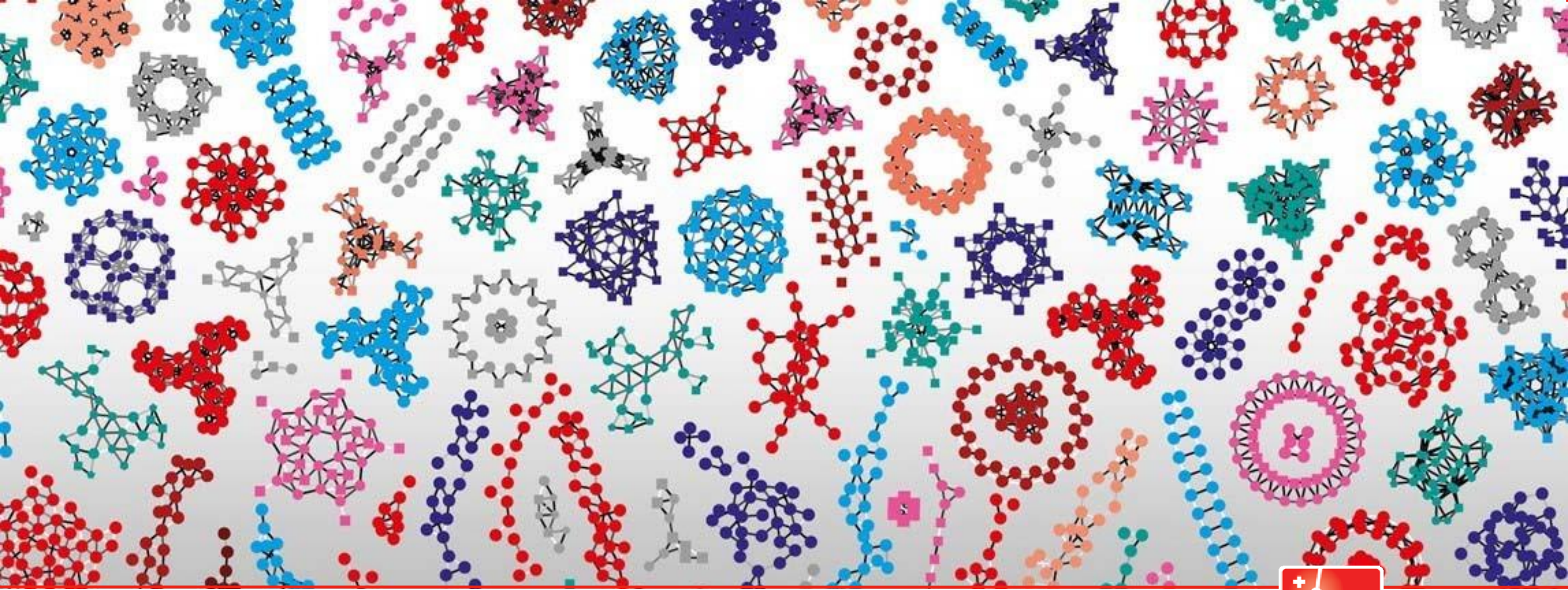


Swiss Institute of Bioinformatics

Regression

Rachel Marcone

www.sib.swiss

birthwt{MASS} Risk Factors Associated with Low Infant Birth Weight

The birthwt data frame has 189 rows and 10 columns. The data were collected at Baystate Medical Center, Springfield, Mass during 1986.

This data frame contains the following columns:

- low: indicator of birth weight less than 2.5 kg
- age: mother's age in years
- lwt: mother's weight in pounds at last menstrual period
- race: mother's race (1 = white, 2 = black, 3 = other)
- smoke: smoking status during pregnancy
- ptl: number of previous premature labours
- ht: history of hypertension
- ui: presence of uterine irritability
- ftv: number of physician visits during the first trimester
- bwt: birth weight in grams

Apply several regression models to the birthwt dataset

1. Inspect the dataset
2. Does the mother's age allow us to predict the birth weight below 2.5 kg ?
3. Can you use a linear regression model ? What’s wrong with this model ?
4. Which model is better adapted to the current situation ?
5. Check the assumptions of the model
6. Can you better predict birth weight using below 2.5 kg using other variables in the dataset, in addition to mother’s age ?

Apply several regression models to the birthwt dataset

1. Inspect the dataset
2. Apply a regression model to predict birth weight below 2.5 kg using mother’s age
3. Start with a simple linear regression model. What’s wrong with this model ?
4. Apply a logistic regression model to predict birth weight below 2.5 kg using mother’s age
5. Check the assumptions of the logistic regression model
6. Apply a multiple logistic regression model to predict birth weight below 2.5 kg using multiple predictors

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

Binary Logistic Regression Model

Y = Binary response: birth weight below 2.5 kg=1, birth weight below 2.5 kg =0

X = Quantitative predictor: mother's age

π = Proportion of success (birth weight below 2.5 kg) at any X

Logit form

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

Logit is the logarithm of the odds

π = Proportion of success

1 - π = Proportion of failure

Proportion of “success”

Y = Binary response: birth weight below 2.5 kg=1, birth weight below 2.5 kg =0

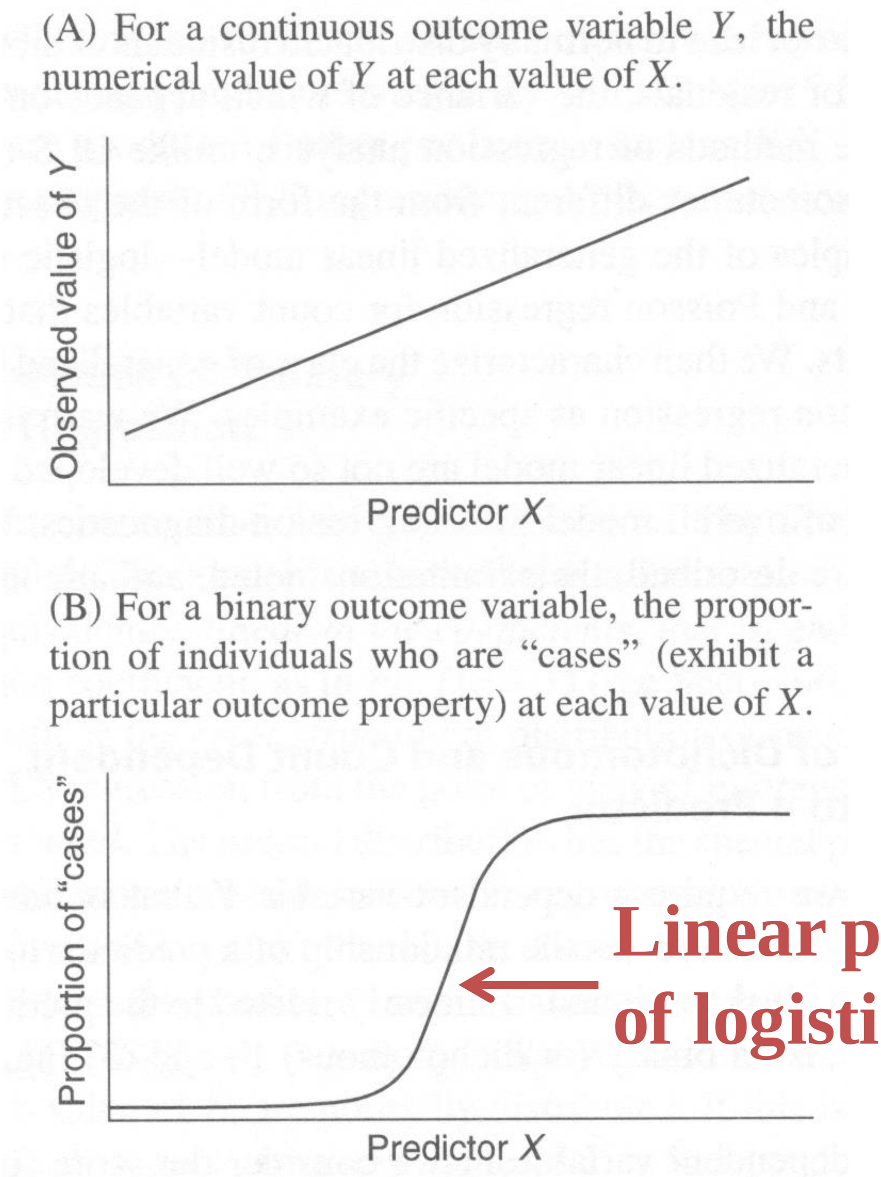
In linear regression the model predicts the mean Y for any combination of prediction.

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

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

Goal of logistic regression: Predict the “true” proportion of success, π, at any value of the predictor.

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

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

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

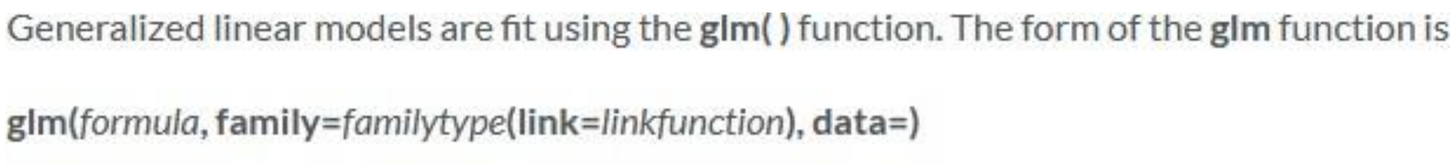
No outliers

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.

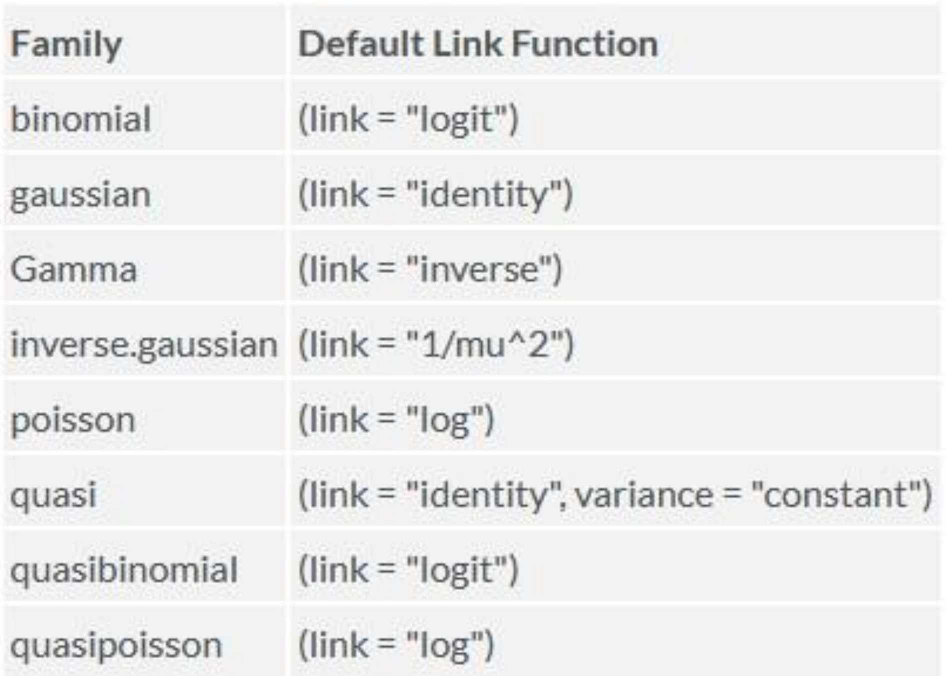
The glm() command is designed to perform generalized linear models (regressions) on binary outcome data, count data, probability data, proportion data and many other data types.

Generalized Linear Models



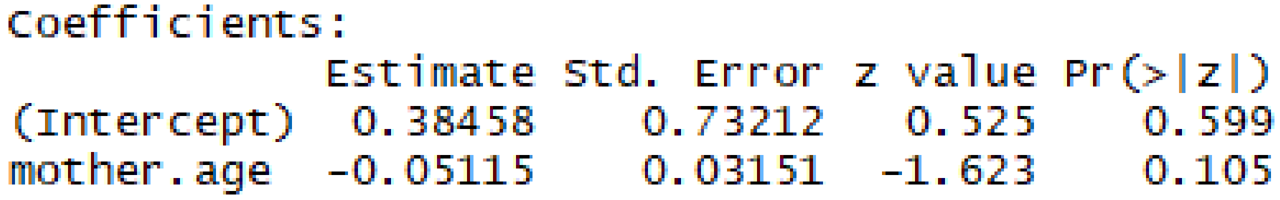
Generalized linear models are fit using the **glm( )** function. The form of the **glm** function is

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")

```
Call:
glm(formula = birthwt.below.2500  ~ mother.age, family =
binomial, data = birthwt)
```



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.38458    0.73212   0.525    0.599
mother.age  -0.05115    0.03151  -1.623    0.105
```

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

Proportion of birth weight below 2.5 kg at mother's age *m*

$$\hat{\pi} = \frac{e^{0.38 - 0.05m}}{1 + e^{0.38 - 0.05m}}$$

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 !

Multiple Logistic Regression

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

With *k* predictors, the model is written:

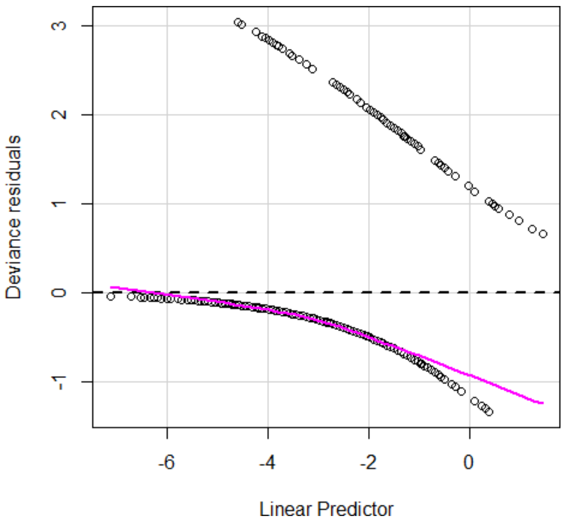
$$\pi = \frac{e^{\beta_0 + \beta_1 x_1 + .. + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + .. + \beta_k x_k}}$$

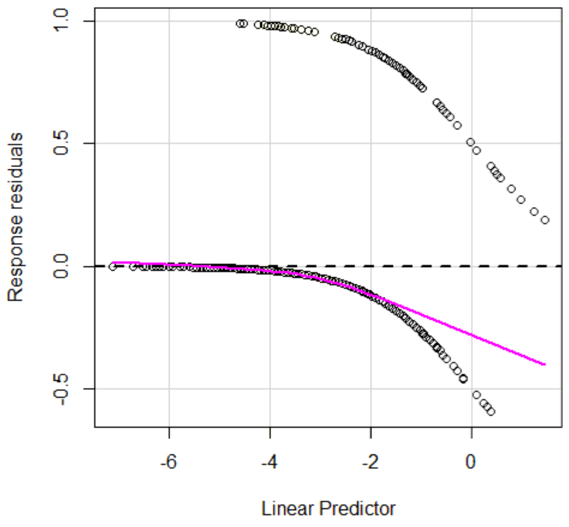
Adjusted Odds ratio for raising x_i by 1 unit, holding all other predictors constant:

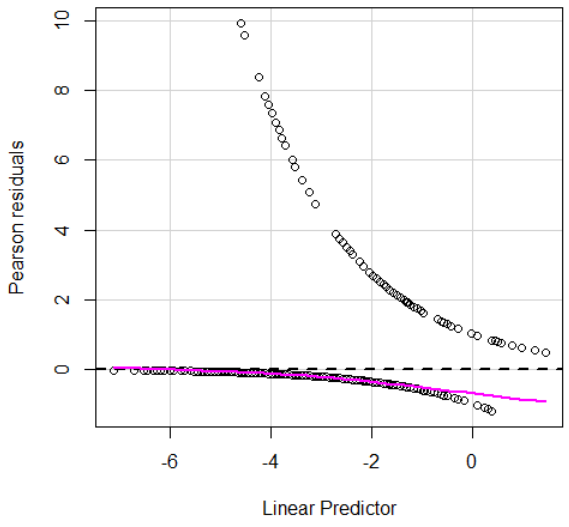
$$OR_i = e^{\beta_i}$$

Residuals

```
> require(car)
> residualPlot(model3, type = "deviance")
> residualPlot(model3, type = "response")
> residualPlot(model3, type = "pearson")
```


Deviance residuals
Linear Predictor


Response residuals
Linear Predictor


Pearson residuals
Linear Predictor