Linear and logistic regression models

Peter Diggle & Emanuele Giorgi

Lancaster Medical School, Lancaster University, Lancaster, UK

Model-based geostatistics: geospatial statistical methods for public health applications, 5-9 October 2015

Overview

1. Linear regression.
   - Assumptions.
   - Least squares.
   - Diagnostic checks.
2. Logistic regression.
   - Assumptions.
   - Maximum Likelihood estimation.
   - Binning with binary outcomes.

Linear model

- $Y_i =$ response variable (random variable).
- $y_i =$ response variable (observations).

Linear model

- Y_i = response variable (random variable).
- y_i = response variable (observations).
- d_i = explanatory variables (or covariates).

Linear model

$$Y_i = \beta_1 + \beta_2 d_i + Z_i, Z_i \sim N(0, \sigma^2) \text{ i.i.d.}$$

Question

Which of these models is linear?

(a) $Y_i = \beta_1 d_i^{\beta_2} e^{Z_i}$

(b) $\log Y_i = \beta_1 + \beta_2 d_i + Z_i$

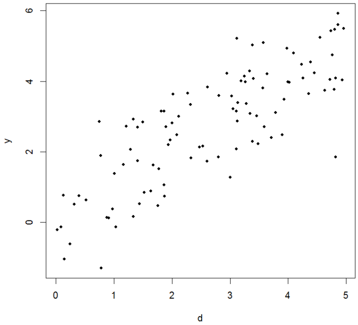
(c) $Y_i = \beta_1 + \beta_2 \log d_i$

Least squares

How to estimate β_1 and β_2?

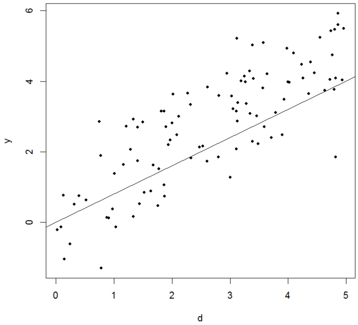
Least squares

How to estimate β_1 and β_2?



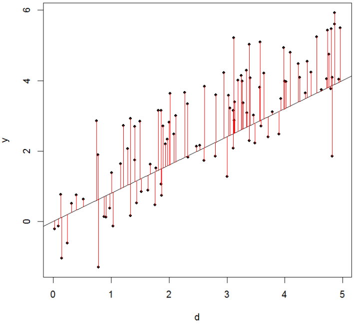
Least squares

How to estimate β_1 and β_2?



Least squares

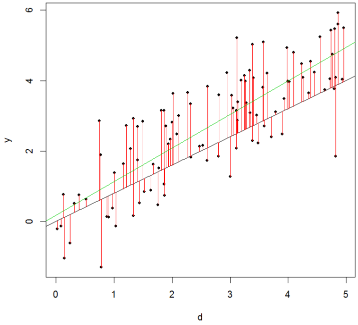
How to estimate β_1 and β_2?



$$RSS = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 d_i)^2$$

Least squares

How to estimate β_1 and β_2?



$$RSS = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 d_i)^2$$

Multiple regression

$$Y_i = d_i^\top \beta + Z_i = \sum_{j=1}^{p} \beta_j d_{ij} + Z_i.$$

Multiple regression

$$Y_i = d_i^\top \beta + Z_i = \sum_{j=1}^{p} \beta_j d_{ij} + Z_i.$$

- $\hat{\beta} = (D^\top D)^{-1} D^\top y.$
- $\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - d_i^\top \hat{\beta})^2 / (n - p)$
- $\hat{\beta} \sim N(\beta, \sigma^2 (D^\top D)^{-1})$

Multiple regression

$$Y_i = d_i^\top \beta + Z_i = \sum_{j=1}^{p} \beta_j d_{ij} + Z_i.$$

- $\hat{\beta} = (D^\top D)^{-1} D^\top y.$
- $\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - d_i^\top \hat{\beta})^2 / (n - p)$
- $\hat{\beta} \sim N(\beta, \sigma^2 (D^\top D)^{-1})$
- $R^2 = 1 - \frac{RSS_p}{RSS_1}.$

Multiple regression

$$Y_i = d_i^\top \beta + Z_i = \sum_{j=1}^{p} \beta_j d_{ij} + Z_i.$$

- $\hat{\beta} = (D^\top D)^{-1} D^\top y.$
- $\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - d_i^\top \hat{\beta})^2 / (n - p)$
- $\hat{\beta} \sim N(\beta, \sigma^2 (D^\top D)^{-1})$
- $R^2 = 1 - \frac{RSS_p}{RSS_1}.$
- $Y_i = \sum_{j=1}^{p} \beta_j d_{ij} + \sum_{k=1}^{q} \beta_k d_{ik} + Z_i$. How to test $\beta_k = 0$ against $\beta_k \neq 0$ for $k = 1, \ldots, q$?

Multiple regression

$$Y_i = d_i^\top \beta + Z_i = \sum_{j=1}^{p} \beta_j d_{ij} + Z_i.$$

- $\hat{\beta} = (D^\top D)^{-1} D^\top y.$
- $\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - d_i^\top \hat{\beta})^2 / (n - p)$
- $\hat{\beta} \sim N(\beta, \sigma^2 (D^\top D)^{-1})$
- $R^2 = 1 - \frac{RSS_p}{RSS_1}.$
- $Y_i = \sum_{j=1}^{p} \beta_j d_{ij} + \sum_{k=1}^{q} \beta_k d_{ik} + Z_i$. How to test $\beta_k = 0$ against $\beta_k \neq 0$ for $k = 1, \ldots, q$?

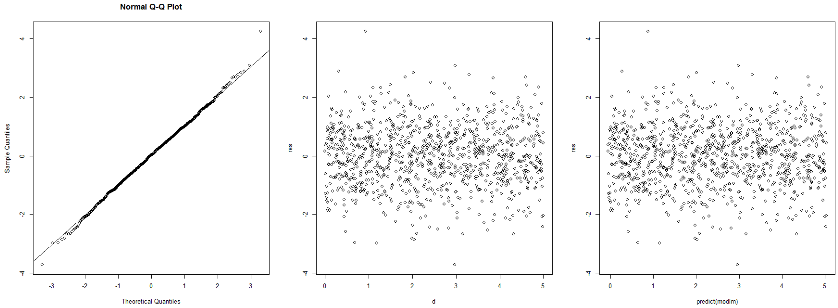
$$\frac{(RSS_p - RSS_{p+q})/q}{RSS_{p+q}/(n-p-q)} \sim F_{q,n-p-q}$$

Diagnostic checks

- $\hat{e}_i = y_i - d_i\hat{\beta}$.

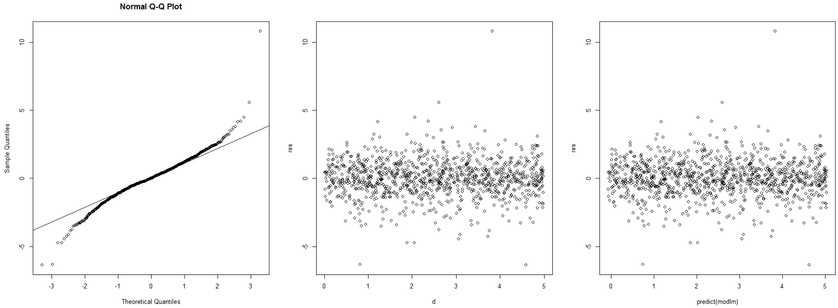
Diagnostic checks

- $\hat{e}_i = y_i - d_i\hat{\beta}$.



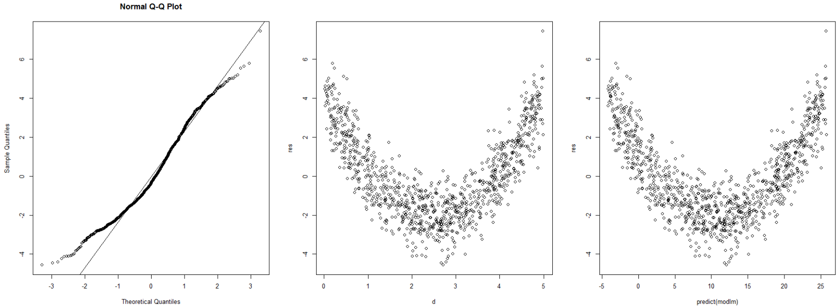
Diagnostic checks

- $\hat{e}_i = y_i - d_i\hat{\beta}.$



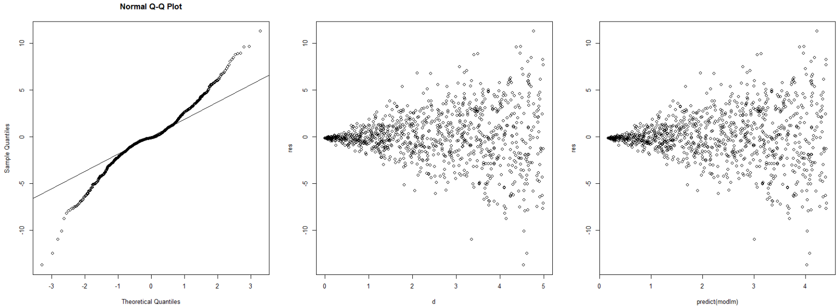
Diagnostic checks

- $\hat{e}_i = y_i - d_i \hat{\beta}.$



Diagnostic checks

- $\hat{e}_i = y_i - d_i\hat{\beta}.$



Logistic regression (1)

- $Y_i \in \{0, \ldots, m\}$ (e.g. number of 6 in m rolls of a die).

Logistic regression (1)

- $Y_i \in \{0, \ldots, m\}$ (e.g. number of 6 in m rolls of a die).
- $P(Y_i = n) = \binom{m}{n} \left(\frac{1}{6}\right)^n \left(\frac{5}{6}\right)^{m-n}$ (balanced die)

Logistic regression (1)

- $Y_i \in \{0, \ldots, m\}$ (e.g. number of 6 in m rolls of a die).
- $P(Y_i = n) = \binom{m}{n} \left(\frac{1}{6}\right)^n \left(\frac{5}{6}\right)^{m-n}$ (balanced die)
- $P(Y_i = n) = \binom{m}{n} p^n (1-p)^{m-n}, p \neq 1/6$ (unbalanced die).

Logistic regression (1)

- $Y_i \in \{0, \ldots, m\}$ (e.g. number of 6 in m rolls of a die).
- $P(Y_i = n) = \binom{m}{n} \left(\frac{1}{6}\right)^n \left(\frac{5}{6}\right)^{m-n}$ (balanced die)
- $P(Y_i = n) = \binom{m}{n} p^n (1-p)^{m-n}, p \neq 1/6$ (unbalanced die).
- $d_i =$ ``number of blue sides on the die''.

Logistic regression (1)

- $Y_i \in \{0, \ldots, m\}$ (e.g. number of 6 in m rolls of a die).
- $P(Y_i = n) = \binom{m}{n} \left(\frac{1}{6}\right)^n \left(\frac{5}{6}\right)^{m-n}$ (balanced die)
- $P(Y_i = n) = \binom{m}{n} p^n (1-p)^{m-n}, p \neq 1/6$ (unbalanced die).
- d_i = ``number of blue sides on the die''.
- Let's roll 200 dice.

		d					
		1	2	3	4	5	6
Y	0	27	23	16	15	8	9
	1	9	8	22	19	20	24

Logistic regression (2)

		d					
		1	2	3	4	5	6
Y	0	0.750	0.742	0.421	0.441	0.286	0.273
	1	0.250	0.258	0.579	0.559	0.714	0.727

Logistic regression (2)

		d					
		1	2	3	4	5	6
Y	0	0.750	0.742	0.421	0.441	0.286	0.273
	1	0.250	0.258	0.579	0.559	0.714	0.727

- Logit-link:

$$\log p_i/(1-p_i) = \beta_1 + \beta_2 d_i$$

Logistic regression (2)

		d					
		1	2	3	4	5	6
Y	0	0.750	0.742	0.421	0.441	0.286	0.273
	1	0.250	0.258	0.579	0.559	0.714	0.727

- Logit-link:

$$\log p_i/(1 - p_i) = \beta_1 + \beta_2 d_i$$

- Balanced die: $\beta_1 = -\log 5$, $\beta_2 = 0$.

Logistic regression (2)

		d					
		1	2	3	4	5	6
Y	0	0.750	0.742	0.421	0.441	0.286	0.273
	1	0.250	0.258	0.579	0.559	0.714	0.727

- Logit-link:

$$\log p_i/(1 - p_i) = \beta_1 + \beta_2 d_i$$

- Balanced die: $\beta_1 = -\log 5$, $\beta_2 = 0$.
- Likelihood:

$$L(\beta_1, \beta_2) = \prod_{i=1}^{200} p_i^{y_i}(1 - p_i)^{(1-y_i)}$$

Logistic regression (2)

		d					
		1	2	3	4	5	6
Y	0	0.750	0.742	0.421	0.441	0.286	0.273
	1	0.250	0.258	0.579	0.559	0.714	0.727

- Logit-link:

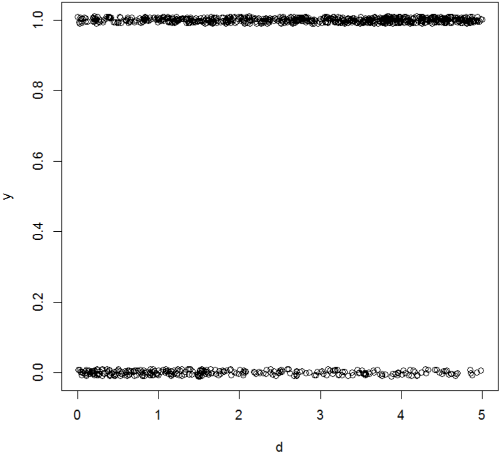
$$\log p_i/(1-p_i) = \beta_1 + \beta_2 d_i$$

- Balanced die: $\beta_1 = -\log 5$, $\beta_2 = 0$.
- Likelihood:

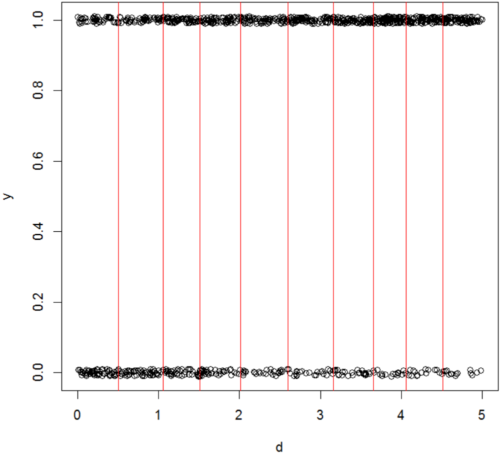
$$L(\beta_1, \beta_2) = \prod_{i=1}^{200} p_i^{y_i}(1-p_i)^{(1-y_i)}$$

- MLE: $(\hat{\beta}_1, \hat{\beta}_2) = (-1.523, 0.458)$
- $\hat{\beta} \sim N(\beta, I(\beta)^{-1})$
- se$(\hat{\beta}_2) = 0.094 \rightarrow$ p-value $= 1.2 \times 10^{-6}$

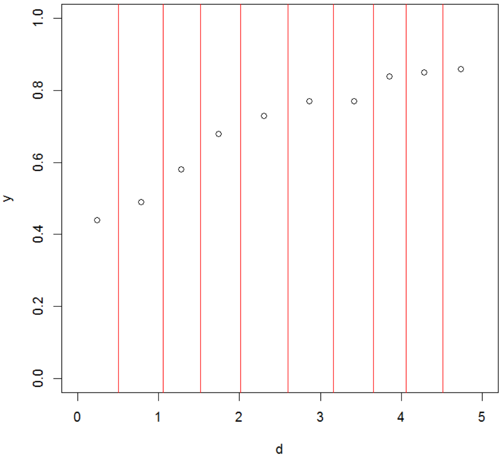
Binning with binary outcome



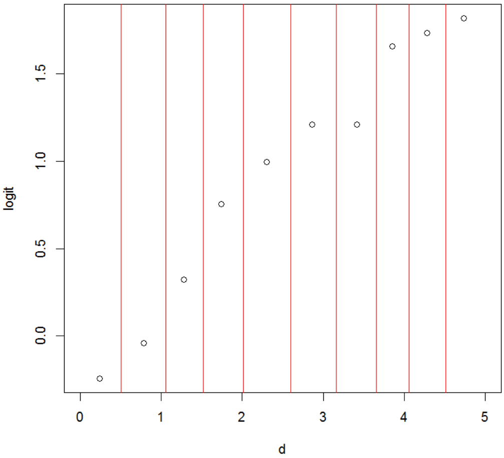
Binning with binary outcome



Binning with binary outcome



Binning with binary outcome



Hypothesis testing and and odds ratio

- $\log\{p_i/(1-p_i)\} = \sum_{j=1}^{p} d_{ij}\beta_j + \sum_{k=1}^{p+q} d_{ik}\beta_k$. How to test $\beta_k = 0$ against $\beta_k \neq 0$?

Hypothesis testing and and odds ratio

- $\log\{p_i/(1-p_i)\} = \sum_{j=1}^{p} d_{ij}\beta_j + \sum_{k=1}^{p+q} d_{ik}\beta_k$. How to test $\beta_k = 0$ against $\beta_k \neq 0$?
- $$2\{\log L(\hat{\beta}_{p+q}) - \log L(\hat{\beta}_p)\} \sim \chi^2_q.$$

Hypothesis testing and and odds ratio

- $\log\{p_i/(1-p_i)\} = \sum_{j=1}^{p} d_{ij}\beta_j + \sum_{k=1}^{p+q} d_{ik}\beta_k$. How to test $\beta_k = 0$ against $\beta_k \neq 0$?
- $$2\{\log L(\hat{\beta}_{p+q}) - \log L(\hat{\beta}_p)\} \sim \chi^2_q.$$
- How to interpret $\exp\{\beta_j\}$?

Hypothesis testing and and odds ratio

- $\log\{p_i/(1-p_i)\} = \sum_{j=1}^{p} d_{ij}\beta_j + \sum_{k=1}^{p+q} d_{ik}\beta_k$. How to test $\beta_k = 0$ against $\beta_k \neq 0$?
- $$2\{\log L(\hat{\beta}_{p+q}) - \log L(\hat{\beta}_p)\} \sim \chi^2_q.$$
- How to interpret $\exp\{\beta_j\}$?
- Confidence intervals of level α for $r_j = \exp\{\beta_j\}$ based on the profile-likelihood:

$$\{r_j : 2[\log L(\hat{r}) - \log L(r_j, \hat{r}_{-j}(r_j))] < \chi^2_{1-\alpha,1}\}$$