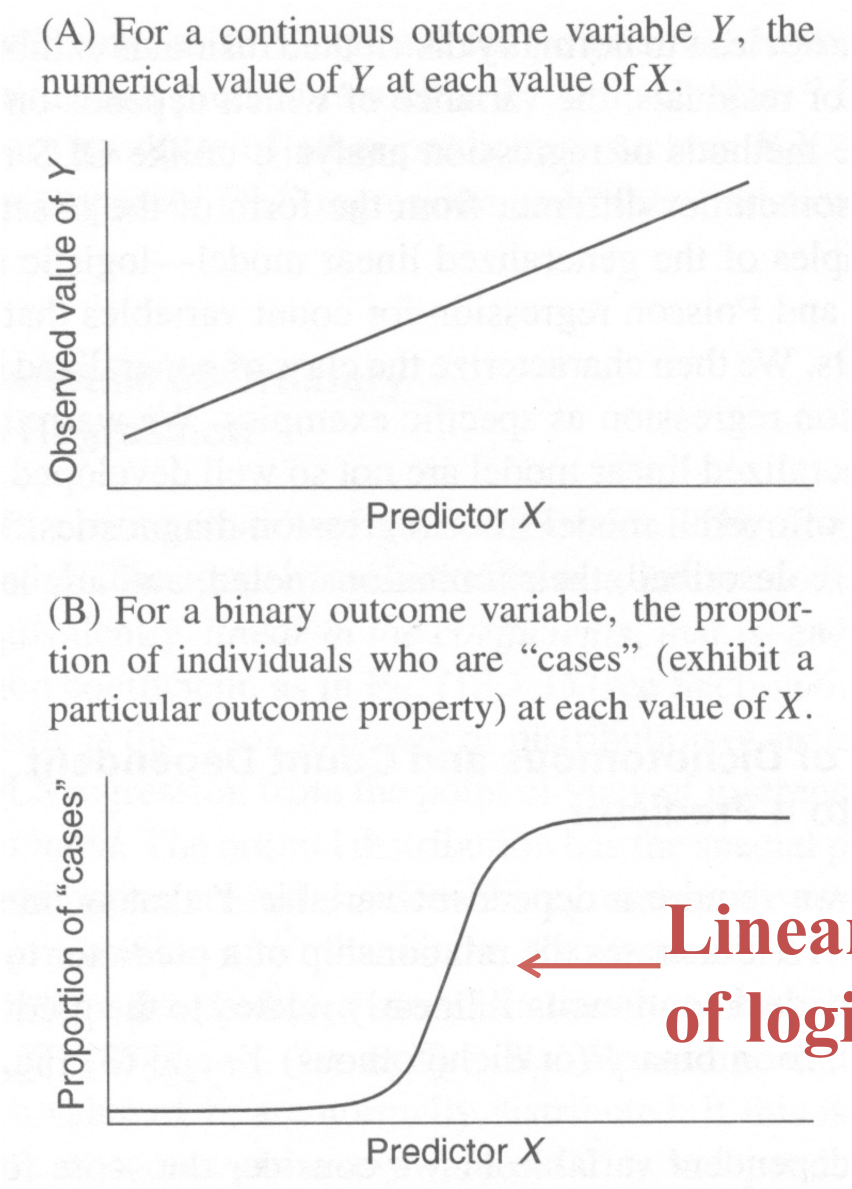
Link - Level

Probability form

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

Response - Level

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

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

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+1 / odds for X):

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

We increase X_1 by one unit **(+1**, **additive**)

The log odds is increased by **ß$_1$** (additive)

The odds is increased by a factor exp(**ß$_1$**) (**multiplicative**)

The probaility is increased by ? (**question!**)

Assumptions

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

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

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

Logistic regression

is a special case of

Generalized Linear Model GLM

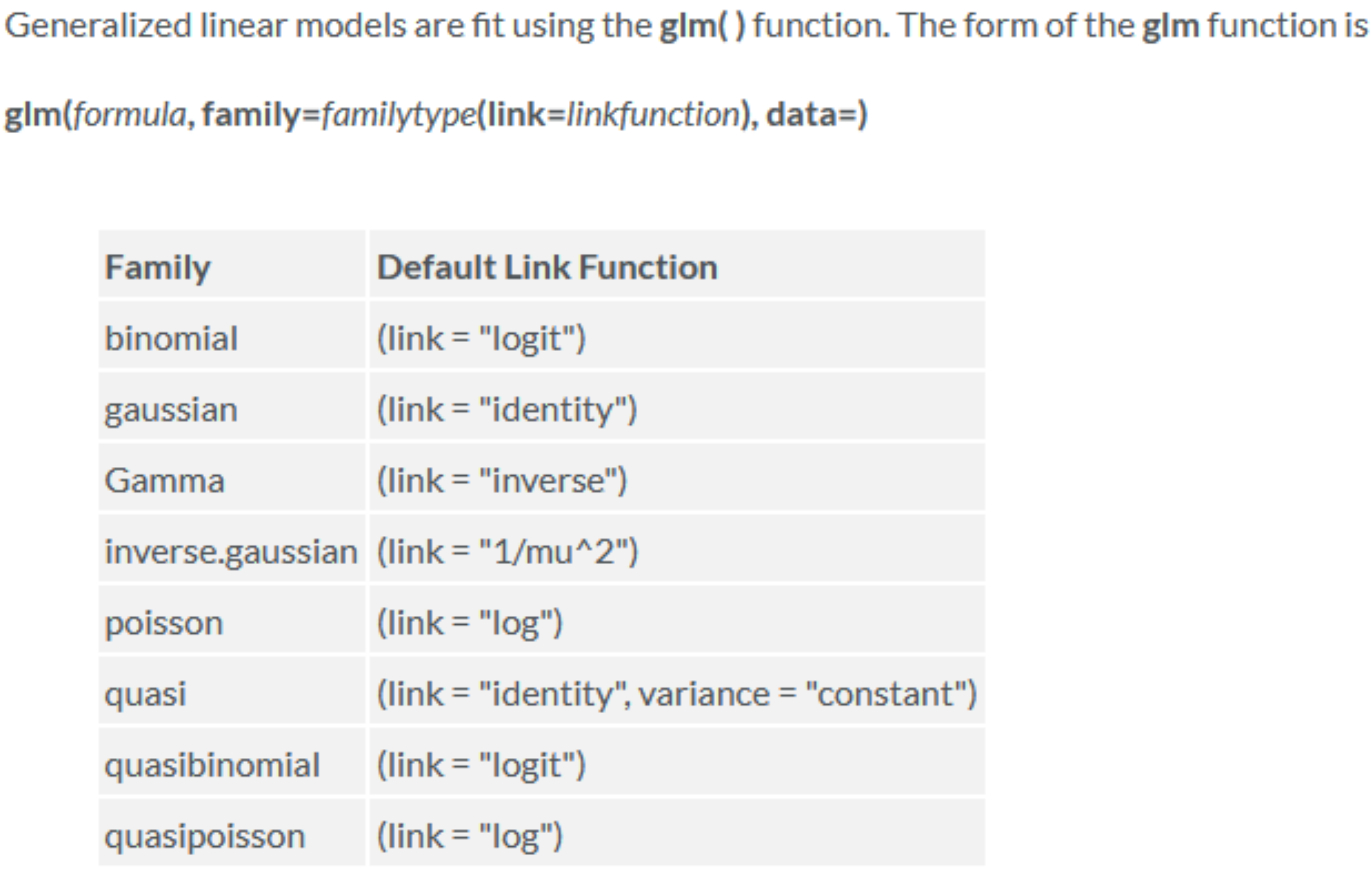
Generalized Linear Models

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.

Generalized Linear 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")

GLM LOGISTIC

Predictors X => E(logit(π) | X) => observations Y_i

=> The assumed model of effects

=> The assumed underlying stochastic process (generating the data)

$\Rightarrow$ ex. Logit is linear in ß's

=> ex. Binomial distribution

GLM POISSON

Predictors X => E(log(mean) | X) => observations Y_i

=> The assumed model of effects

=> The assumed underlying stochastic process (generating the data)

=> ex. Poisson

How to find the »best fit»

Standard method: **maximum likelihood estimation MLE**

Determine parameters so as to :
Probability of observations = maximum
Resp. Called Likelihood of the model Lik (model | data) = MAX

Usually at log level Log (Lik) = MAX

How to find the »best fit»

Standard method: **maximum likelihood estimation MLE**

Log (Lik) = MAX

Where the probability of observations (given parameters), and thus the Likelihood, is given by the **=> process** and the link to the parameters by the **=> model**

How to find the »best fit»

The **maximum likelihood estimation MLE** is in many problems the preferred method

There are several theorems that show how the MLE principle has a series of «desired properties» and in some sense it is the most powerful method for estimation and for statistical testing.

The t-test for example is the maximum likelihood test to compare the mean of two normal distributions.

MLE examples

1) What is the best 1-value model of the center of a normal distribution ?

MLE estimator = mean

(under assumptions of i.i.d)

MLE examples

2) LM Models:
What is the best regression line
fit for a set of Y points given a predictor X ?

MLE estimator = least-square estimator
(the stochastic process is normal that is the real residuals from the real model are i.i.d normal)

The LM is the GLM with the identity as link function
(that is no link function) and with the Gaussian normal distribution as the stochastic process

Major assumptions in linear models:

- (approximate) **linear relationship** between *outcomes* and *predictors*
- 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** $(\epsilon_i \sim N(\mu_i, \sigma_i))|$

How to find the »best fit»

Standard method: **maximum likelihood estimation MLE**

Solution:

Generally there is no closed solution (formula) for the parameters in function of the data

How to find the »best fit»

Standard method: **maximum likelihood estimation MLE**

The point estimated are determined by multi-step iterative algorithms that improves the solution until it is «good enough»

The standard errors of the estimates are than derived (approximatively) , also an (analogon of the) hat matrix and various types of residuals

How to test »significance» and determine CI ?

Given standard errors SE of ß :

test-statistics = estimate / SE = z

approx. Normal (under the null hypothesis)

called a **Wald-test**

CI width = approx. 1.96 * SE

CI symmetric for ß and the log odds scale

⇒ not symmetric for the multiplicative effect exp(ß) on the odds scale

⇒ not symmetric for the effect on the probability π

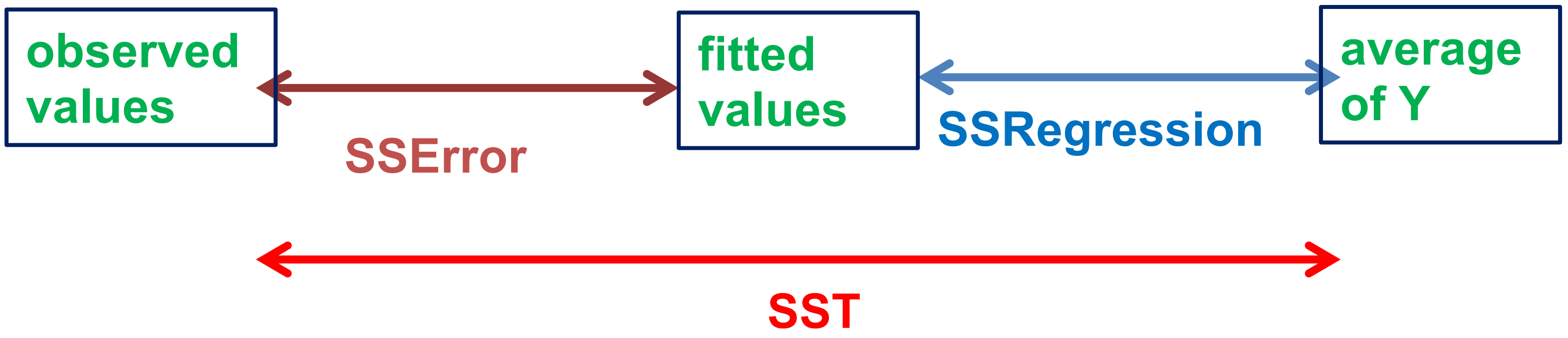
SST = SSR + SSE

Total sum of squares = regression SS + residual SS

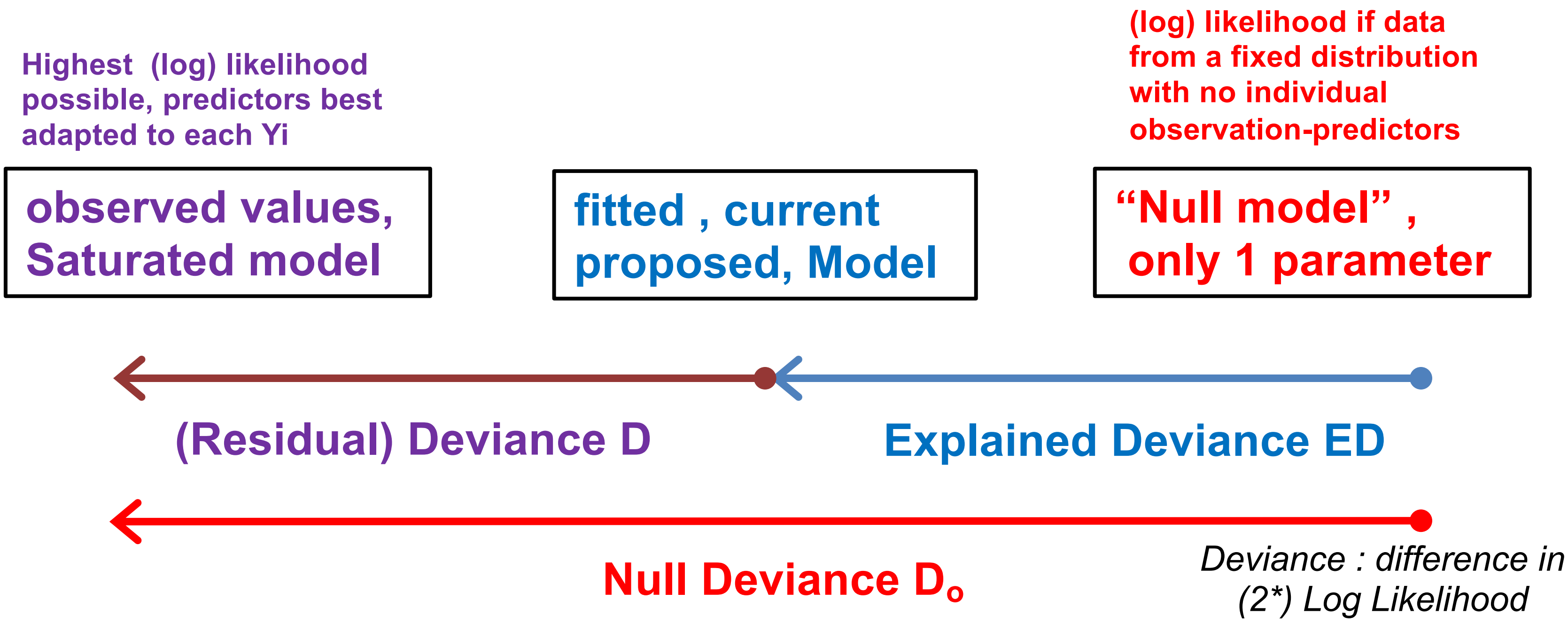
$$\underset{\text{SST}}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \underset{\text{SSR}}{\sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2} + \underset{\text{SSE}}{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}$$

R^2 = SSR / SST = 1 – (SSE / SST)

HEURISTIC REPRESENTATION



MLE likelihood and deviance



Highest (log) likelihood possible, predictors best adapted to each Yi

observed values, Saturated model

fitted , current proposed, Model

(log) likelihood if data from a fixed distribution with no individual observation-predictors

“Null model” , only 1 parameter

(Residual) Deviance D

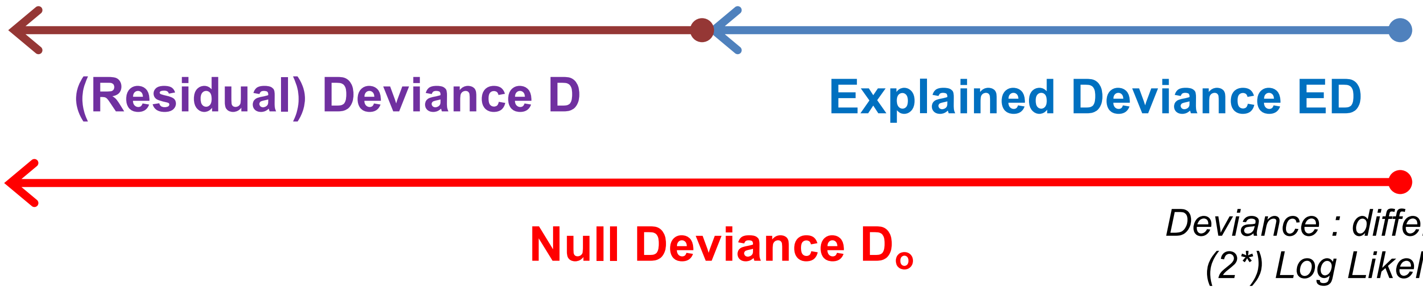
Explained Deviance ED

Null Deviance D_o

Deviance : difference in (2) Log Likelihood*

Analogous of SST^2 in LM $D_o = ED + D$; $ED = D_o - D$

MLE likelihood and deviance



(Residual) Deviance D

Explained Deviance ED

Null Deviance D_o

Deviance : difference in (2) Log Likelihood*

Analogous of SST^2 in LM $D_o = ED + D$; $ED = D_o - D$

Analogous of R^2 in LM $R^2 = ED / D_o = (D_o – D) / D_o) = \mathbf{1} - (D / D_o)$

called a pseudo- R^2

Reminder:
log (a / b) = log a - log b; log of the ratio equal difference of the logs

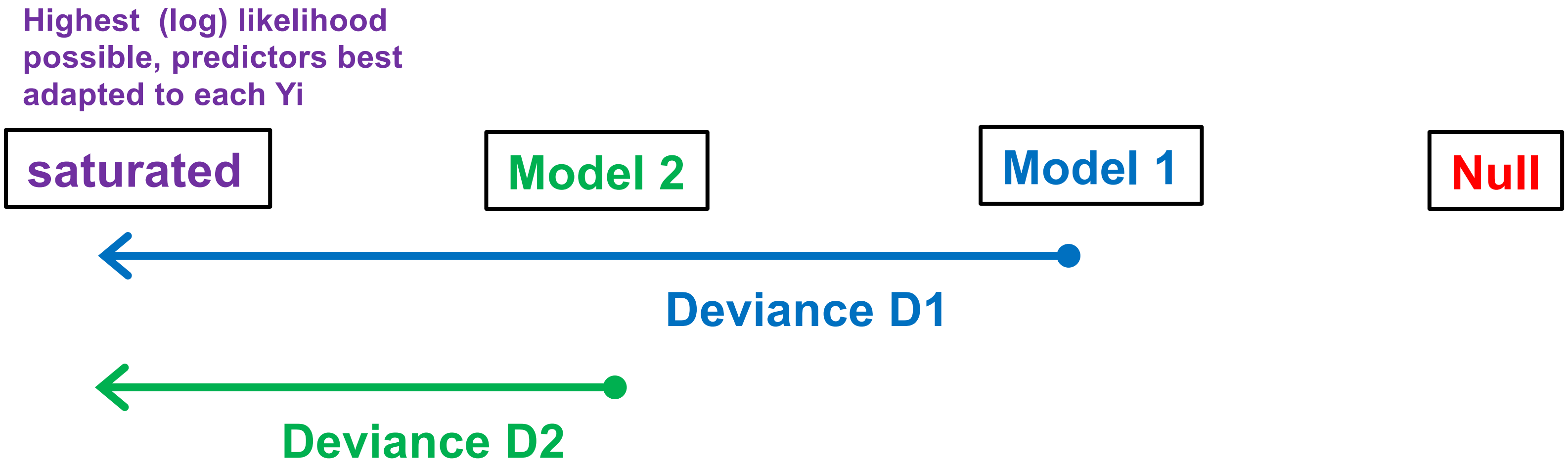
Notes:
log Lik ≤ 0 ; good Log Lik is close to 0;

Deviance > 0 , a measure of ”lack of fitting”, good is small positive close to 0
Maximal (Log) Likelihood ~ Minimal Deviance

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 in the sense of lack of fit.
- MLE is equivalent to finding parameter values that minimize the deviance.
- 2 values of deviance usually reported
  - ❑ **Null deviance**: how well (or bad) the response variable is predicted by a model that includes only the intercept (overall mean, logistic: binomial with fixed p) compared to the best possible model
  - ❑ **Residual deviance**: how much deviance is missing compared to the best model after including the proposed set of independent variables (residual lack of fit)

model comparison tests with deviance



Highest (log) likelihood possible, predictors best adapted to each Yi

saturated

Model 2

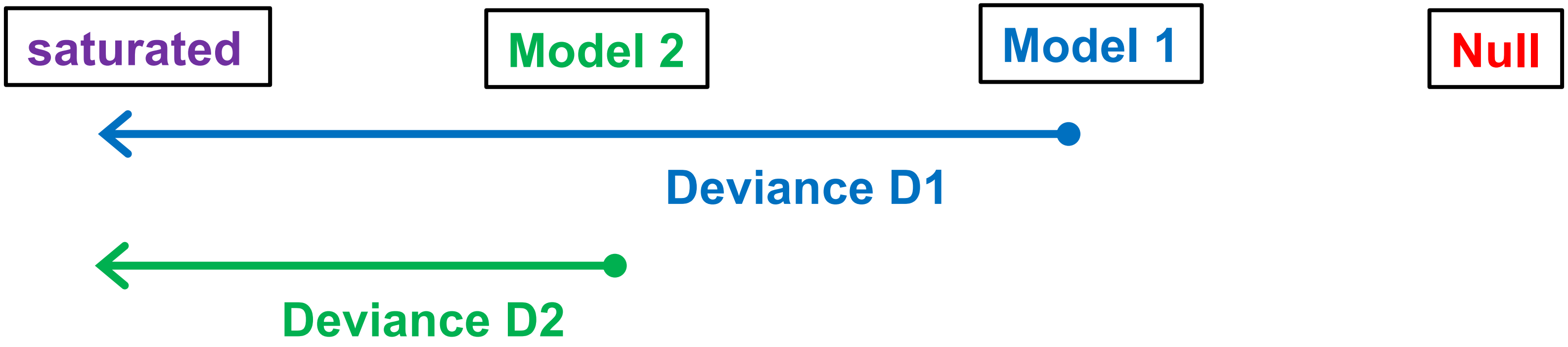
Model 1

Null

Deviance D1

Deviance D2

model comparison tests with deviance



Likelihood Ratio Test LRT

If a pair of models is **nested** (i.e. the smaller model 1 is a special case of the larger model 2; larger model has some additional predictors (and degrees of freedom))

then we can test if the improvement is statistically significant (more than expected by random effects) with a **likelihood ratio test = deviance test = Wilks test**

model comparison tests with deviance

Likelihood Ratio Test LRT

If a pair of models is **nested** (i.e. the smaller model 1 is a special case of the larger model 2;   larger model has some additional predictors (and degrees of freedom))

then we can test if the improvement is statistically significant (more than expected by random effects) with a **likelihood ratio test = deviance test = Wilks test**

Test statistic = LRTS = 2 x Log Lik Ratio =   **Deviance D1** **-** **Deviance D2**

~ **chi2 distribution**   with degrees of freedom =   df for larger model - df for smaller model

Example R code :
Anova (model1, model2, test = "Chisq")

R squared

Analogous of R^2 in LM

$$R^2 = \mathbf{ED} / \mathbf{D_o} = (\mathbf{D_o} - \mathbf{D}) / \mathbf{D_o}) = \mathbf{1} - (\mathbf{D} / \mathbf{D_o})$$

called a pseudo- R^2

Many different R-Squared and adjusted R-Squared have been proposed for GLM
Some are fairly widely used but generally model selection is best done with LRT

Akaike Information Criterion (AIC)

- allows to assess the quality of a model through comparison of related models
- based on the Deviance, but penalizes for the number of parameters (like adjusted R-squared, it's intent is to correct for irrelevant predictors)

Model selection

Nested Models: LRT

Otherwise: complicated
¿ Nothing simple works reliably ?
Resampling methods (learning-testing, cross-validation, bootstraps)

See statistical learning / machine learning

Problems

Non-Linearity in the logit

Poor fit overall
Outliers

Influential points

Multi-collinearity among predictors

Questions

Is the model appropriate ?

Does another **link function** give a better fit ?
(example: binomial family regression: logit or complementary log-log which can better fit cases asymmetric about 0.5 , …)

Does another model type («**family»)** give a better fit ?
(example: binomial vs. Poisson vs. quasi…)

Complementary Log-Log transformation

$\log \{-\log [1- \pi(x)] \}$ linear in X , $=X\text{ß}$

$\pi(x) = 1 - \exp (- \exp (X\text{ß}))$

[redacted]

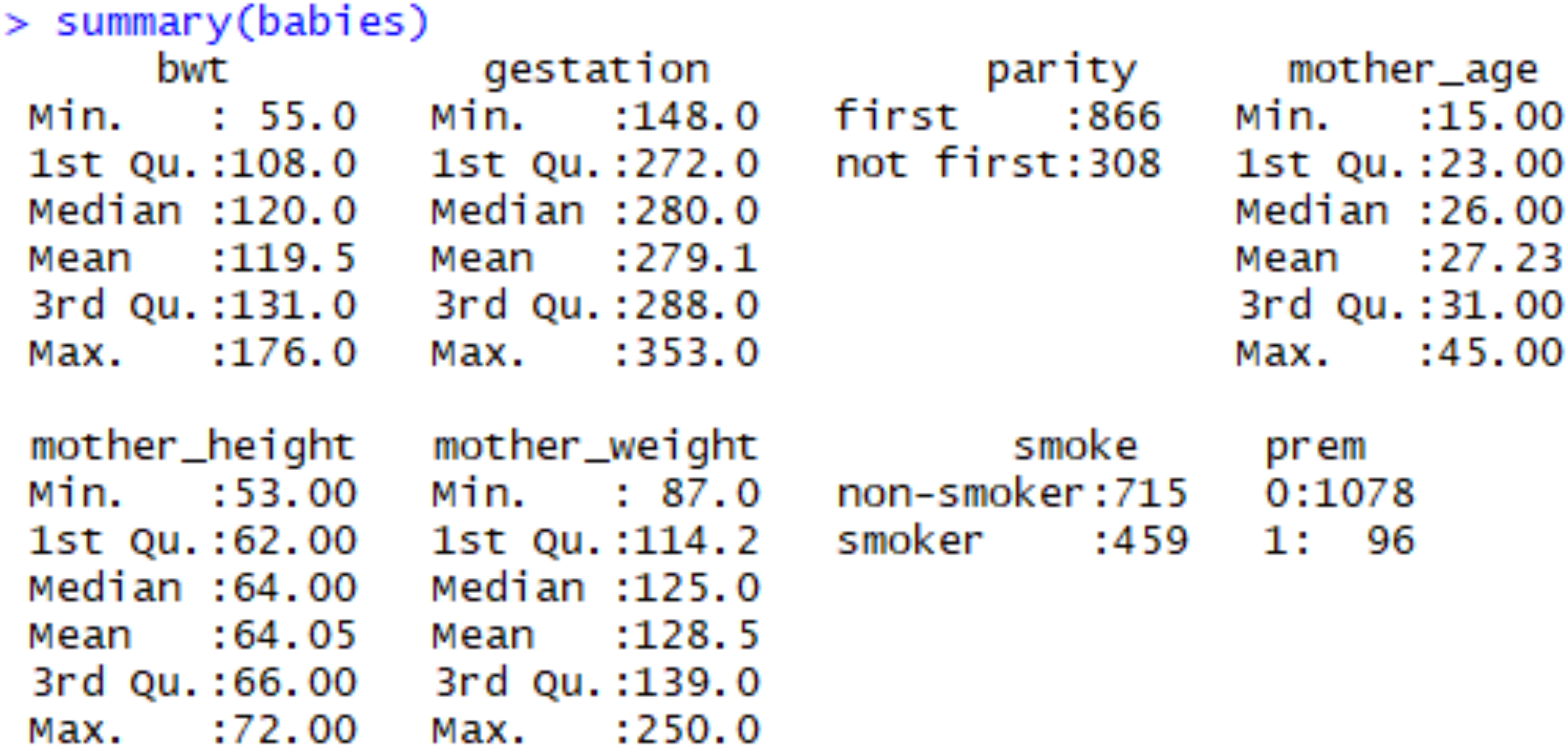
Warm up

Load and explore the dataset babies.

```
load("exercises/babies.RData")
```

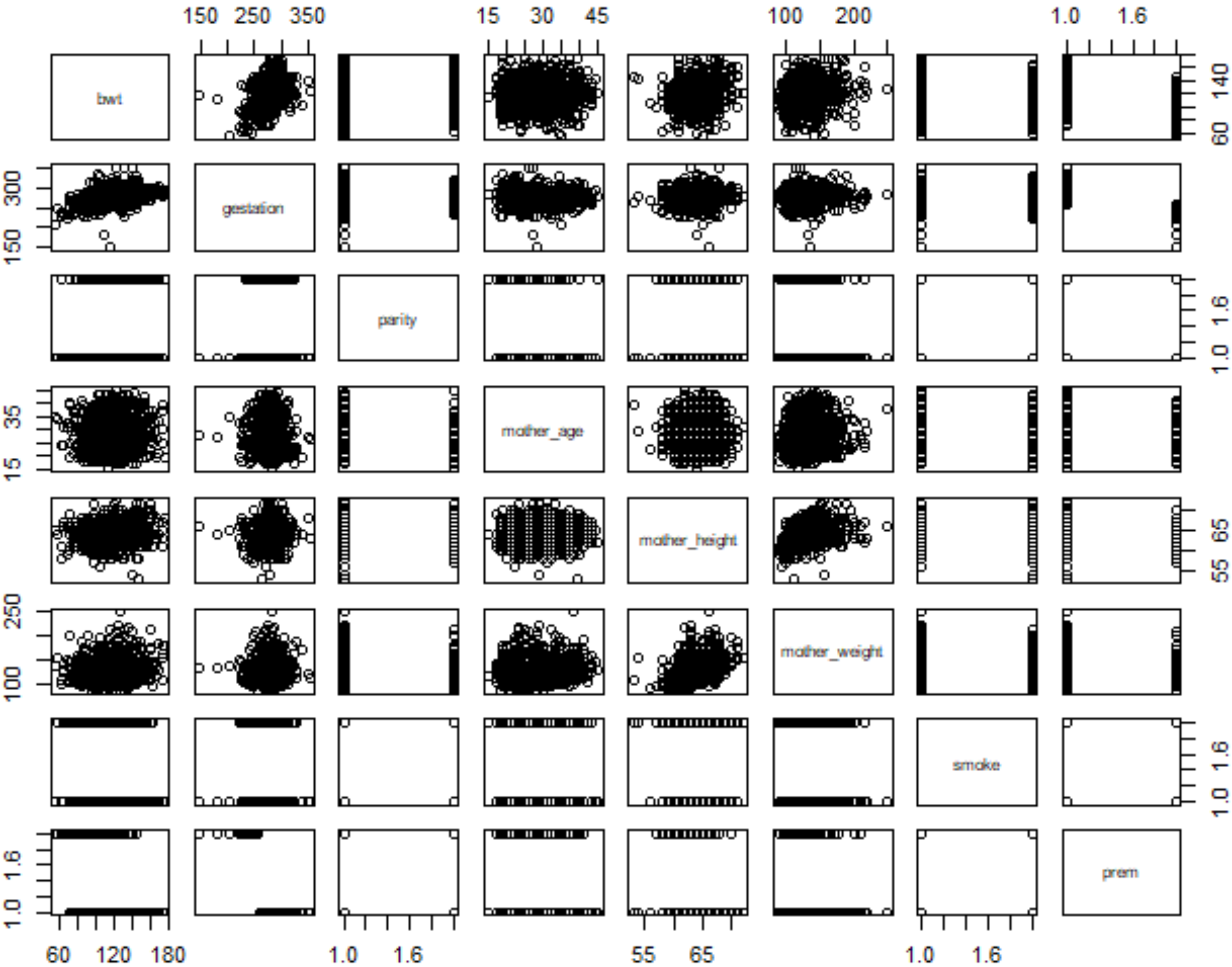
The data records the birth weight of 1174 babies along with information on the mother and the pregnancy.

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

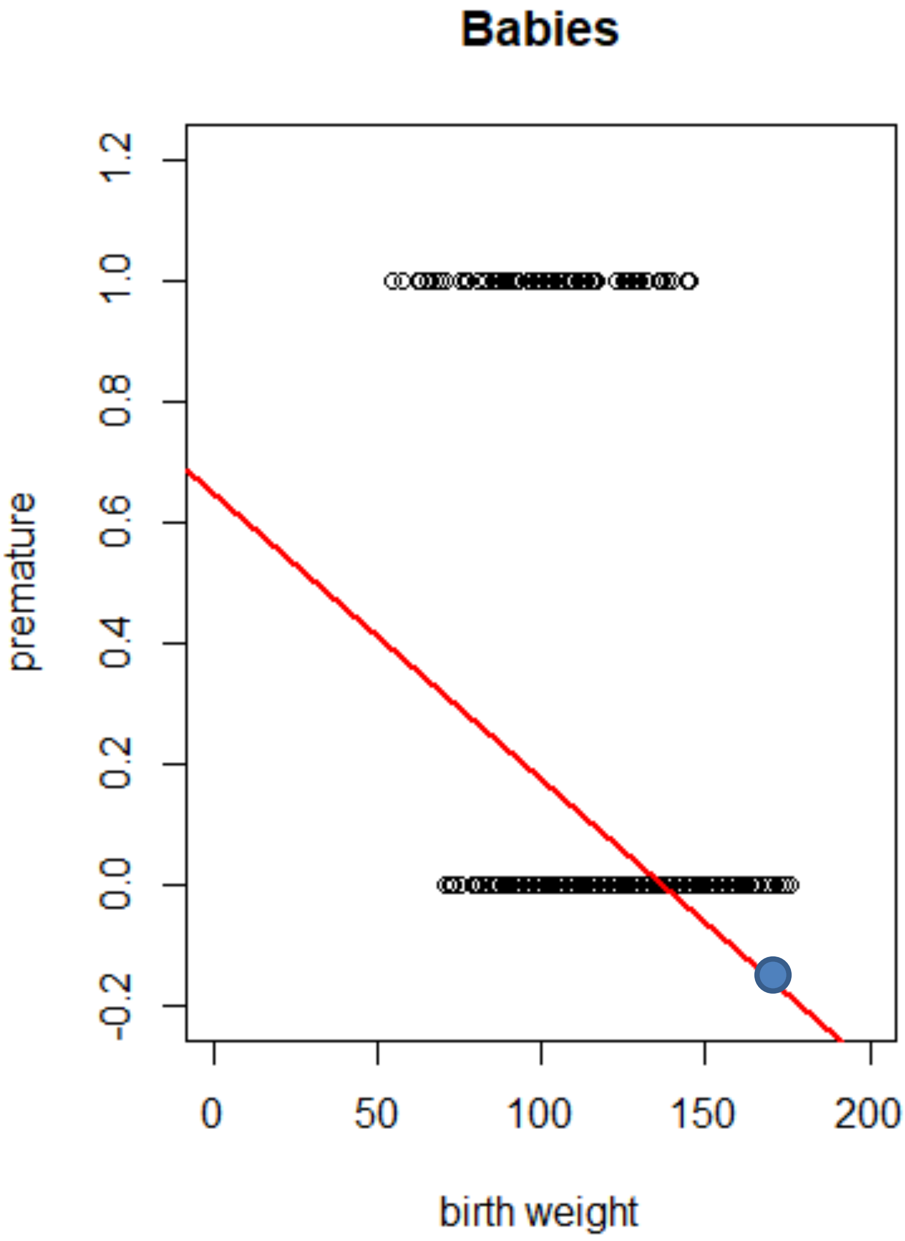


```
> summary(babies)
      bwt          gestation             parity       mother_age   
 Min.   : 55.0   Min.   :148.0   first    :866    Min.   :15.00  
 1st Qu.:108.0   1st Qu.:272.0   not first:308    1st Qu.:23.00  
 Median :120.0   Median :280.0                    Median :26.00  
 Mean   :119.5   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
140
60
300
150
1.6
1.0
35
15
65
55
250
100
1.6
1.0
1.6
1.0
60
120
180
1.0
1.6
55
65
1.0
1.6

Prediction if birth weight=170 ?



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)
```

Example 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
```

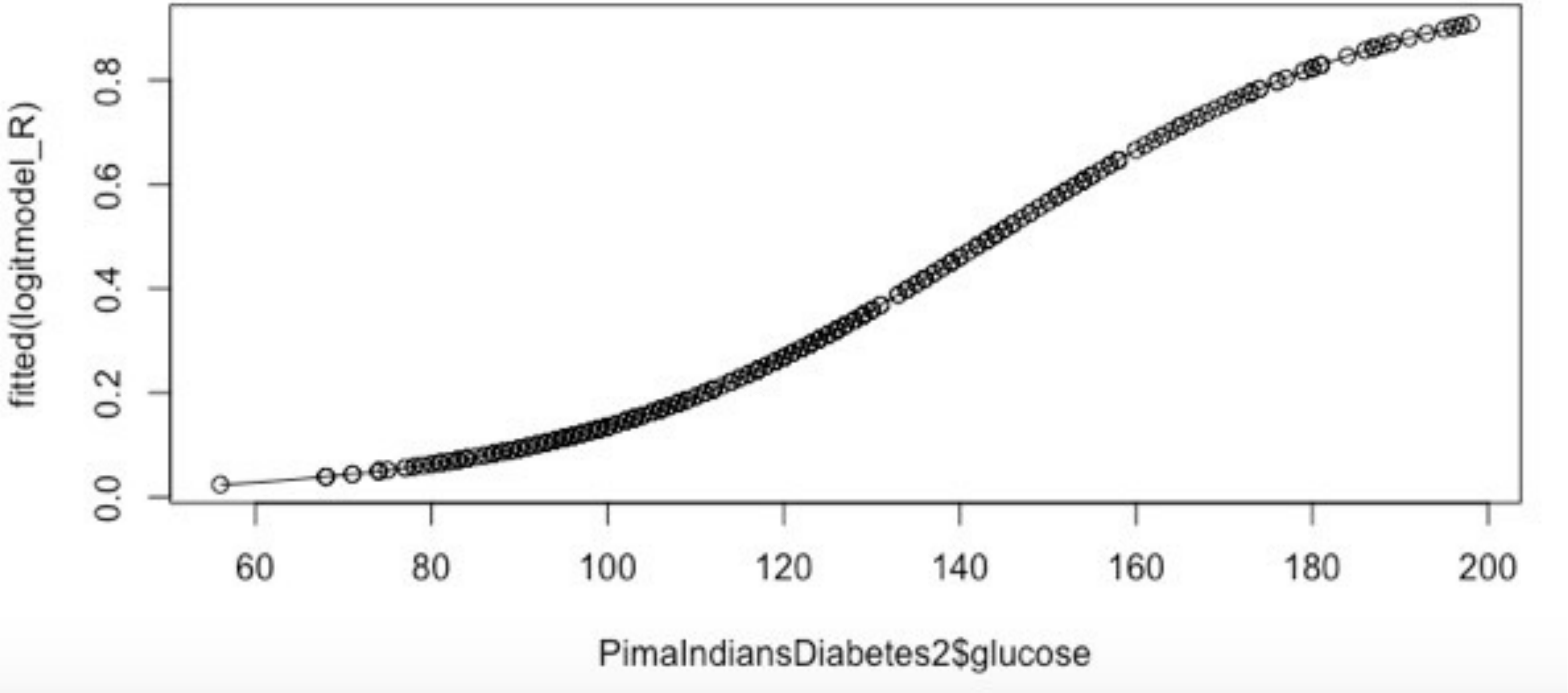
```
Call:
glm(formula = diabetes ~ glucose, family = binomial, data
= PimaIndiansDiabetes2)

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**

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

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

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



Example 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              or z value Pr(>|z|)
(Intercept) -6.095521              787  -9.679   <2e-16 ***
glucose      0.042421              761   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
```

[redacted]

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:

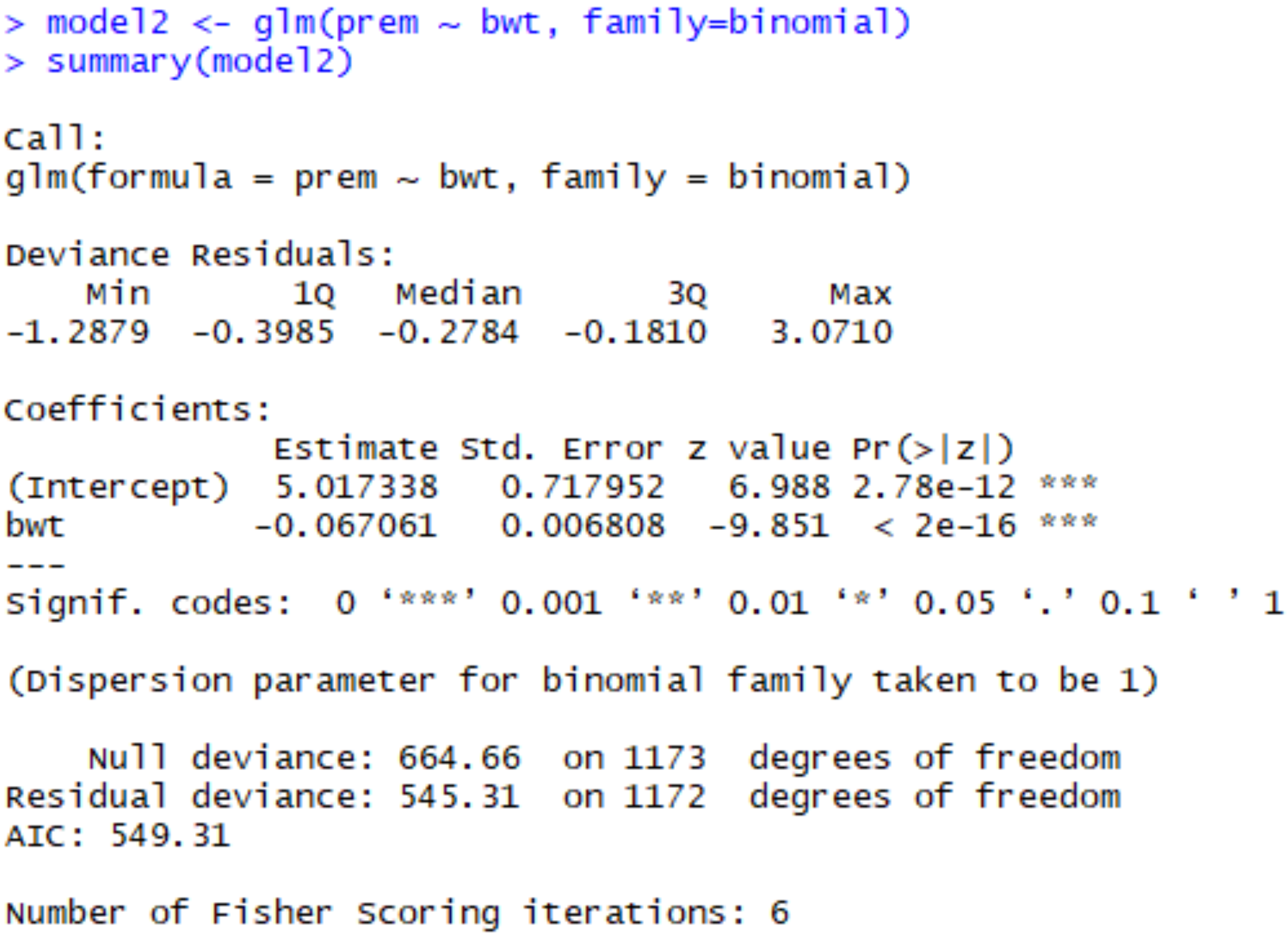
$$OR_i = e^{\beta_i}$$

Challenge 1

Using the babies dataset

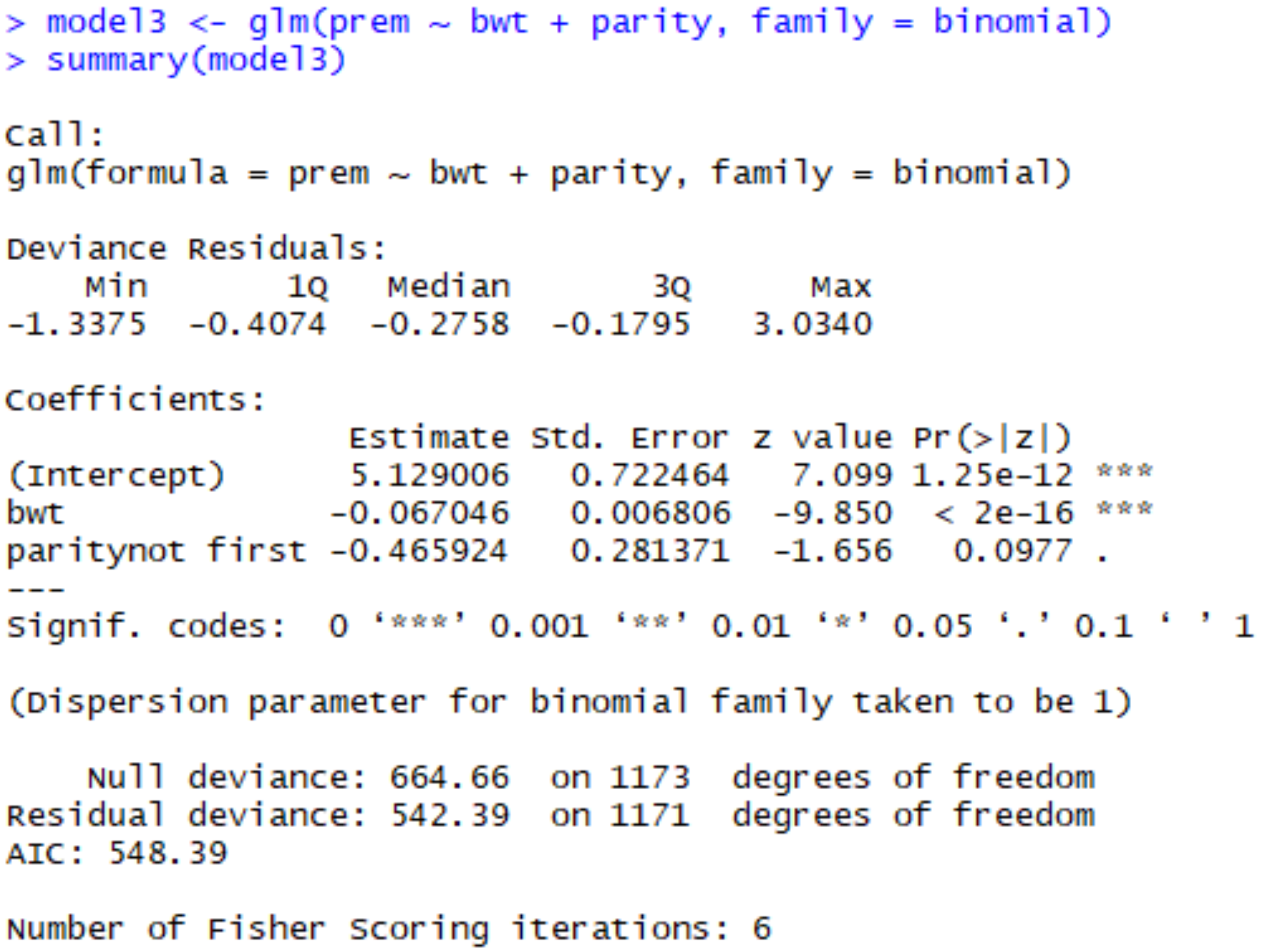
- Fit a logistic regression to find parameters explaining the probability of prematurity ?
- What is the effect of birth weight on the probability of prematurity ?
- What about parity ?

Solution



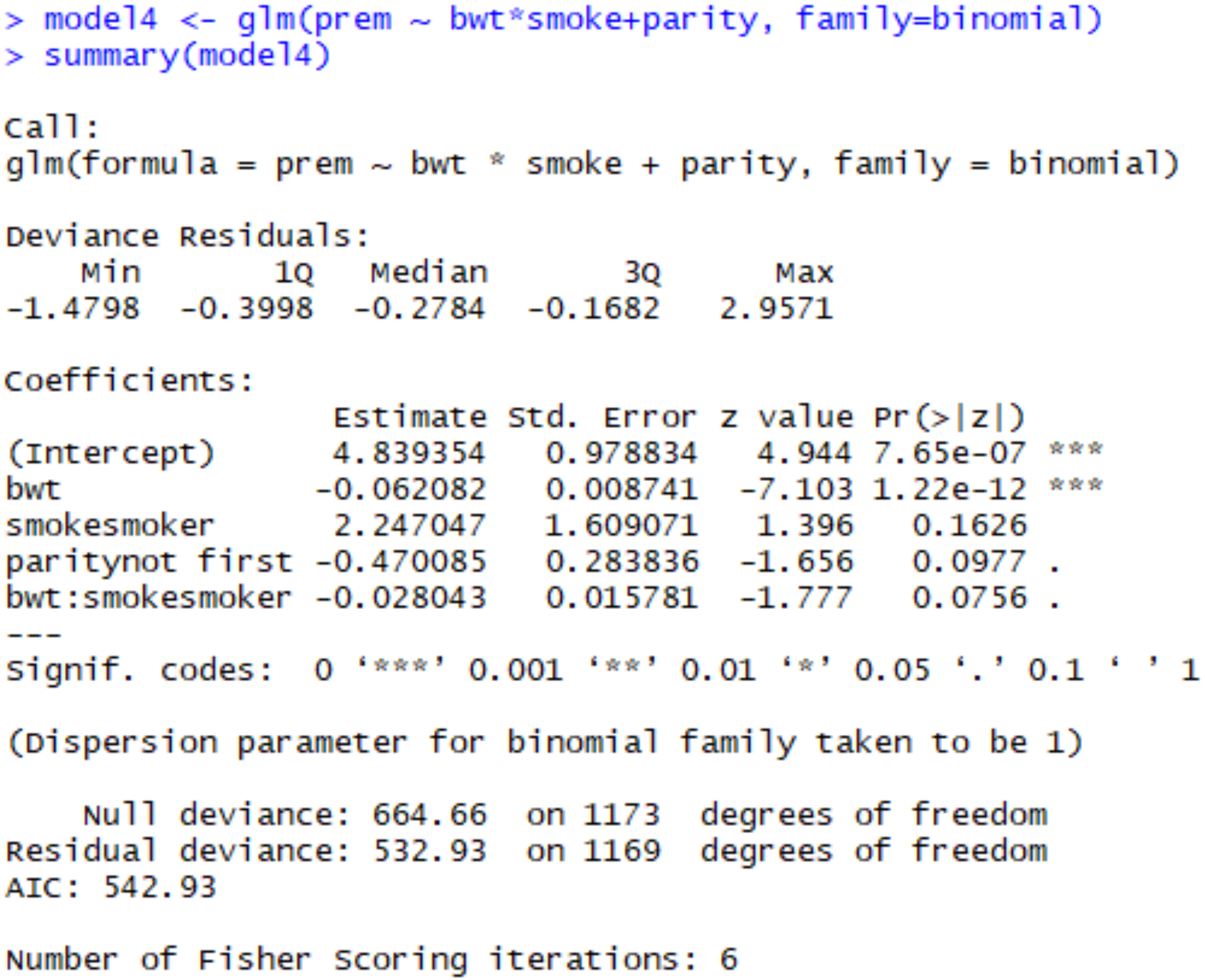
```
> model2 <- glm(prem ~ bwt, family=binomial)
> summary(model2)

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

[redacted]

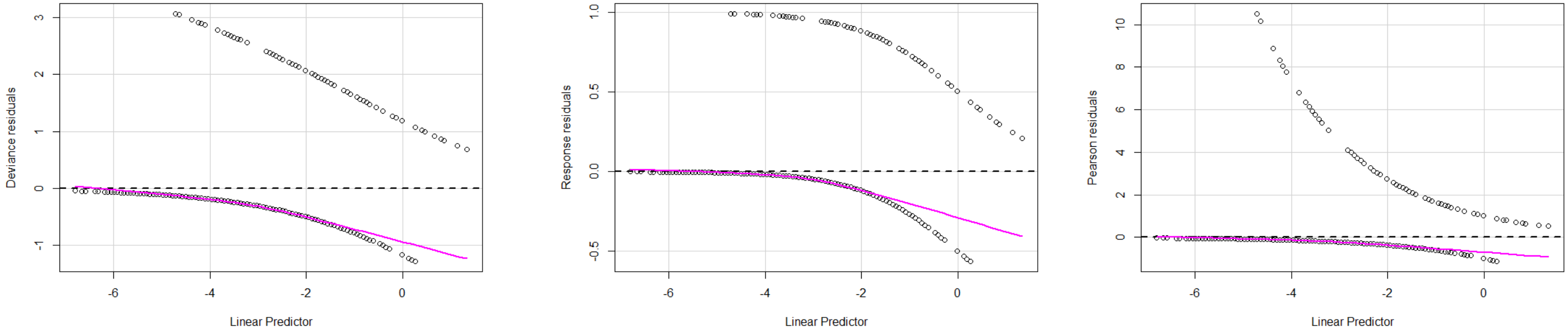
Challenge-1 b

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

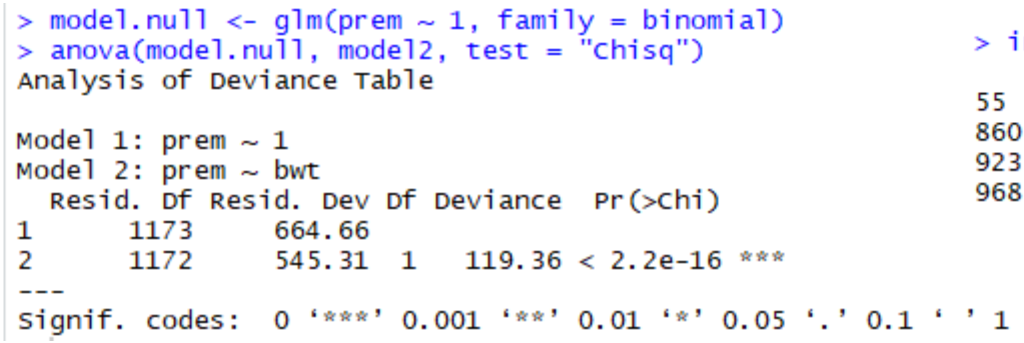
Residuals

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

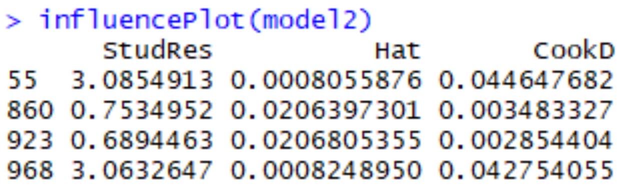


Residuals

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

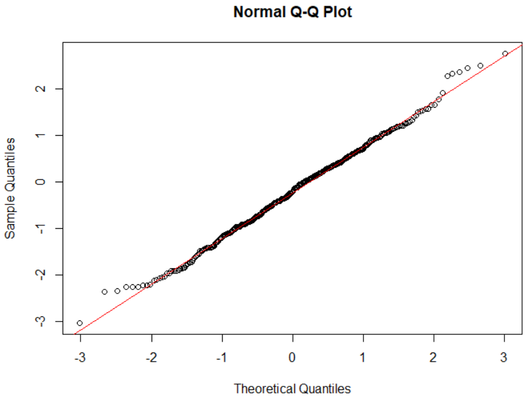


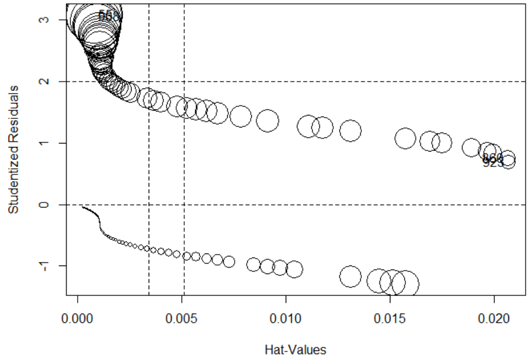
```
> model.null <- glm(prem ~ 1, family = binomial)
> anova(model.null, model2, test = "Chisq")
Analysis of Deviance Table

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
```

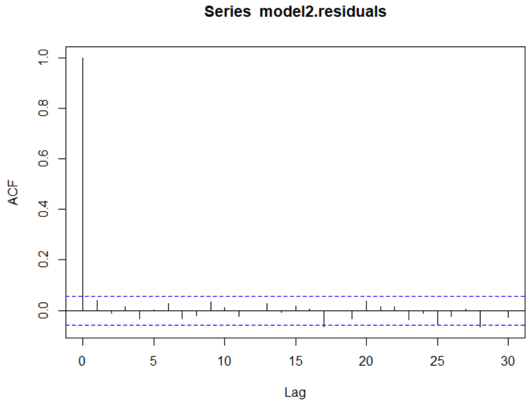


```
> influencePlot(model2)
      StudRes          Hat       CookD
55  3.0854913 0.0008055876 0.044647682
860 0.7534952 0.0206397301 0.003483327
923 0.6894463 0.0206805355 0.002854404
968 3.0632647 0.0008248950 0.042754055
```

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







Challenge 2: baby food

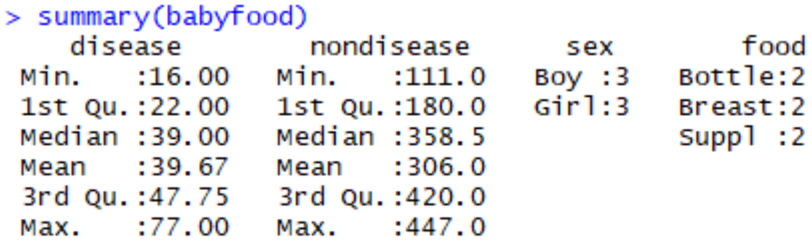
The data for this exercise study infant respiratory disease, namely the proportions of children developing bronchitis or pneumonia in their first year of life by type of feeding, and sex. Data may be found in Payne (1987) and Faraway (2006)

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

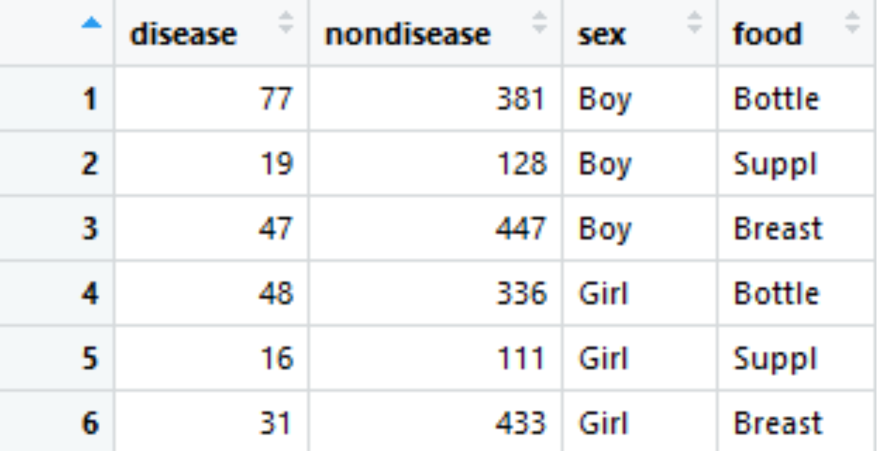
1. Explore the data
2. Fit a logistic regression to explain the probability of disease by sex and food.

Challenge: baby food: solution

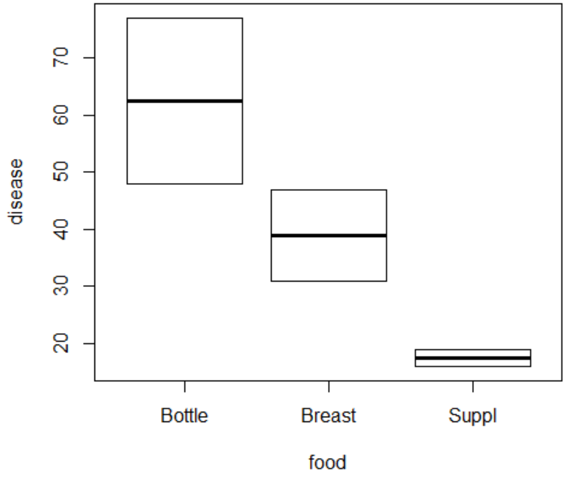
1. summary(babyfood)
boxplot(disease ~ food, babyfood)
boxplot(disease ~ sex, babyfood)

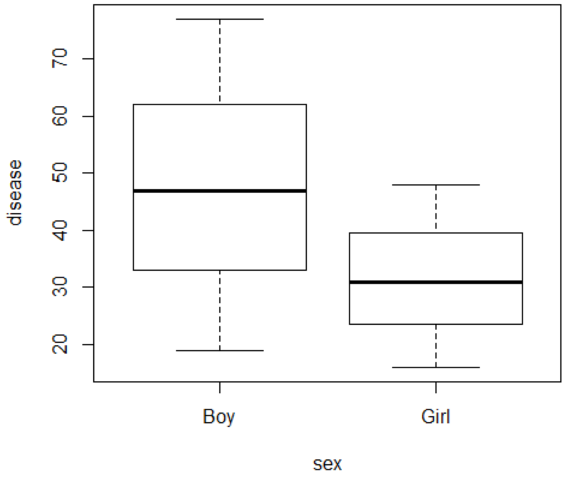


```
> summary(babyfood)
    disease          nondisease        sex          food   
 Min.   :16.00   Min.   :111.0   Boy :3   Bottle:2  
 1st Qu.:22.00   1st Qu.:180.0   Girl:3   Breast:2  
 Median :39.00   Median :358.5            Suppl :2  
 Mean   :39.67   Mean   :306.0                      
 3rd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



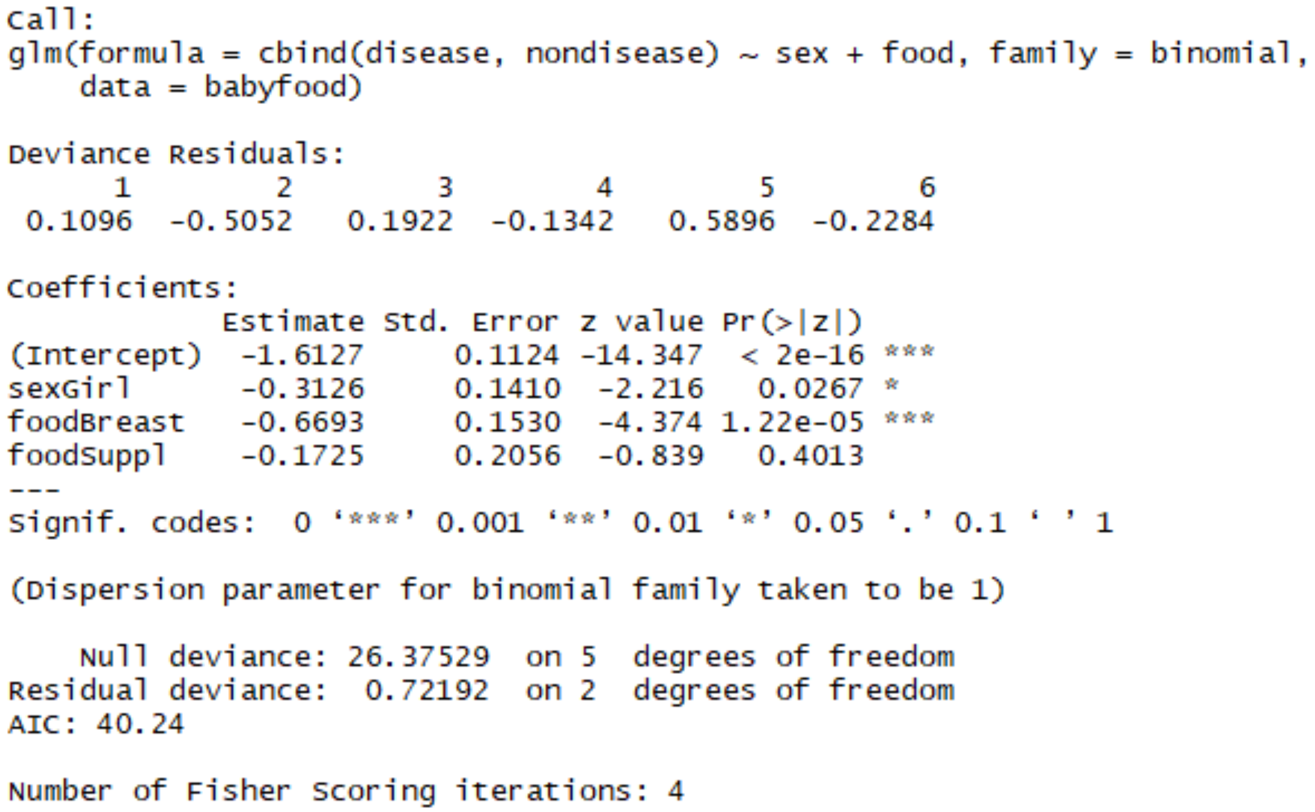
	disease	nondisease	sex	food
1	77	381	Boy	Bottle
2	19	128	Boy	Suppl
3	47	447	Boy	Breast
4	48	336	Girl	Bottle
5	16	111	Girl	Suppl
6	31	433	Girl	Breast





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
```

[redacted]

GLM diagnostic: analysis of residuals

- Detection of influential observations
- Anomalies and Outliers: Deviance residuals vs fitted values
- Missing patterns: Deviance residuals vs each of the available covariates
- Dispersion check: Quantile Residuals

Many more checking procedures are known, but **interpretation** and recommended actions rarely straightforward

GLM diagnostic: Hat and Cook

- **Detection of influential observations**

1) **Hat values h_i** In analogy to LM there is a definition of a hat matrix for logistic regression fits and large diagonal values suggest a potential high influence of a point on the obtained fit. Limit ~ 2p / n or 3p / n.

2) **Cook's distance Cd_i** is a measure of a change in estimated coefficients when the observation i is ignored. Large values (> ~ 4/n) suggest a large influence, pointing to observations one might want to "investigate".

3) The square of the individual **deviance residuals** (see below) can also indicate single observation points with high influence.

GLM diagnostic: analysis of residuals

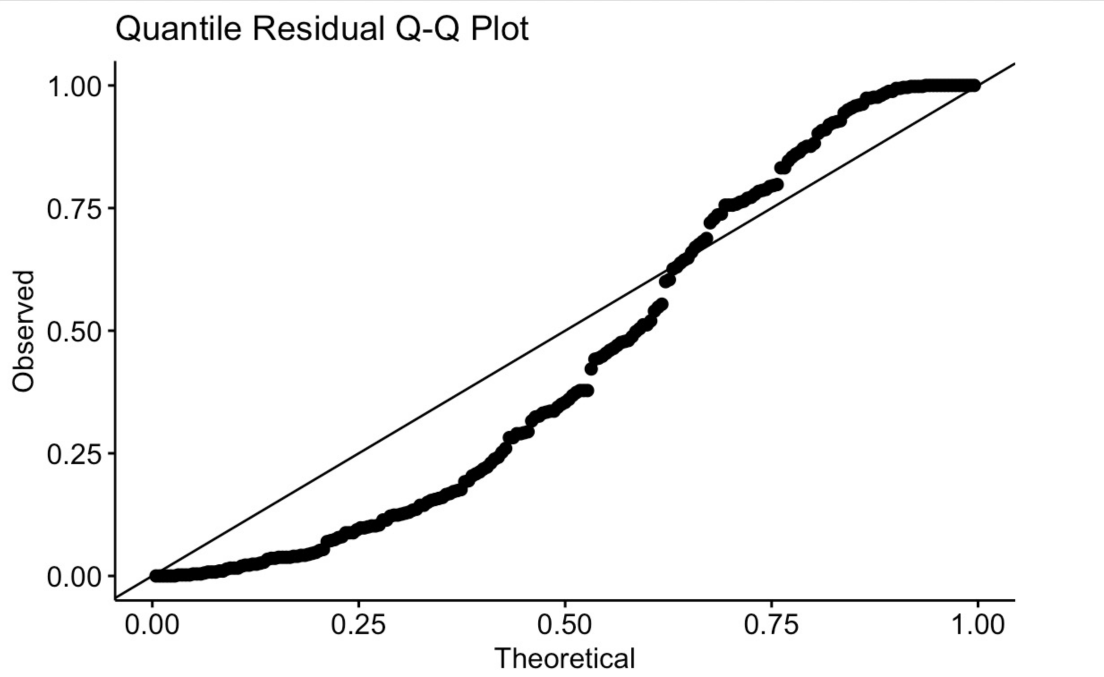
Quantile Residuals $\mathbf{QR_i}$

One can generate faked simulated data from the fitted model, like distribution of predicted new values and compare to the observed values (for each data point, each $\mathbf{x}_i$). The nb of fakes < obs, y_i is called quantile residual $\mathbf{Qr_i}$. If the data are distributed as specified by the model these follow a uniform U[0,1] distribution. A Q-Q plot of calculated vs. expected quantile residuals can detect significant departures and suggest modifications to the model.

GLM diagnostic: analysis of residuals

Quantile Residuals QR_i

Example: (Poisson GLM QRs)
Several excessively extreme (larger and smaller than expected) QRs in checking
suggests overdispersion of data and might suggest the use of a **quasi-Poisson** or a **Negativ Binomial** approach instead of the Poisson.



GLM diagnostic: analysis of residuals

“**Raw Residuals” RR_i = Y_i - fitted E [Y | X)** , where **Y_i** = 0 or 1
response residuals are not informative for assessing the characteristics of a GLM fit (Note: given the model the variance can depend on the mean so even their spread is not directly informative. GLM : no assumption on constant variance or normality of these residuals). Generally rather use the Pearson residuals.

Pearson Residuals Pr_i
Studentized (Pearson) Residuals” Sr_i

Deviance Residuals dr_i:

GLM diagnostic: analysis of residuals

“**Raw Residuals” $\mathbf{RR_i}$ = $\mathbf{Y_i}$ - fitted E [Y | X)** , where $\mathbf{Y_i}$ = 0 or 1

Pearson Residuals $\mathbf{PR_i}$: are adjusted for expected variance (given X) and are expected to follow approximately a normal distribution at each X_i (under assumptions).
Can reveal potential outliers. Large residuals (in absolute value) are “somewhat strange” compared to their “neighbour points”, but not necessarily to be considered outliers (in general some large residuals have to be expected).
A (linear) trend in a plot of **$\mathbf{PR_i}$** against covariates might identify predictors that should have been omitted in the model but should maybe be included.
Trends: add a loess to the graph to see trends.
A curved trend might indicate that adding a higher order term of the covariate could be useful (ex. x^2).

Studentized (Pearson) Residuals” $\mathbf{Sr_i}$

Deviance Residuals $\mathbf{dr_i}$:

GLM diagnostic: analysis of residuals

“Raw Residuals” RR_i = Y_i - fitted E [Y | X) , where **Y_i** = 0 or 1

Pearson Residuals PR_i :

“Studentized (Pearson) Residuals” SR_i = modified Pearson residuals so that their information is independently informative than influence measures (hat)

Deviance Residuals dr_i:
deviance contribution by point i to the residual deviance of the model. Typically similar trend like the **RR_i** .

[redacted]

Poisson Regression
for
count data

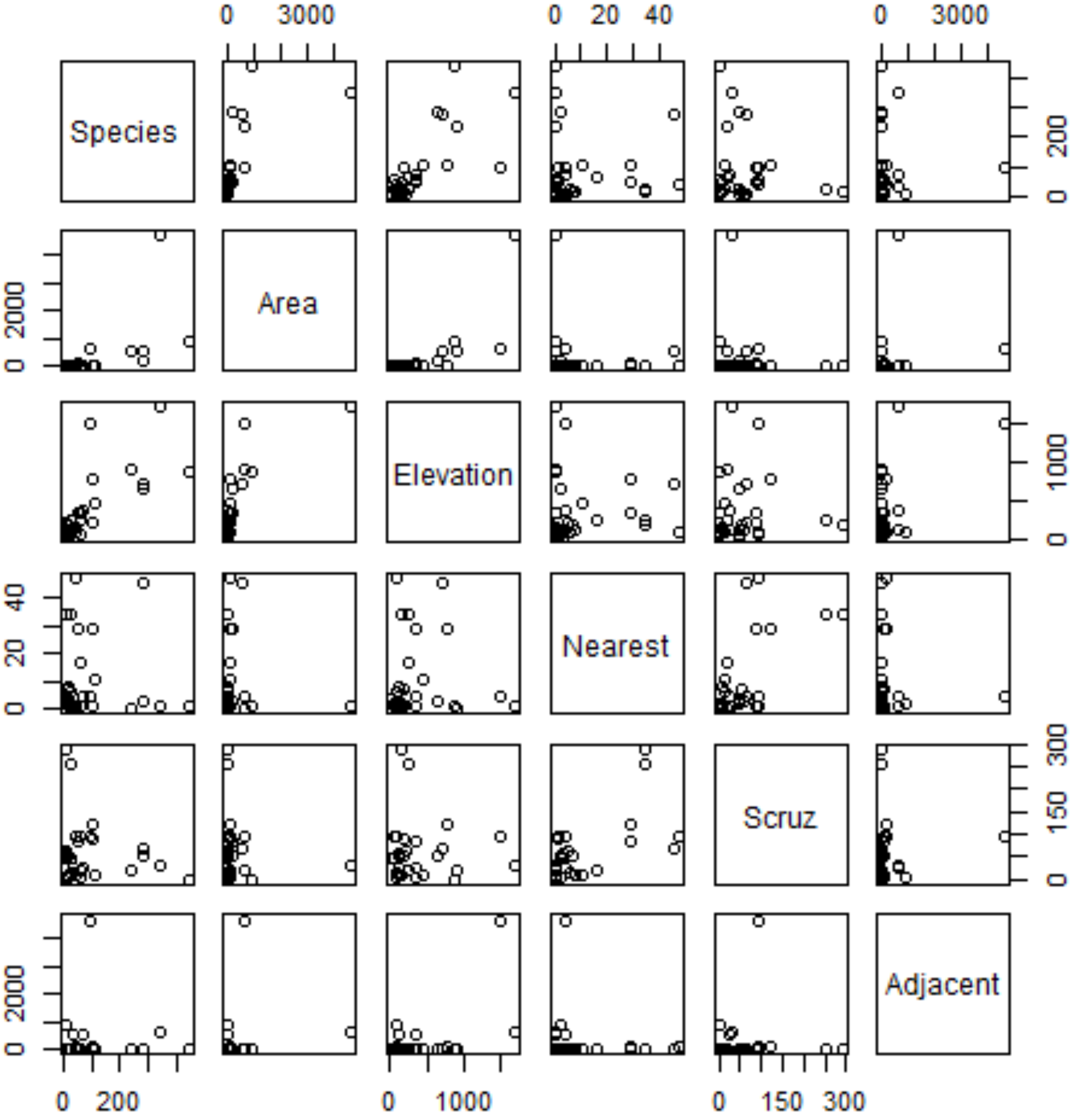
Warmup

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
200
0
2000
0
1000
0
20
40
0
150
300
0
2000
0
200
0
1000
0
150
300

GLM Poisson

Predictors X => E(link(Y) | X) **=>** observations Y_i

=> The assumed model of effects

Log (λ_i) linear in ß's

$\Rightarrow$ **stochastic process**

Data Y_i ~ Poisson distribution
Poi (mean λ =E[Y])
Stdev = sqrt (λ)

The dispersion is the one expected for a »pure random sampling» that is without any factor of variability increasing the dispersion.

Stddev = sqrt(mean)

ML-estimation, deviance, LRT,
Wald test on coefficients etc:
Like Logistic Regression

Poisson regression

- Basic standard model used for Count data
- Distribution: Poisson, (Restriction: mean = variance : E(Y)=V(Y)=λ)
- Default 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

Poisson regression - assumptions

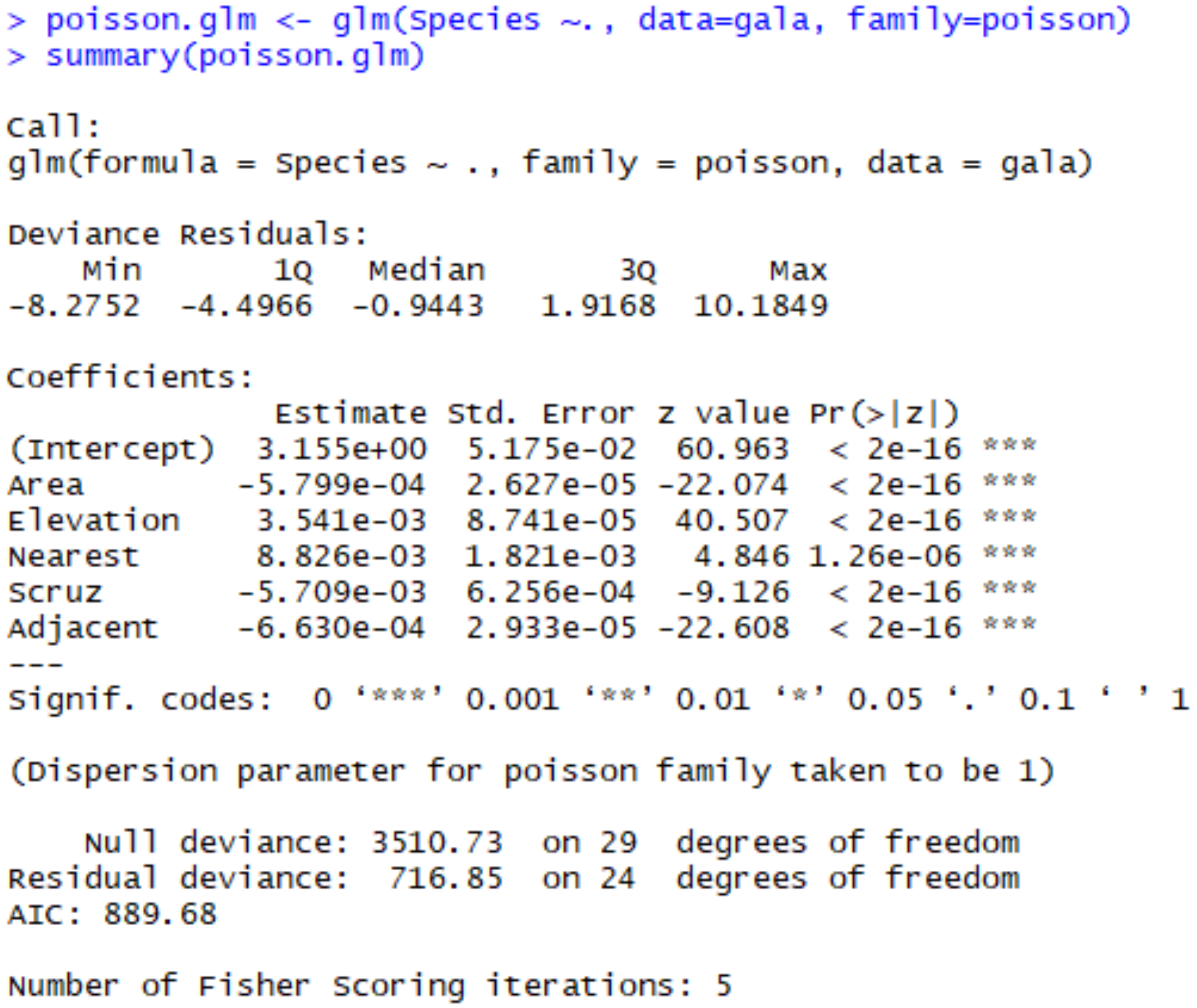
- Poisson at the Response Level : the response variable is a count per unit of time or space, described by a Poisson distribution.
- Linearity at Link Level : the log of the mean rate, log(λ), is be a linear function of the predictor x.
- Independence: the observations are independent of one another.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

Challenge

Using the glm function with family=poisson,

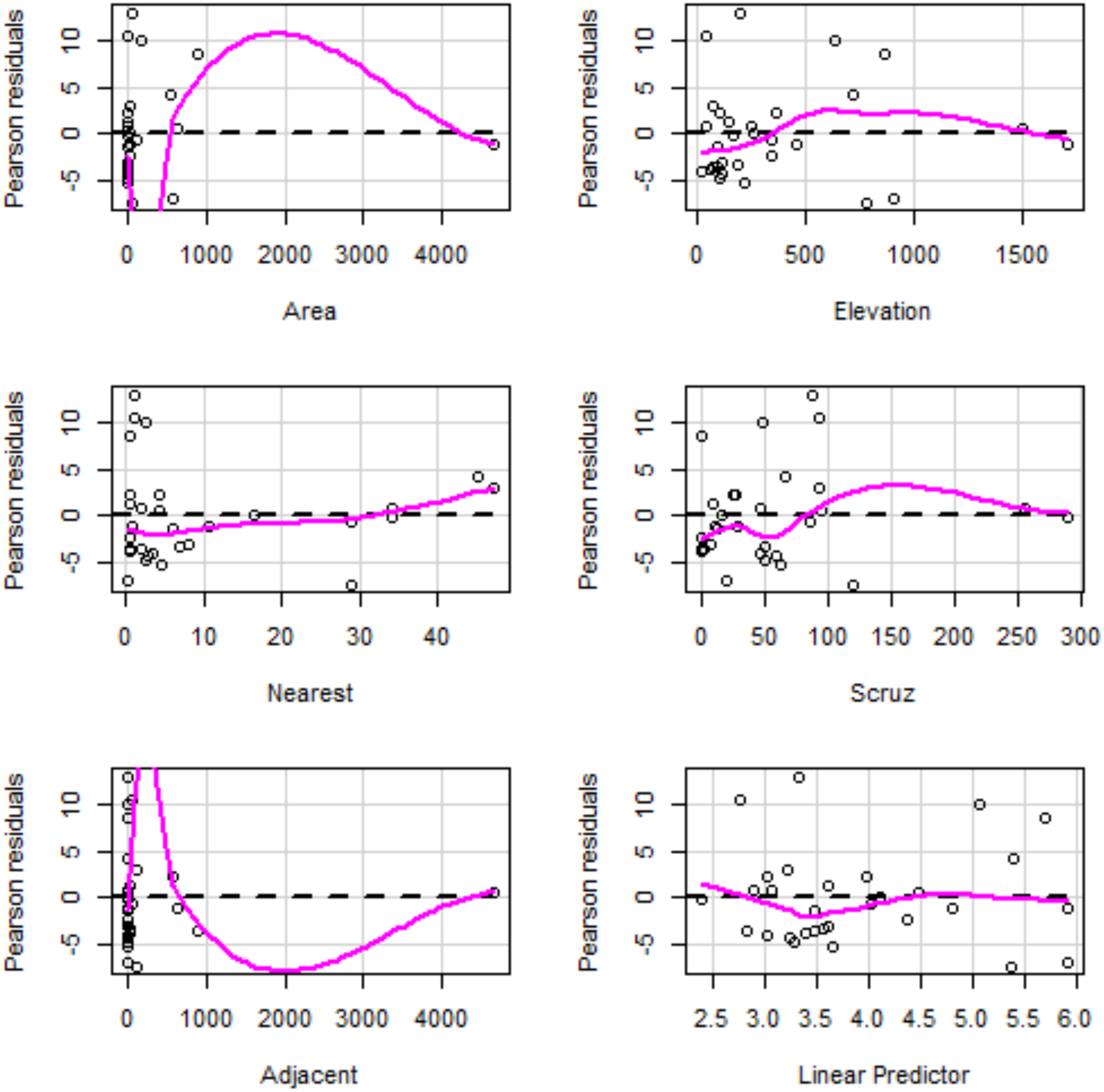
- Fit a poisson model to the galapagos data.
- Which variables are significant ?
- Check the deviance of the model

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