Advanced Regression: 1c Simple generalised linear models

Verena Zuber

Epidemiology and Biostatistics, Imperial College London

17th January 2019

The linear model

- Basic definition
- Assumptions
- Diagnostic plots in the linear model
- Prediction using linear models

Generalised linear model

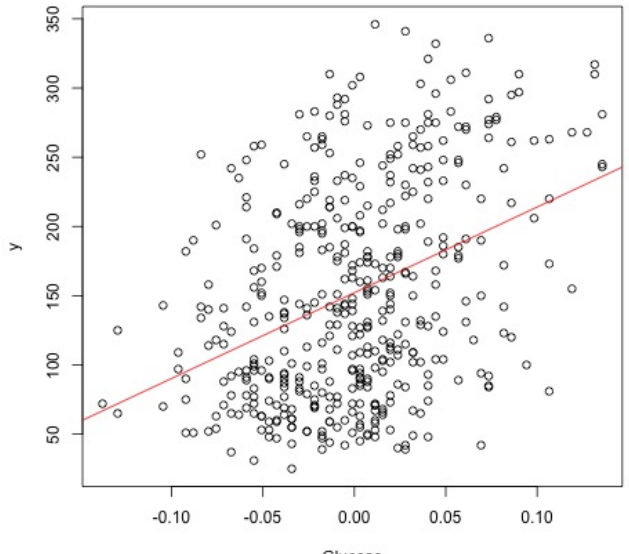
- Basic definition
- Logistic regression and binary outcomes
- Generalised linear models in R

Main aim

Regression models are used to investigate association between

- an outcome variable y
- potential explanatory variables (or predictors) $x = (x_1, x_2, \ldots, x_p)$

The statistical idea is to see if the $x_1, x_2, \ldots, x_p$ can give an adequate description of the variability of the outcome y.



Motivations

1. **Understand** how the predictors affect the outcome.
   - Example: We conduct an observational study focusing on type 2 diabetes as outcome. Our aim is to understand which risk factors are associated with the risk of type 2 diabetes.
2. **Predict** the outcome of new observations, where only the predictors are observed, but not the outcome.
   - Example: We study type 2 diabetes and want to predict disease progression. Our aim is to identify individuals with poor prognosis and improve their treatment.

The linear model

$$y = \alpha + x\beta + \epsilon$$

- y: Outcome, response, dependent variable
  Dimension: $n \times 1$
- x: Regressors, exposures, covariates, input, explanatory, or independent variables
  Dimension: $n \times p$
- ϵ: Residuals, error
- α: Intercept
- β: Regression coefficients

Parameters to estimate:

- α: intercept
  Baseline level, the expected mean value of y when all $x = 0$
- $\beta = (\beta_1, ..., \beta_p)$: vector of regression coefficients
- β_j: regression coefficients of variable x_j
  The expected change in y for a one-unit change in x_j when the other covariates are held constant.

Observed data:

- y: Outcome or response
- x: Regressors, exposures, covariates, input, explanatory or independent variables
  - $i = 1, ..., n$ samples
  - $j = 1, ..., p$ variables

Estimates: Ordinary least squares (OLS)

$$\hat{\beta}_{LS} = \underbrace{(x^t x)^{(-1)}}_{p \times p} \underbrace{x^t}_{p \times n} \underbrace{y}_{n \times 1}$$

- Inversion of $\underbrace{(x^t x)^{(-1)}}_{p \times p}$ requires $x^t x$ to be of full rank (Lecture 2b).
- Alternative representation:

$$\hat{\beta} = \frac{\hat{cov}(x, y)}{\hat{cov}(x)}$$

where the sample covariance is defined as:

- $\hat{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
- $\hat{cov}(x) == \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})$

Example: Diabetes data

- y: quantitative measure of disease progression one year after baseline (vector)
- x: predictor matrix
  - clincal parameters: age, sex, bmi
  - map: blood pressure
  - tc: total cholesterol
  - ldl: low-density lipoprotein
  - hdl: high-density lipoprotein
  - tch: total cholesterol
  - ltg: triglycerides
  - glu: glucose
- $n = 442$: sample size

The `lm()` command in R

- `lm(y ~ age+sex+glu+map+ltg, data = x)`
- Formula `y ~ x1+x2+x3`
  - left of ∼: outcome
  - right of ∼: predictors
- It is also possible to enter a full matrix x (transform by `as.matrix`) as multivariable set of predictors: `y ~ x`
- An intercept is always included, to turn off add `-1`

Interpreting the summary.lm() command

- `summary.lm(lm(y ~ age+sex+glu+map+ltg, data = x))`

```
> summary(lm1)

Call:
lm(formula = y ~ age + sex + glu + map + ltg, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-165.128  -43.025   -5.232   42.446  182.050

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  152.133      2.871  52.990  < 2e-16 ***
age          -54.227     65.854  -0.823  0.41071
sex         -166.066     62.903  -2.640  0.00859 **
glu          175.377     71.671   2.447  0.01480 *
map          426.532     70.342   6.064 2.89e-09 ***
ltg          706.395     70.929   9.959  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.36 on 436 degrees of freedom
Multiple R-squared:  0.3939,    Adjusted R-squared:  0.387
F-statistic: 56.68 on 5 and 436 DF,  p-value: < 2.2e-16
```

Difference between univariable and multivariable regression

- `summary.lm(lm(y ~glu))`

```
> summary(lm0)

Call:
lm(formula = y ~ glu, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-153.069  -57.716   -5.466   54.656  186.839

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  152.133      3.392  44.851   <2e-16 ***
glu          619.223     71.312   8.683   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71.31 on 440 degrees of freedom
Multiple R-squared:  0.1463,    Adjusted R-squared:  0.1444
F-statistic:  75.4 on 1 and 440 DF,  p-value: < 2.2e-16
```

- Reduction of the regression coefficient from 619 to 175 after conditioning on other covariates. → attenuation of the effect.

Further estimates

- Weighted least squares

$$\hat{\beta}_{WLS} = \underbrace{(x^t w x)^{(-1)}}_{p \times p} \underbrace{x^t}_{p \times n} \underbrace{w}_{n \times n} \underbrace{y}_{n \times 1}$$

  where w is a $n \times n$ diagonal weight matrix.
- Maximum likelihood (Lecture 1b)
- Bayesian linear regression (Module: Bayesian Statistics)

Fitted values and residuals

- Fitted values

$$\hat{y} = x\hat{\beta} = \underbrace{x(x^t x)^{-1} x^t}_{h} y$$

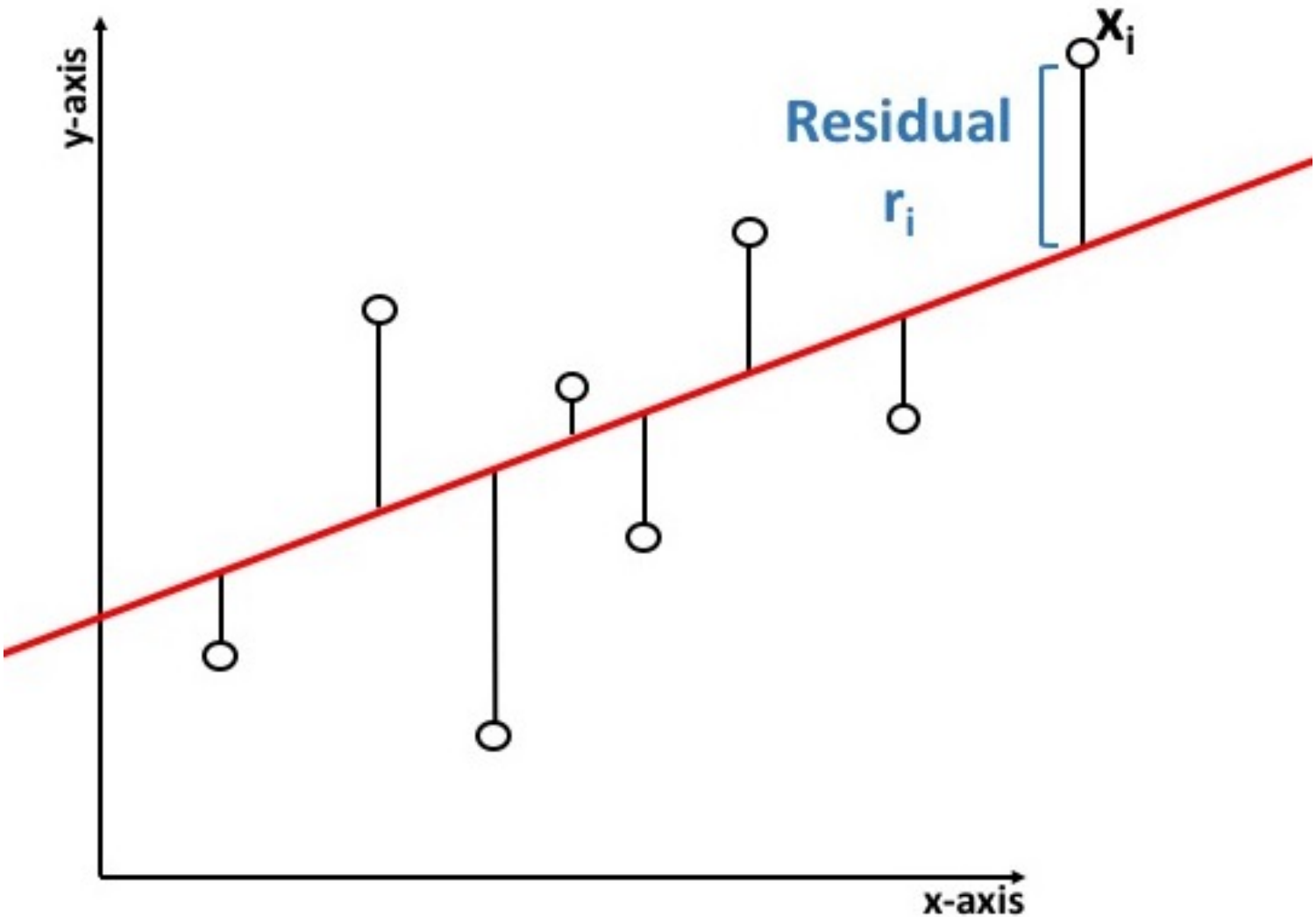
- Hat matrix h
- Residuals are the difference between the fitted values (predicted by the model) and the actual observed outcome.

$$r_i = \hat{y}_i - y_i$$

- The residuals are a vector $r = (r_1, ..., r_n)$ of length n.

Residuals are an important quantity for model diagnostics.

Fitted values and residuals



lm(): Fitted values and residuals

- First fit a linear model and save it in the object lm0
  `lm0 = lm(y~x)`
- The linear model object lm0 contains
  - Regression coefficients: `lm0$coefficients`

```
> lm0$coefficients
(Intercept)         glu
   152.1335    619.2228
```

  - Fitted values: `lm0$fitted.values`
  - Residuals: `lm0$residuals`

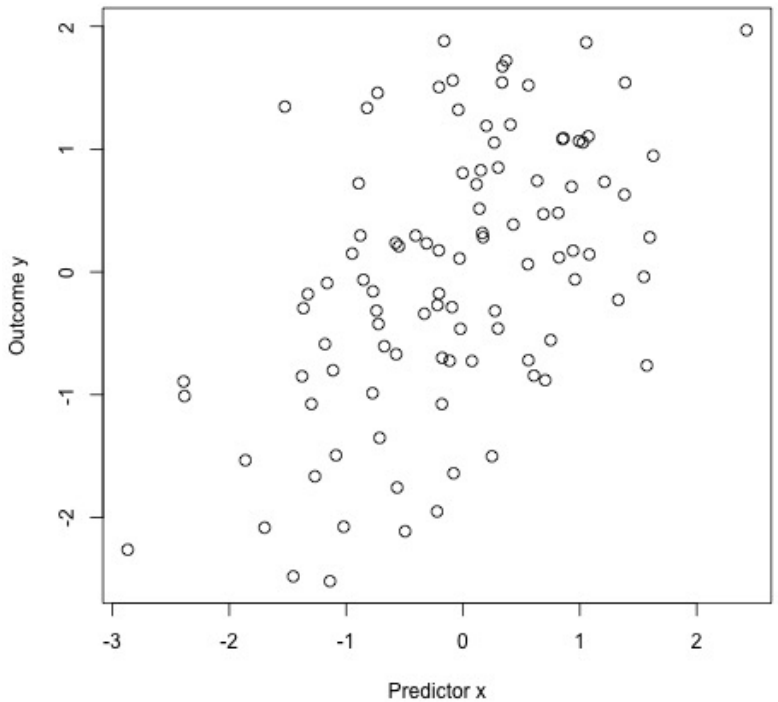
Assumptions

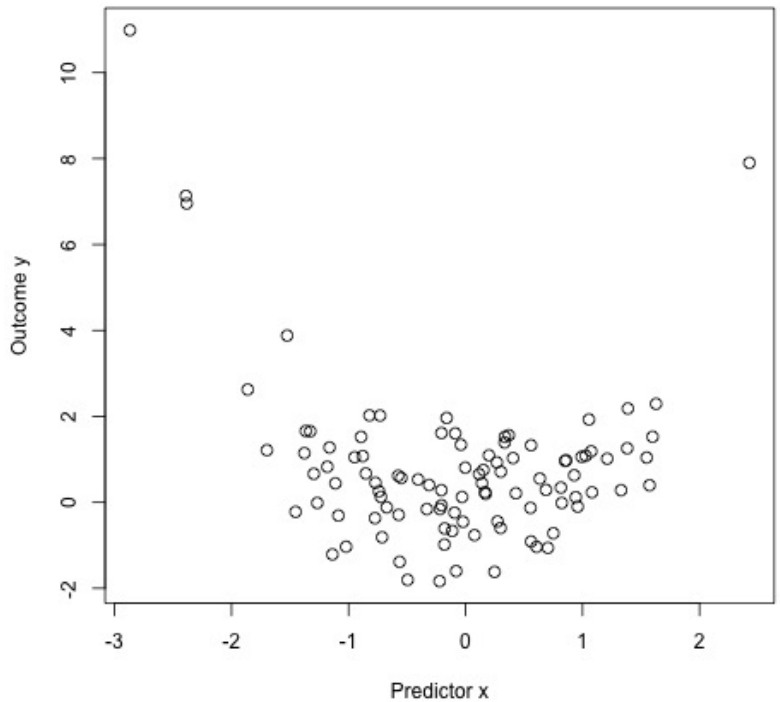
I. Linearity: There is a linear relationship between x and y
II. Weak exogeneity: The predictors x are viewed as fixed variables; there is no measurement error on x.
III. Constant variance (homoscedasticity):
IV. No perfect multicollinearity: No predictor can be expressed as a linear combination of the other predictors (Lecture 2b).
V. Independent errors: The residuals are uncorrelated (e.g. in time-series the error of time point t will depend on the error of time point $t-1$) and independent of x.

Further assumptions

- Normal-distributed errors:
  The residuals are normal-distributed.
  Note: This is not required for the OLS estimate, but for the Maximum Likelihood estimation.
- Outlier: observation point that is distant from other observations.
  It is recommended to check the data for outliers, which can arise because of many reasons:
  - Measurement error (remove)
  - Errors in the pre-processing steps (fix or remove)
  - 'True' biological outliers (follow-up)
- Cook's distance

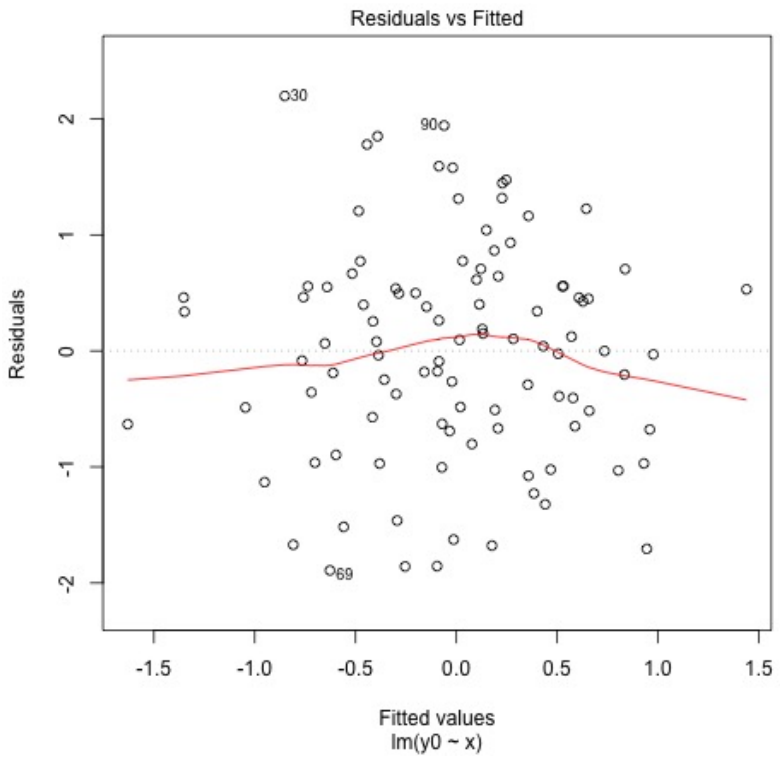
Diagnostic plots: Linear relationship

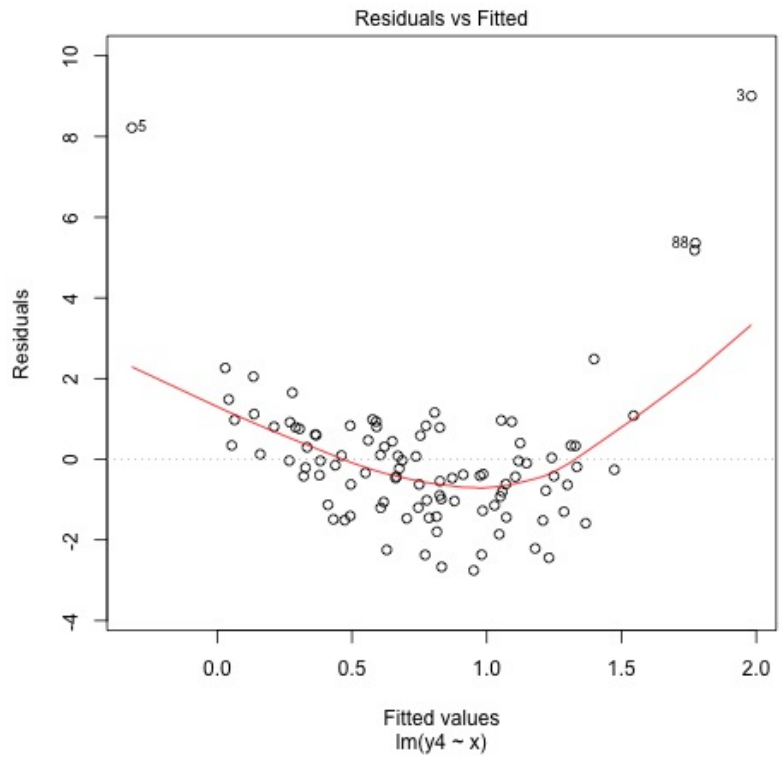




- Scatterplot of y against x

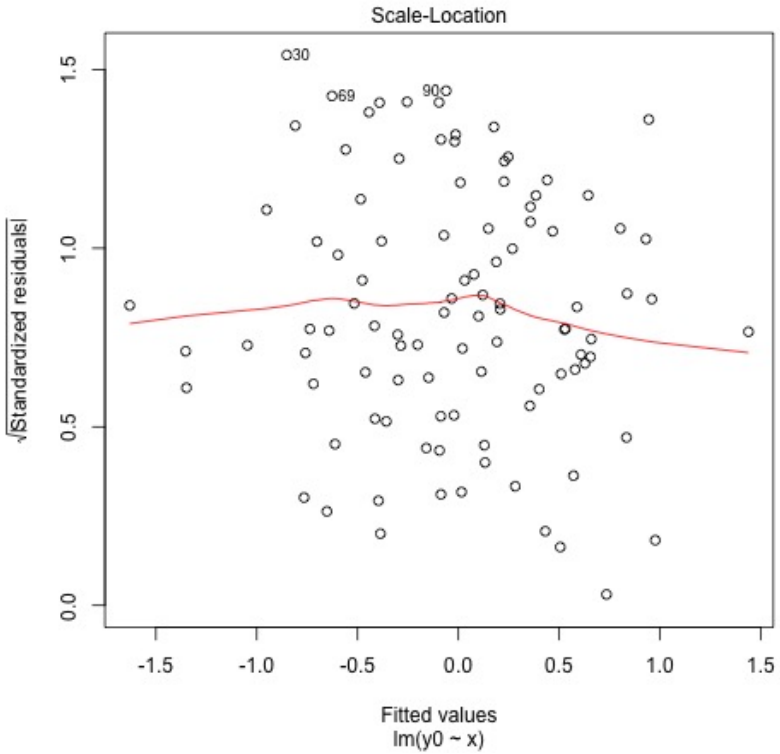
Diagnostic plots: Linear relationship

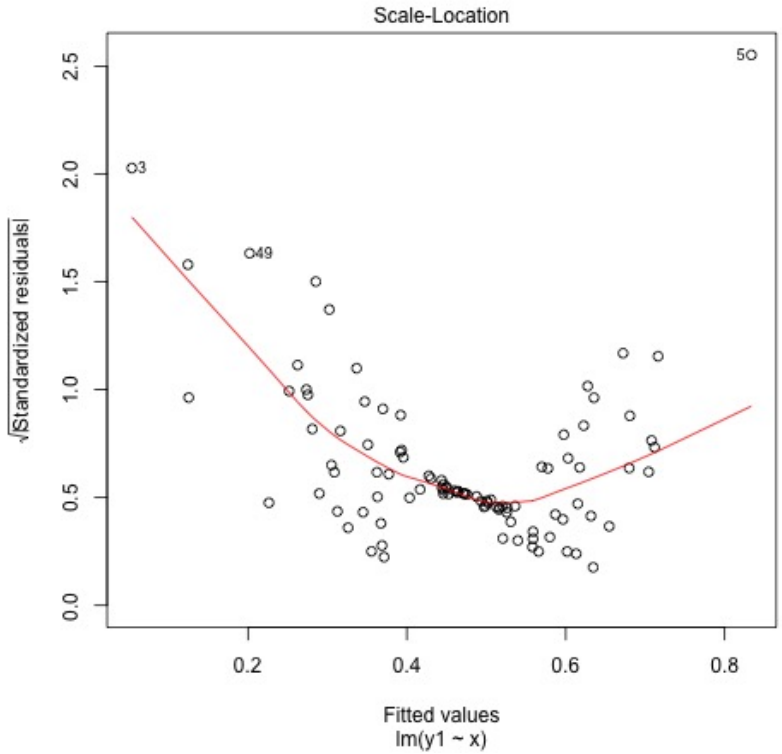




- Scatterplot of residuals (y-axis) against fitted values (x-axis)

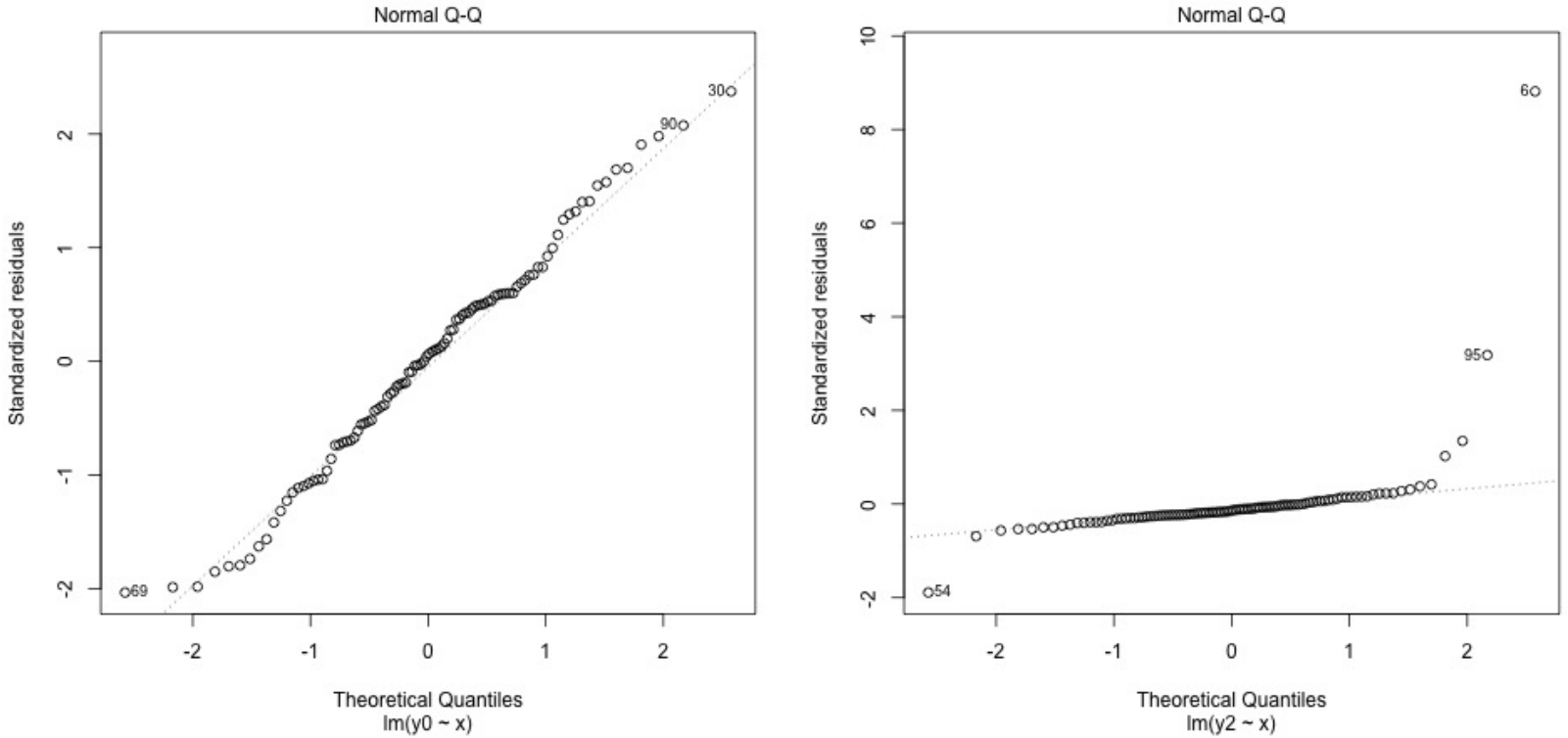
Diagnostic plots: Homoscedasticity





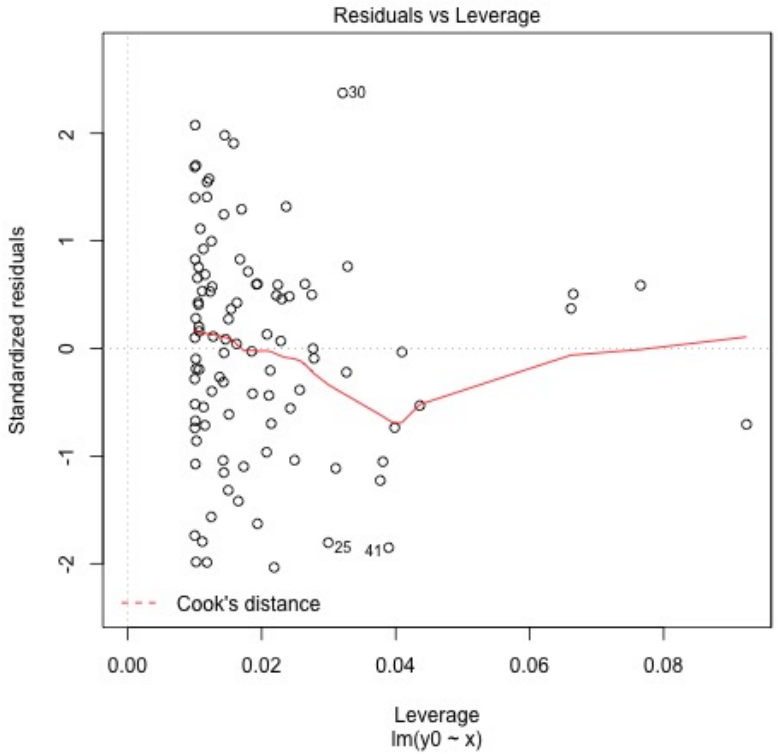
- Scatterplot of standardised residuals (y-axis) against fitted values (x-axis)

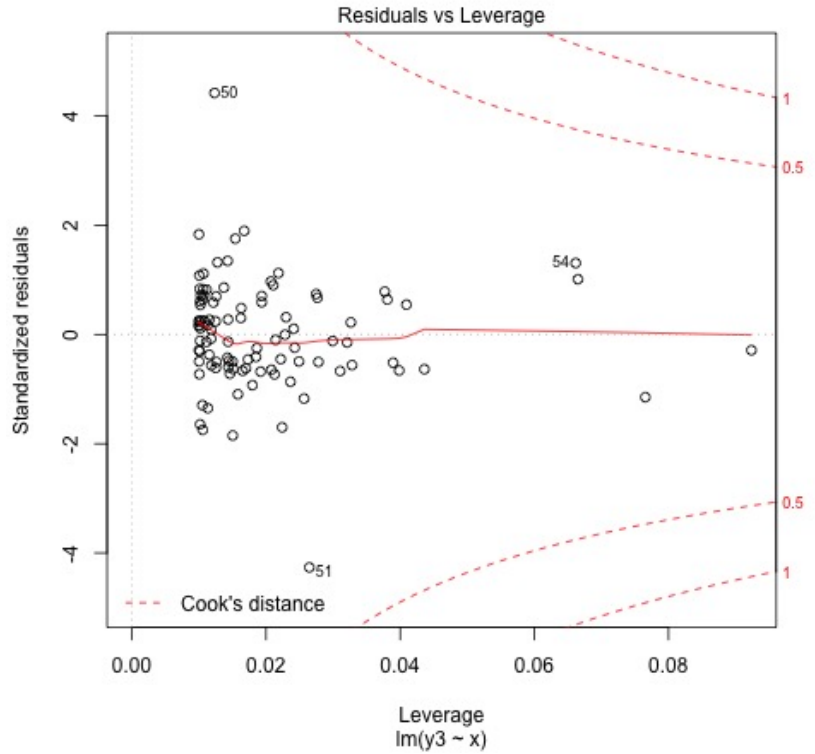
Diagnostic plots: Normal-distribution of residuals



- Q-Q plots of observed residuals (y-axis) against theoretical values under the Normal distribution (x-axis)

Diagnostic plots: Outliers





- Scatterplot of standardised residuals (y-axis) against Cook's distance (x-axis)

lm(): Diagnostics

- Linear relationship (Scatterplot of y against x)
  `plot(x,y)`
  `abline(lm0, col='red')`
- Linear relationship (Residuals against fitted values)
  `plot(lm0, which=1)`
- Homoscedasticity (Standardised residuals against fitted values)
  `plot(lm0, which=3)`
- Normal-distribution of residuals (Q-Q plots of observed residuals against theoretical values under the Normal distribution)
  `plot(lm0, which=2)`
- Outliers (Standardised residuals against Cook's distance)
  `plot(lm0, which=5)`

Prediction using linear models

1. Assume we have a database with n type 2 diabetes cases, where we have measured the following data:
   - y: quantitative measure of disease progression one year after baseline (vector)
   - x: predictor matrix including clinical data (age, sex, bmi), blood pressure and triglycerides

   This is our training data y_train and x_train.
2. For a new case we only have the predictor matrix x_{new}, but not y_{new}.
3. Goal: For each new type 2 diabetes case we want to predict y_{new}, his/her progression one year later.

lm(): Predictions

1. Use the linear model to learn a prediction rule from the training data, where both x and y are observed on the same individuals.
   ```
   lm_train =
   lm(formula=y_train~age+sex+bmi+map+ltg,
   data=x_train)
   ```
2. Predict the outcome based on the prediction rule and the predictors of the new data.
   ```
   predict.lm(lm_train,x_new)
   ```

Generalised linear model (GLM)

- Flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

The GLM consists of three elements:

1. A probability distribution from the exponential family.
2. A linear predictor $\eta = x\beta$.
3. A link function g such that $E(y) = \mu = g^{-1}(\eta)$.

Exponential families

- An exponential family is a set of probability distributions of the following form

$$f_x(x \mid \theta) = h(x) \exp\{\eta(\theta) \times T(x) - A(\theta)\}$$

where

- θ is our parameter of interest
- $T(x)$ is a sufficient statistic.
- $\eta(\theta)$ is the natural parameter or link function.

Exponential families

- The table below gives some of the regression models that are accommodated in this framework.
- Each of these distributions has a location parameter, e.g. μ for the Gaussian, p for the Bernoulli and Binomial.
- The natural link function between the location parameter and the linear predictor can be derived from the mathematical form of the distribution.

Response	Distribution	$E(y)$	Link (g)
Continuous	Gaussian	μ	identity
Dichotomous	Bernoulli	p	$\log\text{it}$
Counts	Binomial	p	$\log\text{it}$
Counts	Poisson	λ	log

Gaussian distribution as exponential distribution

Gaussian distribution with unknown μ, but known σ

$$\begin{aligned} f_\sigma(x \mid \mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\} \end{aligned}$$

- $\theta = \mu$
- $h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2}\}$
- $T(x) = \frac{x}{\sigma}$
- $\eta(\mu) = \frac{\mu}{\sigma}$
- $A(\mu) = \frac{\mu^2}{2\sigma^2}$

Logistic regression and binary outcomes

Binomial distribution with known number of trials n, but unknown probability p

$$\begin{aligned} f(x \mid p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} \exp\{x \log(\frac{p}{1-p}) + n \log(1-p)\} \end{aligned}$$

- $\theta = p$
- $h(x) = \binom{n}{x}$
- $T(x) = x$
- $\eta(p) = \log(\frac{p}{1-p})$
- $A(p) = -n \log(1-p)$

Logistic regression and binary outcomes

Formulate model: Three elements

1. Error distribution for response variable
2. Linear predictor
3. Link function

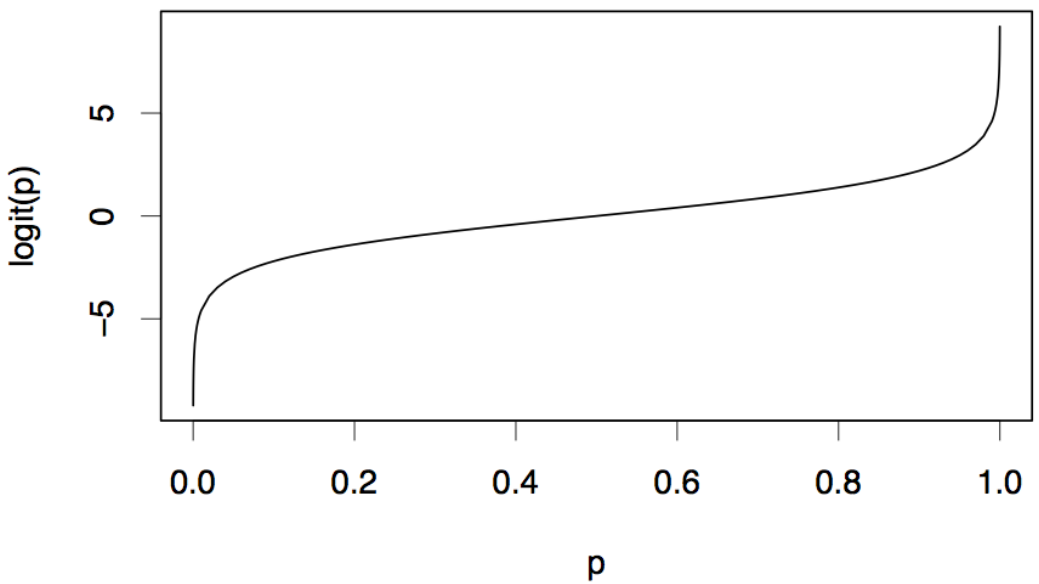
The three elements of the logistic regression model are

1. The Bernoulli probability distribution modelling the data: $\mathbb{P}(y_i = 1 \mid x_i) = p_i$
2. The linear predictor: $\alpha + \sum_{j=1}^{p} \beta_j x_{ij}$
3. The link function g associating the mean of y, $\mathbb{P}(y_i = 1 \mid x_i)$ to the linear predictor: here the link is the logistic link as we set $g(\mathbb{P}(y_i = 1 \mid x_i) = \mathrm{logit}(\mathrm{p_i}) = \beta_0 + \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_{\mathrm{j}} \mathrm{x_{ij}}$

Logistic function

Transform the linear predictor to lie in the Interval [0,1] for probabilities.

$$\text{logit}(p_i) = \log(p_i/(1-p_i)) = \alpha + \beta x_i$$



- Interpretation: The regression coefficient β in logistic regression represents the log odds ratio between $y = 0$ and $y = 1$.

glm(): in R

- GLMs can be called in R just as linear models.

`glm(y_binary ~ age+sex+bmi+map+ltg, data = x, family=binomial)`

```
> summary(glm_out)

Call:
glm(formula = y_binary ~ age + sex + bmi + map + ltg, family = binomial,
    data = x)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9203  -0.5727  -0.2611   0.3643   2.9926

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6505     0.1753  -9.417  < 2e-16 ***
age          -1.6660     3.3488  -0.497    0.619
sex          -2.2971     3.0168  -0.761    0.446
bmi          21.2383     3.5556   5.973 2.33e-09 ***
map          13.3619     3.3562   3.981 6.85e-05 ***
ltg          22.2722     3.6066   6.175 6.60e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

glm(): in R

- Different types of exponential families can be called using the `family` option:
  - binomial(link = 'logit')
  - gaussian(link = 'identity')
  - Gamma(link = 'inverse')
  - inverse.gaussian(link = '$1/\mu^2$')
  - poisson(link = 'log')
- There are similar return values as for the lm function:
  - coefficients
  - residuals
  - fitted.values
  - linear.predictors: the linear fit on link scale

Take away: Linear models

- Motivation why to use linear models
  (To understand and to predict)
- Model fit using ordinary least squares
- Interpretation of the regression coefficients
- Residuals and fitted values
- Model diagnostics
- Using the linear model to predict

Take away: Generalised linear models

The model formulation in GLMs consists of three elements:

1. Error distribution for response variable
2. Linear predictor
3. Link function

Most common data types can be modeled using GLMs

- Continuous $\rightarrow$ Gaussian distribution
- Dichotomous or binary $\rightarrow$ Bernoulli distribution
- Counts $\rightarrow$ Poisson or Binomial (with known number of trials) distribution

Outlook: Next week

Practical:

- Linear model and diagnostics
- Prediction using the linear model
- Fitting and interpreting GLMs

Lectures:

- Variable ranking and variable selection
- How does correlation impact the analysis?
- High-dimensional data analysis
- Multiple testing and false discovery rates