Methods of Applied Statistics I

STA2101H F LEC9101

Week 6

October 19 2022

Alison Horst Data Science Art

Today

Start Recording

1. Upcoming events
2. Recap
3. Binary and binomial responses; logistic regresson ELM-2
4. In the News

5. More logistic regression 3rd hour

6. Sections for Project
   - a description of the scientific problem of interest
   - how (and why) the data being analyzed was collected
   - preliminary description of the data (plots and tables)
   - models and analysis
   - summary for a statistician of the analysis and conclusions
   - non-technical summary for a non-statistician of the analysis and conclusions

Upcoming

- Monday October 24 3.30-4.30 : DoSS Seminar Room 9014 (Hydro Building)
- Data Science Seminar Series
- Daniel McDonald, U Chicago
- Markov-Switching State Space Models for Uncovering Musical Interpretation

Recap

- Design of studies: systematic error (bias); random variation; scale of effort; plan of analysis; pre-specified methods and exploration
- unit of study; randomized controlled trials; ecological bias (unit of interpretation $\neq$ unit of study)
- factor variables; analysis of covariance ($=$ interaction between dummy variable and continuous variable)
- observational studies; confounding; support for causality ("Bradford-Hill criteria")

- unit of analysis – “smallest subdivision of the experimental material such that two distinct units might be randomized to different treatments”
    - example: patient in a clinical trial
    - example: plot of land in an agricultural trial
    - example: units of material in a quality control trial

- advantages of randomization?
    - balances other potential influences on responses
    - avoidance of systematic error
    - a systematic difference in response not due to treatment under study

… Recap: randomized experiments

- unit of analysis – “smallest subdivision of the experimental material such that two distinct units might be randomized to different treatments”
  - example: patient in a clinical trial
  - example: plot of land in an agricultural trial
  - example: units of material in a quality control trial
- advantages of randomization?
  - balances other potential influences on responses
  - avoidance of systematic error
  - a systematic difference in response not due to treatment under study
- randomization can make causal interpretation more plausible

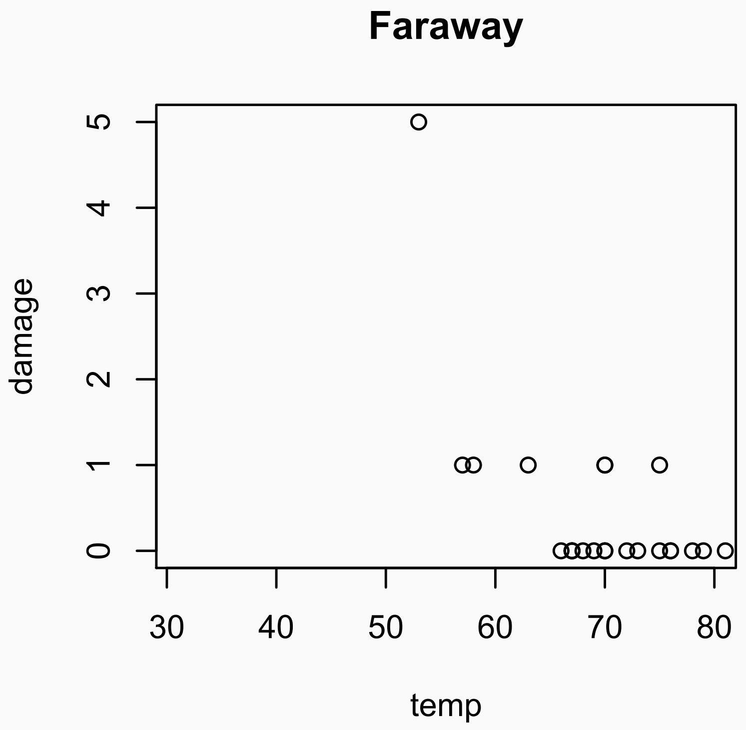
permutation test LM-2 §5.3

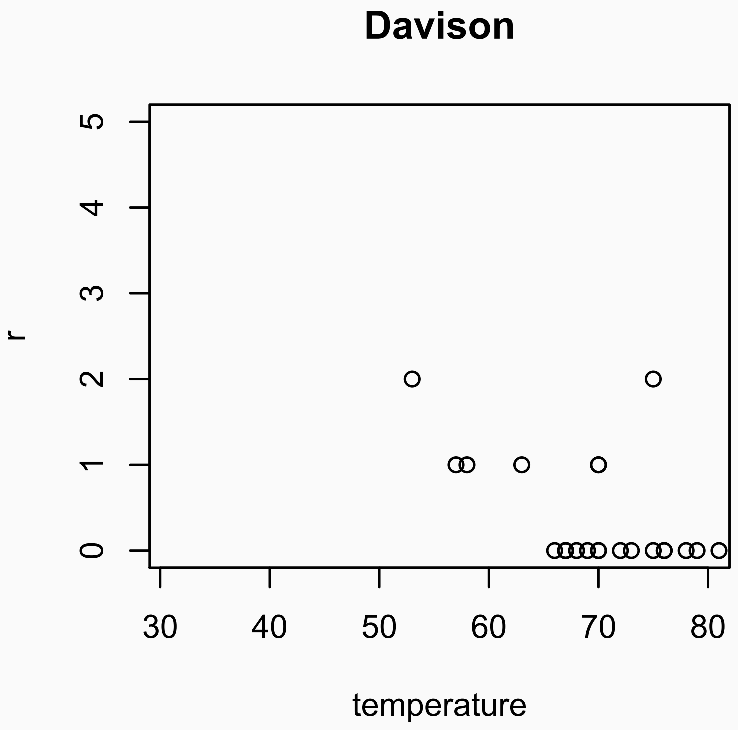
Table 1.3 O-ring thermal distress data. r is the number of field-joint O-rings showing thermal distress out of 6, for a launch at the given temperature (°F) and pressure (pounds per square inch) (Dalal *et al.*, 1989).

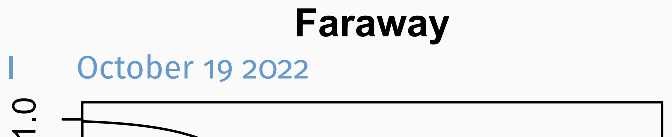
Flight	Date	Number of O-rings with thermal distress, r	Temperature (°F) x_1	Pressure (psi) x_2
1	21/4/81	0	66	50
2	12/11/81	1	70	50
3	22/3/82	0	69	50
5	11/11/82	0	68	50
6	4/4/83	0	67	50
7	18/6/83	0	72	50
8	30/8/83	0	73	100
9	28/11/83	0	70	100
41-B	3/2/84	1	57	200
41-C	6/4/84	1	63	200
41-D	30/8/84	1	70	200
41-G	5/10/84	0	78	200
51-A	8/11/84	0	67	200
51-C	24/1/85	2	53	200
51-D	12/4/85	0	67	200
51-B	29/4/85	0	75	200
51-G	17/6/85	0	70	200
51-F	29/7/85	0	81	200
51-I	27/8/85	0	76	200
51-J	3/10/85	0	79	200
61-A	30/10/85	2	75	200
61-B	26/11/86	0	76	200
61-C	21/1/86	1	58	200

Challenger Shuttle Disaster Jan 28 1986

video


Faraway
damage
0
1
2
3
4
5
30
40
50
60
70
80
temp


Davison
r
0
1
2
3
4
5
30
40
50
60
70
80
temperature


Faraway
1.0


Davison
1.0

Table 1. O-Ring Thermal-Distress Data

		Field			Nozzle				Leak-check pressure	
Flight	Date	Erosion	Blowby	Erosion or blowby	Erosion	Blowby	Erosion or blowby	Joint temperature	Field	Nozzle
1	4/12/81							66	50	50
2	11/12/81	1		1				70	50	50
3	3/22/82							69	50	50
5	11/11/82							68	50	50
6	4/04/83				2		2	67	50	50
7	6/18/83							72	50	50
8	8/30/83							73	100	50
9	11/28/83							70	100	100
41-B	2/03/84	1		1	1		1	57	200	100
41-C	4/06/84	1		1	1		1	63	200	100
41-D	8/30/84	1		1	1	1	1	70	200	100
41-G	10/05/84							78	200	100
51-A	11/08/84							67	200	100
51-C	1/24/85	2, 1*	2	2		2	2	53	200	100
51-D	4/12/85				2		2	67	200	200
51-B	4/29/85				2, 1*	1	2	75	200	100
51-G	6/17/85				2	2	2	70	200	200
51-F	7/29/85				1			81	200	200
51-I	8/27/85				1			76	200	200
51-J	10/03/85							79	200	200
61-A	10/30/85		2	2	1			75	200	200
61-B	11/26/85				2	1	2	76	200	200
61-C	1/12/86	1		1	1	1	2	58	200	200
61-I	1/28/86							31	200	200
	Total	7, 1*	4	9	17, 1*	8	17			

*Secondary O-ring.

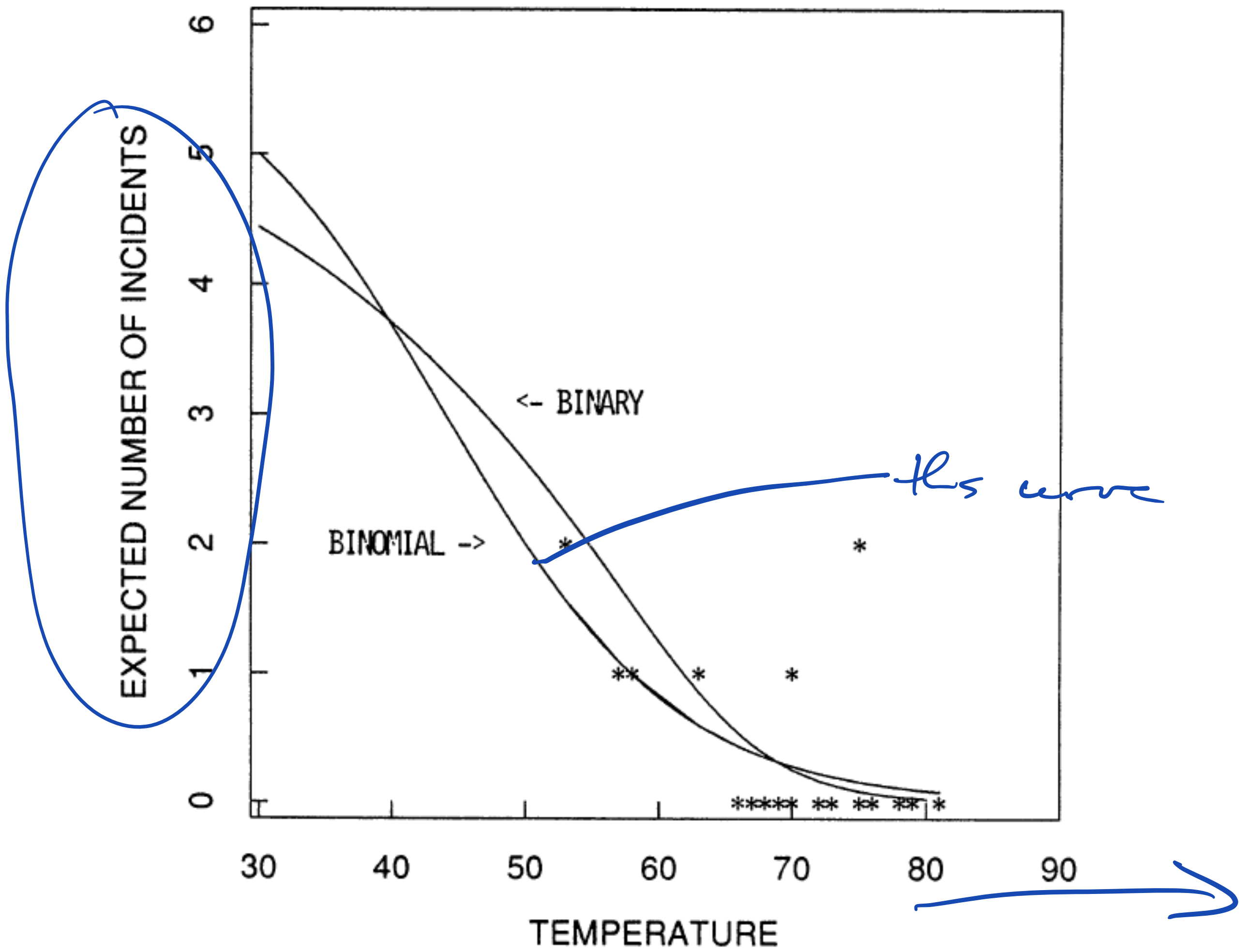


Figure 4. O-Ring Thermal-Distress Data: Field-Joint Primary O-Rings, Binomial-Logit Model, and Binary-Logit Model.

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$

Modelling numbers/proportions of events

Pr of ≥ 1 O-ring failure on launch i

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$
- in general: n_i trials, y_i successes, probability of success p_i

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$
- in general: n_i trials, y_i successes, probability of success p_i
- for regression: associated covariate vector x_i, e.g. temperature , p_i.

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$
- in general: n_i trials, y_i successes, probability of success p_i
- for regression: associated covariate vector x_i, e.g. temperature
- SM uses m_i and r_i instead of n_i and y_i

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$
- in general: n_i trials, y_i successes, probability of success p_i
- for regression: associated covariate vector x_i, e.g. temperature
- SM uses m_i and r_i instead of n_i and y_i
- each y_i could in principle be the sum of n_i independent Bernoulli trials w prob p_i

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$
- in general: n_i trials, y_i successes, probability of success p_i
- for regression: associated covariate vector x_i, e.g. temperature
- SM uses m_i and r_i instead of n_i and y_i
- each y_i could in principle be the sum of n_i independent Bernoulli trials
- each of the n_i trials having the same probability p_i

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$
- in general: n_i trials, y_i successes, probability of success p_i
- for regression: associated covariate vector x_i, e.g. temperature
- SM uses m_i and r_i instead of n_i and y_i
- each y_i could in principle be the sum of n_i independent Bernoulli trials
- each of the n_i trials having the same probability p_i
- with the same covariate vector x_i

ELM 'covariate classes'

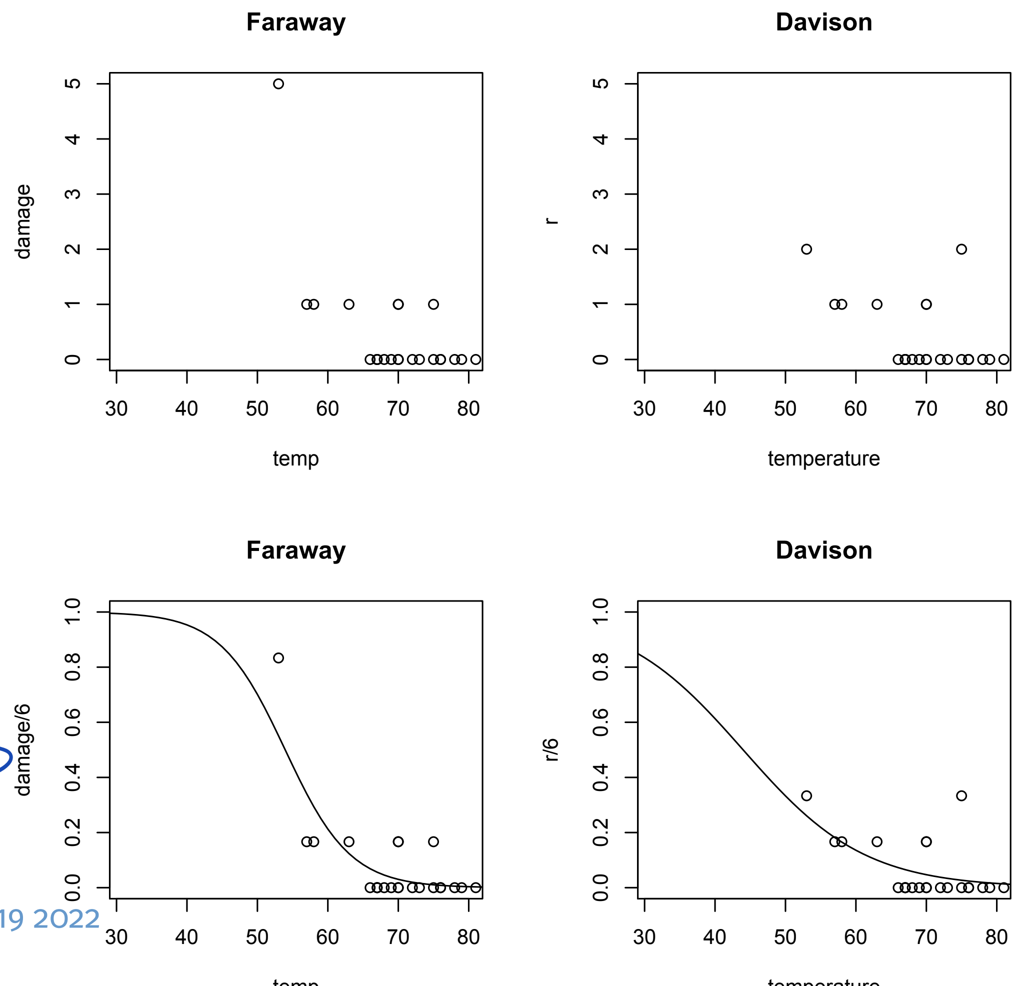
Challenger data: Faraway

```
> library(faraway); data(orings)
> logitmod <- glm(cbind(damage,6-damage) ~ temp, family = binomial, data = orings)
> summary(logitmod)
Call:
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial,
    data = orings)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.66299    3.29626   3.538 0.000403 ***
temp        -0.21623    0.05318  -4.066 4.78e-05 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
```

Challenger data: Davison

```
> library(SMPracticals) # this is for datasets in
                        #Statistical Models by Davison
> data(shuttle) # same example, different name
> shuttle2 <- data.frame(as.matrix(shuttle)) # this is a kludge to avoid
                                   #an error with head(shuttle)
> head(shuttle2)
  m r temperature pressure
1 6 0          66       50
2 6 1          70       50
3 6 0          69       50
4 6 0          68       50
5 6 0          67       50
6 6 0          72       50
> par(mfrow=c(2,2)) # puts 4 plots on a page


> with(orings,plot(temp,damage,main="Faraway",xlim=c(31,80)))
> with(shuttle,plot(temperature,r,main="Davison",xlim=c(31,80),
+ ylim=c(0,5)))
```

Challenger data fits



→ Challenger.html

Regression modelling with binomial

- model:

$$y_i \sim Bin(n_i, p_i)$$

$n_i = 6, i = 1, \ldots, n$

Regression modelling with binomial

- model:

$$y_i \sim Bin(n_i, p_i)$$

$$n_i = 6, i = 1, \ldots, n$$

- regression: link the p_i's through x_i

Regression modelling with binomial

- model:

$$y_i \sim Bin(n_i, p_i)$$

$n_i = 6, i = 1, \ldots, n$

- regression: link the p_i's through x_i
- for example,

$$p_i = \frac{\exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q)}{1 + \exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q))}$$

ilogit$(x_i^T\beta)$

$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + x_{i1}\beta_1 + \ldots + x_{iq}\beta_q$

"log-odds scale"

Regression modelling with binomial

- model:

$$y_i \sim Bin(n_i, p_i)$$

$n_i = 6, i = 1, \ldots, n$

- regression: link the p_i's through x_i
- for example,

$$p_i = \frac{\exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q)}{1 + \exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q))}$$

- more concisely

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

Regression modelling with binomial

- model:

$$y_i \sim Bin(n_i, p_i)$$

$n_i = 6, i = 1, \dots, n$

- regression: link the p_i's through x_i
- for example,

$$p_i = \frac{\exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q)}{1 + \exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q)}$$

- more concisely

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

- $x_i^{\mathrm{T}} = (1, x_{i1}, \dots, x_{iq}); \quad \beta = (\beta_0, \beta_1, \dots, \beta_q)^{\mathrm{T}}$

all vectors are column vectors

… regression modelling with binomial

- Probability of event:

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

… regression modelling with binomial

- Probability of event:

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

- Linear on the logit scale:

$$\log \frac{p_i}{1 - p_i} = x_i^{\mathrm{T}}\beta$$

… regression modelling with binomial

- Probability of event:

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1+\exp(x_i^{\mathrm{T}}\beta)}$$

- Linear on the logit scale:

$$\log \frac{p_i}{1-p_i} = x_i^{\mathrm{T}}\beta$$

- linear predictor:

$$x_i^{\mathrm{T}}\beta = \eta_i$$

$$\hat{\beta} \rightarrow x_i^{\mathrm{T}}\hat{\beta} = \hat{\eta}_i$$

$$\hat{p}_i = \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}}$$

… regression modelling with binomial

- Probability of event:

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1+\exp(x_i^{\mathrm{T}}\beta)}$$

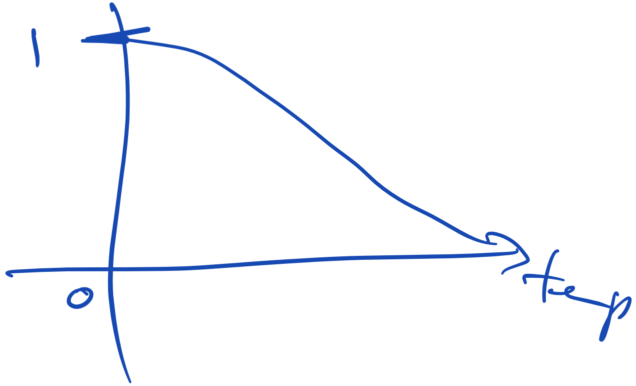
- Linear on the logit scale:

$$\log\frac{p_i}{1-p_i} = x_i^{\mathrm{T}}\beta$$

- linear predictor:

$$x_i^{\mathrm{T}}\beta = \eta_i$$

- p_i is always between 0 and 1



… regression modelling with binomial

- Probability of event:

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

- Linear on the logit scale:

$$\log \frac{p_i}{1 - p_i} = x_i^{\mathrm{T}}\beta$$

- linear predictor:

$$x_i^{\mathrm{T}}\beta = \eta_i$$

- p_i is always between 0 and 1
- see ELM-1 §2.1 for a linear fit

… regression modelling with binomial

- Probability of event:

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1+\exp(x_i^{\mathrm{T}}\beta)}$$

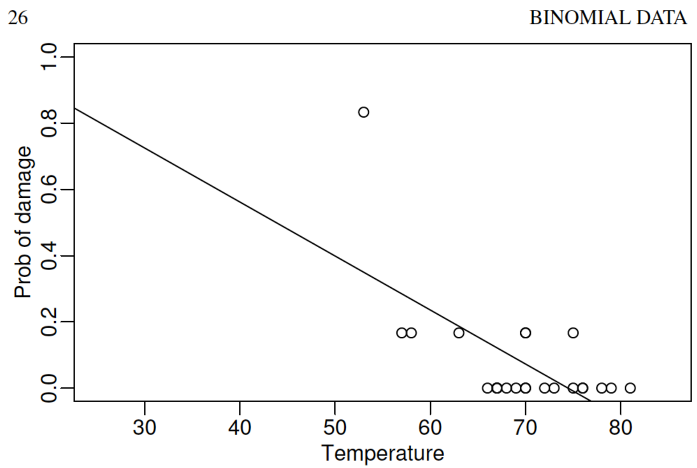
- Linear on the logit scale:

$$\log\frac{p_i}{1-p_i} = x_i^{\mathrm{T}}\beta$$

- linear predictor:

$$x_i^{\mathrm{T}}\beta = \eta_i$$

- p_i is always between 0 and 1
- see ELM-1 §2.1 for a linear fit



… regression modelling with binomial

```
> summary(logitmodcorrect)

Call:
glm(formula = cbind(r, m - r) ~ temperature, family = binomial,  data = shuttle2)


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.08498    3.05247   1.666   0.0957 .
temperature -0.11560    0.04702  -2.458   0.0140 *
```

… regression modelling with binomial

```
> summary(logitmodcorrect)

Call:
glm(formula = cbind(r, m - r) ~ temperature, family = binomial,  data = shuttle2)


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.08498    3.05247   1.666   0.0957 .
temperature -0.11560    0.04702  -2.458   0.0140 *
```

linear predictor:

$$\text{logit}(p_i) = \log(\frac{p_i}{1 - p_i}) = \beta_0 + \beta_1 \texttt{temp}_i$$

$$p_i = \frac{\exp\{\beta_0 + \beta_1 \texttt{temp}_i\}}{1 + \exp\{\beta_0 + \beta_1 \texttt{temp}_i\}}$$

Estimation

- $\ell(\beta; y) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - n_i \log\{1 + \exp(\beta_0 + \beta_1 x_i)\}] + c(y_i, n_i)$

$y_i \sim Bin(n_i, p_i)$ $\quad i = 1, \dots, n$ ind't

$f(y_i; p_i) = \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$ $\quad y_i = 0, \dots, n_i$

$f(y) = \prod_{i=1}^{n} f(y_i; p_i)$

$= \prod_{i=1}^{n} \binom{n_i}{y_i} \left(\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)^{y_i} \left(\frac{1}{1+e^{x_i^T\beta}}\right)^{n_i - y_i} = f(y; \beta)$ fn

$= \text{Likelihood} = L(\beta; y)$

$L(\beta; y) = \prod_{i=1}^{n} \binom{n_i}{y_i} e^{x_i^T\beta y_i} \cdot \left(\frac{1}{1+e^{x_i^T\beta}}\right)^{n_i}$

Estimation

- $\ell(\beta; y) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - n_i \log\{1 + \exp(\beta_0 + \beta_1 x_i)\}]$
- maximum likelihood estimate $\hat{\beta}_0, \hat{\beta}_1$

$\partial \ell(\beta; y)/\partial \beta = 0$

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n} \left(y_i - n_i \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}} \right) = 0$$

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n} \left(y_i x_i - n_i \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}} x_i \right) = 0$$

$$\begin{cases} \sum y_i x_i = \sum n_i \hat{p}_i x_i \\ \sum y_i = \sum n_i \hat{p}_i \end{cases}$$

$\hat{\beta}_0 = \ldots$

$\hat{\beta}_1 = \ldots$

Estimation

- $\ell(\beta; y) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - n_i \log\{1 + \exp(\beta_0 + \beta_1 x_i)\}]$

- maximum likelihood estimate $\hat{\beta}_0$, $\hat{\beta}_1$ $\qquad$ $\partial\ell(\beta; y)/\partial\beta = 0$
-

$$\hat{\beta}_0 = 5.08498, \quad \hat{\beta}_1 = -0.11560 \qquad j(\beta) \equiv -\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}$$

Estimation

- $\ell(\beta; y) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - n_i \log\{1 + \exp(\beta_0 + \beta_1 x_i)\}]$

- maximum likelihood estimate $\hat{\beta}_0, \hat{\beta}_1$ $\quad\quad$ $\partial\ell(\beta; y)/\partial\beta = 0$

- $$\hat{\beta}_0 = 5.08498, \quad \hat{\beta}_1 = -0.11560 \qquad j(\beta) \equiv -\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}$$

- $\text{var}(\hat{\beta}) \doteq j^{-1}(\hat{\beta})$

Estimation

- $\ell(\beta; y) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - n_i \log\{1 + \exp(\beta_0 + \beta_1 x_i)\}]$
- maximum likelihood estimate $\hat\beta_0$, $\hat\beta_1$ $\qquad \partial\ell(\beta; y)/\partial\beta = 0$
- $$\hat\beta_0 = 5.08498, \quad \hat\beta_1 = -0.11560 \qquad j(\beta) \equiv -\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}$$
- $\mathrm{var}(\hat\beta) \doteq j^{-1}(\hat\beta)$

```
> vcov(logitmodcorrect)
            (Intercept)  temperature
(Intercept)   9.3175983 -0.142564339
temperature  -0.1425643  0.002211221
```

Interpretation of estimated coefficients

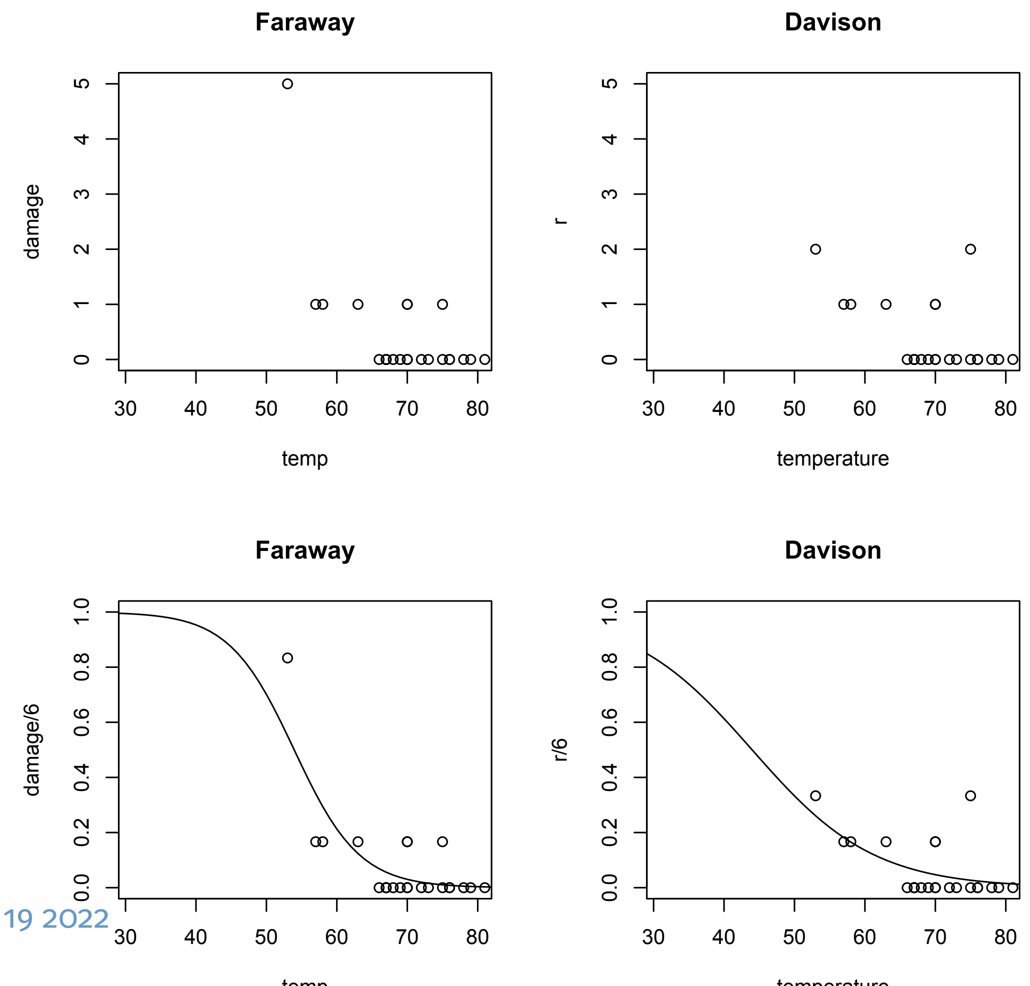
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.08498    3.05247   1.666   0.0957 .
temperature -0.11560    0.04702  -2.458   0.0140 *
```

"a unit increase in temperature is associated with an increase in log-odds of O-ring damage of -0.116"

"an increase in the odds of $\exp(-0.116) = 0.89$" so actually a decrease

" an increase in the probability of ?? depends on the baseline probability

Predicted probabilities



Nested models

- Comparing two models:

Nested models

- Comparing two models:
- likelihood ratio test

$$2\{\ell_A(\hat{\beta}_A) - \ell_B(\hat{\beta}_B)\}$$

  compares the maximized log-likelihood function under model A and model B

Nested models

- Comparing two models:
- likelihood ratio test

$$2\{\ell_A(\hat{\beta}_A) - \ell_B(\hat{\beta}_B)\}$$

  compares the maximized log-likelihood function under model A and model B
- example
  model A: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad \beta_A = (\beta_0, \beta_1, \beta_2)$
  model B: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i}, \quad \beta_B = (\beta_0, \beta_1)$

Nested models

- Comparing two models:
- likelihood ratio test

$$2\{\ell_A(\hat{\beta}_A) - \ell_B(\hat{\beta}_B)\}$$

  compares the maximized log-likelihood function under model A and model B
- example
  model A: $\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad \beta_A = (\beta_0, \beta_1, \beta_2)$
  model B: $\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i}, \quad \beta_B = (\beta_0, \beta_1)$
- when model B is nested in model A, LRT is approximately χ^2_ν distributed, under model B

Nested models

- Comparing two models:
- likelihood ratio test
$$2\{\ell_A(\hat{\beta}_A) - \ell_B(\hat{\beta}_B)\}$$
  compares the maximized log-likelihood function under model A and model B
- example
  model A: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad \beta_A = (\beta_0, \beta_1, \beta_2)$
  model B: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i}, \quad \beta_B = (\beta_0, \beta_1)$
- when model B is nested in model A, LRT is approximately χ^2_ν distributed, under model B
- $\nu = dim(A) - dim(B)$

... nested models

```
> logitmodcorrect2 <- glm(cbind(r,m-r) ~ temperature + pressure, family = binomial, data = shuttle2)

> summary(logitmodcorrect2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.520195   3.486784   0.723   0.4698
temperature -0.098297   0.044890  -2.190   0.0285 *
pressure     0.008484   0.007677   1.105   0.2691
---
    Null deviance: 24.230  on 22  degrees of freedom
Residual deviance: 16.546  on 20  degrees of freedom
AIC: 36.106
Number of Fisher Scoring iterations: 5
```

... nested models

```
> logitmodcorrect2 <- glm(cbind(r,m-r) ~ temperature + pressure, family = binomial, data = shuttle2)

> anova(logitmodcorrect,logitmodcorrect2)
Analysis of Deviance Table

Model 1: cbind(r, m - r) ~ temperature
Model 2: cbind(r, m - r) ~ temperature + pressure
  Resid. Df Resid. Dev Df Deviance
1        21     18.086
2        20     16.546  1   1.5407
```

...nested models

- Model A: $\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{pressure}_i$

...nested models

- Model A: $\operatorname{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{pressure}_i$
- Model B: $\operatorname{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i$

...nested models

- Model A: $\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{pressure}_i$
- Model B: $\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i$
- nested: Model B is obtained by setting $\beta_2 = 0$

...nested models

- Model A: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 \mathtt{temp}_i + \beta_2 \mathtt{pressure}_i$
- Model B: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 \mathtt{temp}_i$
- nested: Model B is obtained by setting $\beta_2 = 0$
- Under Model B, the change in deviance is (approximately) an observation from a χ^2_1

…nested models

- Model A: $\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{pressure}_i$
- Model B: $\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i$
- nested: Model B is obtained by setting $\beta_2 = 0$
- Under Model B, the change in deviance is (approximately) an observation from a χ^2_1
- $\Pr(\chi^2_1 \geq 1.5407) = 0.22$; this is a p-value for testing $H_0 : \beta_2 = 0$

...nested models

- Model A: $\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{pressure}_i$
- Model B: $\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i$
- nested: Model B is obtained by setting $\beta_2 = 0$
- Under Model B, the change in deviance is (approximately) an observation from a χ^2_1
- $\Pr(\chi^2_1 \geq 1.5407) = 0.22$; this is a p-value for testing $H_0 : \beta_2 = 0$
- so is $1 - \Phi\{\frac{\hat{\beta}_2}{\widehat{s.e.}(\hat{\beta}_2)}\} = 1 - \Phi(1.105) = 0.27$

ELM-1 p.30

Inference

- confidence intervals for β_1

Inference

- confidence intervals for β_1

- based on normal approximation: $\hat{\beta}_1 \pm \widehat{\text{s.e.}}(\hat{\beta}_1) * 1.96$

Inference

- confidence intervals for β_1

- based on normal approximation: $\hat{\beta}_1 \pm \widehat{\text{s.e.}}(\hat{\beta}_1) * 1.96$
- `(-0.208, -0.023)`

Inference

- confidence intervals for β_1

- based on normal approximation: $\hat{\beta}_1 \pm \widehat{\text{s.e.}}(\hat{\beta}_1) * 1.96$
- `(-0.208, -0.023)`

- based on profile log-likelihood

$\ell_p(\beta_1)$, details to follow

Inference

- confidence intervals for β_1

- based on normal approximation: $\hat{\beta}_1 \pm \widehat{\text{s.e.}}(\hat{\beta}_1) * 1.96$
- `(-0.208, -0.023)`

- based on profile log-likelihood

$\ell_p(\beta_1)$, details to follow

- `confint(logitmodcorrect):`

  (-0.212~~2262~~, -0.024~~4701~~)

ELM-1 p. 31

Binary data

- each response is $y_i = 0, 1$ instead of $0, 1, \ldots, m_i$
- explanatory variables x_i^T as usual
- same model

$$\mathrm{pr}(y_i = 1 \mid x_i) = p_i(\beta) = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}$$

- example `wcgs` data, ELM-2, Ch.2
- example `nodal` data in `SMPracticals`, SM Example 10.18

$\longrightarrow$ BinaryELM2.Rmd

Binary data

```
> data(wcgs, package="faraway")
> head(wcgs); help(wcgs) #latter not shown
     age height weight sdp dbp chol behave cigs
2001  49     73    150 110  76  225     A2   25
2002  42     70    160 154  84  177     A2   20
2003  42     69    160 110  78  181     B3    0
2004  41     68    152 124  78  132     B4   20
2005  59     70    150 144  86  255     B3   20
2006  44     72    204 150  90  182     B4    0
     dibep chd  typechd timechd   arcus
2001     B  no     none    1664  absent
2002     B  no     none    3071 present
2003     A  no     none    3071  absent
2004     A  no     none    3064  absent
2005     A yes  infdeath    1885 present
```

Binary regression ELM-2 Ch.2

18/10/2022

Binary data

```
data(wcgs, package = "faraway")
dim(wcgs)
```

```
## [1] 3154   13
```

```
head(wcgs) #not run: str(wcgs); plot(wcgs); help(wcgs)
```

```
##      age height weight sdp dbp chol behave cigs dibep chd  typechd timechd
## 2001  49     73    150 110  76  225     A2   25     B  no     none    1664
## 2002  42     70    160 154  84  177     A2   20     B  no     none    3071
## 2003  42     69    160 110  78  181     B3    0     A  no     none    3071
## 2004  41     68    152 124  78  132     B4   20     A  no     none    3064
## 2005  59     70    150 144  86  255     B3   20     A yes infdeath    1885
## 2006  44     72    204 150  90  182     B4    0     A  no     none    3102
##        arcus
## 2001  absent
## 2002 present
## 2003  absent
## 2004  absent
## 2005 present
```

... Binary responses

- where’s the epsilon?

… Binary responses

- where's the epsilon? There isn't one

… Binary responses

- where's the epsilon? There isn't one
- what's the model? It has two parts

... Binary responses

- where's the epsilon? There isn't one
- what's the model? It has two parts
- Regression.

$$\mathbb{E}(y_i) = p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

- Probability distribution.

$$y_i \sim Bernoulli(p_i)$$

… Binary responses

- where's the epsilon? There isn't one
- what's the model? It has two parts
- Regression.

$$\mathbb{E}(y_i) = p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

- Probability distribution.

$$y_i \sim \mathit{Bernoulli}(p_i)$$

- What are these parts in linear regression?
- Regression

$$\mathbb{E}(y_i) = \mu_i = x_i^{\mathrm{T}}\beta$$

- Probability distribution

$$y_i \sim \mathit{Normal}(\mu_i, \sigma^2)$$

Binomial responses

- if you add a lot of Bernoulli's together, all with the same p_i, you get
- how could they have the same p_i in our model?

Binomial responses

- if you add a lot of Bernoulli's together, all with the same p_i, you get
- how could they have the same p_i in our model?
- $p_i = \mathit{function}(x_i^{\mathrm{T}}\beta)$
- different observations with the same p_i are called covariate classes
- Example 10.18 in SM – Table 10.8 has 23 rows of binomials

  sample sizes vary from 1 to 6
- `data(nodal)` in `library(SMPracticals)` has 53 rows of binary observations

Binomial responses

- if you add a lot of Bernoulli's together, all with the same p_i, you get
- how could they have the same p_i in our model?
- $p_i = function(x_i^{\mathrm{T}}\beta)$
- different observations with the same p_i are called covariate classes
- Example 10.18 in SM – Table 10.8 has 23 rows of binomials
  sample sizes vary from 1 to 6
- `data(nodal)` in `library(SMPracticals)` has 53 rows of binary observations
- `R` expects `cbind(r, m-r)` in `glm` with binomial data
- but if all observations are binary you can get away with `r` only
- see `?family` (check `Details`)
- you can also specify proportions y_i/n_i, but then you need to use `weights`

Binomial/Binary

Table 10.8 Data on nodal involvement (Brown, 1980).

m	r	age	stage	grade	xray	acid
6	5	0	1	1	1	1
6	1	0	0	0	0	1
4	0	1	1	1	0	0
4	2	1	1	0	0	1
4	0	0	0	0	0	0
3	2	0	1	1	0	1
3	1	1	1	0	0	0
3	0	1	0	0	0	1
3	0	1	0	0	0	0
2	0	1	0	0	1	0
2	1	0	1	0	0	1
2	1	0	0	1	0	0
1	1	1	1	1	1	1
1	1	1	1	0	1	1
1	1	1	0	1	1	1
1	1	1	0	0	1	1
1	0	1	0	1	0	0
1	1	0	1	1	1	0
1	0	0	1	1	0	0
1	1	0	1	0	1	0
1	1	0	0	1	0	1

Can we predict nodal involvement from other measurements?

Binary regression ELM-2 Ch.2

18/10/2022

Binary data

```
data(wcgs, package = "faraway")
dim(wcgs)
```

```
## [1] 3154   13
```

```
head(wcgs) #not run: str(wcgs); plot(wcgs); help(wcgs)
```

```
##      age height weight sdp dbp chol behave cigs dibep chd  typechd timechd
## 2001  49     73    150 110  76  225     A2   25     B  no     none    1664
## 2002  42     70    160 154  84  177     A2   20     B  no     none    3071
## 2003  42     69    160 110  78  181     B3    0     A  no     none    3071
## 2004  41     68    152 124  78  132     B4   20     A  no     none    3064
## 2005  59     70    150 144  86  255     B3   20     A yes infdeath    1885
## 2006  44     72    204 150  90  182     B4    0     A  no     none    3102
##        arcus
## 2001  absent
## 2002 present
## 2003  absent
## 2004  absent
## 2005 present
## 2006  absent
```

→ BinaryELM2.Rmd

```
sum(is.na(wcgs)) # there are a few missing values; much further sleuthing finds arcus and chol have some missing
 values
```

Inference based on the likelihood function

- model: $y_i \sim f(y_i; \theta), i = 1, \ldots, n$ independent
- joint density: $f(\underline{y}; \theta) = \prod_{i=1}^{n} f(y_i; \theta)$
- likelihood function $L(\theta; \underline{y}) = f(\underline{y}; \theta)$

Inference based on the likelihood function

- model: $y_i \sim f(y_i; \theta), i = 1, \dots, n$ independent
- joint density: $f(\underline{y}; \theta) = \prod_{i=1}^{n} f(y_i; \theta)$
- likelihood function $L(\theta; \underline{y}) = f(\underline{y}; \theta)$

- log-likelihood function $\ell(\theta; \underline{y}) = \log L(\theta; \underline{y}) = \sum_{i=1}^{n} \log f(y_i; \theta)$
- maximum likelihood estimate $\hat{\theta} = \arg\sup \ell(\theta; \underline{y})$; $\ell'(\hat{\theta}) = 0$
- Fisher information $j(\theta) = -\ell''(\theta)$

Inference based on the likelihood function

- model: $y_i \sim f(y_i; \theta), i = 1, \ldots, n$ — independent
- joint density: $f(\underline{y}; \theta) = \prod_{i=1}^{n} f(y_i; \theta)$
- likelihood function $L(\theta; \underline{y}) = f(\underline{y}; \theta)$

- log-likelihood function $\ell(\theta; \underline{y}) = \log L(\theta; \underline{y}) = \sum_{i=1}^{n} \log f(y_i; \theta)$
- maximum likelihood estimate $\hat{\theta} = \arg\sup \ell(\theta; \underline{y})$; $\ell'(\hat{\theta}) = 0$
- Fisher information $j(\theta) = -\ell''(\theta)$

- two theorems:

$$(\hat{\theta} - \theta) j^{1/2}(\hat{\theta}) \xrightarrow{d} N(0, I)$$

asymptotically normal

- likelihood ratio statistic

$$w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi^2_p$$

p is dimension of θ

… Inference based on the likelihood function

- two theorems:

$$(\hat{\theta} - \theta) j^{1/2}(\hat{\theta}) \xrightarrow{d} N(0, I)$$

$$w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi^2_p$$

... Inference based on the likelihood function

- two theorems:

$$
\begin{aligned}
(\hat\theta - \theta) j^{1/2}(\hat\theta) &\xrightarrow{d} N(0, I) \\
w(\theta) = 2\{\ell(\hat\theta) - \ell(\theta)\} &\xrightarrow{d} \chi^2_p
\end{aligned}
$$

- two approximations

$$
\begin{aligned}
\hat\theta_k &\dot\sim N(\{\theta_k, j^{-1}(\hat\theta)_{kk}\} \\
w(\theta) &\dot\sim \chi^2_p
\end{aligned}
$$

... Inference based on the likelihood function

- two theorems:

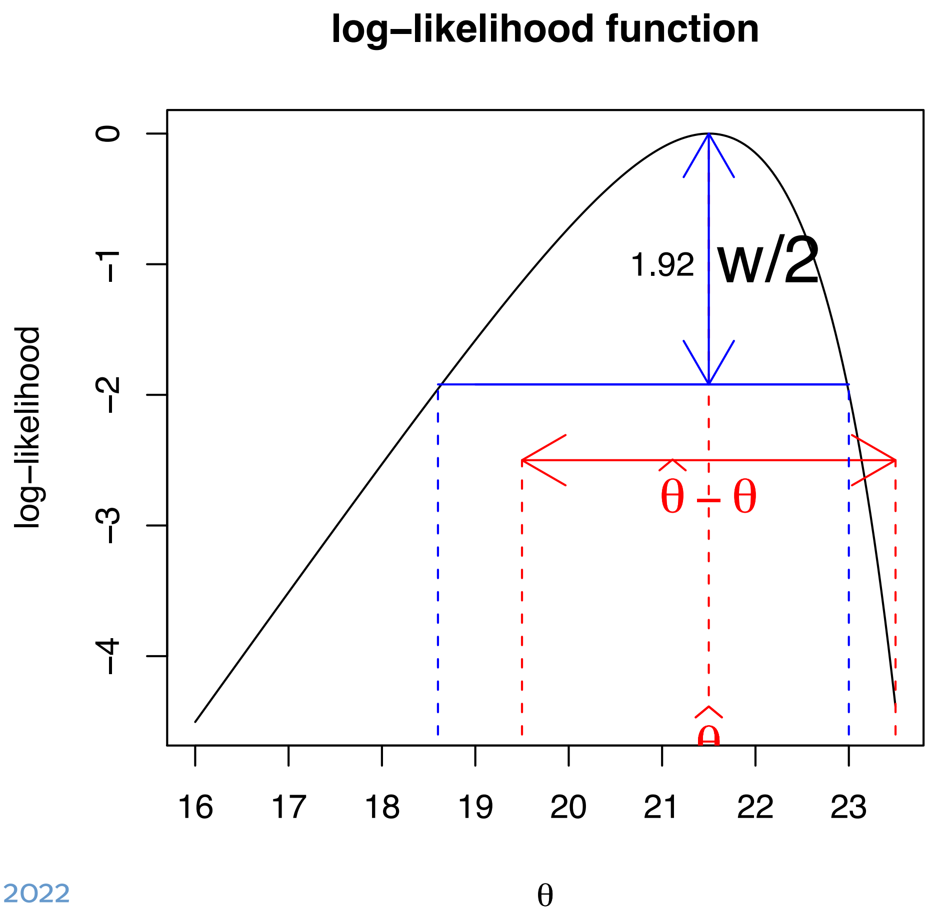
$$
\begin{aligned}
(\hat\theta - \theta) j^{1/2}(\hat\theta) &\stackrel{d}{\to} N(0, I) \\
w(\theta) = 2\{\ell(\hat\theta) - \ell(\theta)\} &\stackrel{d}{\to} \chi^2_p
\end{aligned}
$$

- two approximations

$$
\begin{aligned}
\hat\theta_k &\dot\sim N(\{\theta_k, j^{-1}(\hat\theta)_{kk}\} \\
w(\theta) &\dot\sim \chi^2_p
\end{aligned}
$$

- compare two models using change in likelihood ratio statistic — nested models

… Inference based on the likelihood function



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.08498    3.05247   1.666   0.0957 .
temperature -0.11560    0.04702  -2.458   0.0140 *
```

maximum likelihood estimate $\partial\ell(\beta; y)/\partial\beta = 0$

$$\hat{\beta}_0 = 5.08498, \quad \hat{\beta}_1 = -0.11560 \qquad j(\beta) \equiv -\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}$$

$$\text{var}(\hat{\beta}) \doteq j^{-1}(\hat{\beta})$$

```
> vcov(logitmodcorrect)
            (Intercept)  temperature
(Intercept)   9.3175983 -0.142564339
temperature  -0.1425643  0.002211221
```

Special to the binomial

- likelihood ratio test for logistic model $p_i = p_i(\beta) = \text{expit}(x_i^{\mathrm{T}}\beta), \quad \hat{p}_i = p_i(\hat{\beta})$
- this model is nested in the saturated model $\tilde{p}_i = y_i/n_i$

Special to the binomial

- likelihood ratio test for logistic model $p_i = p_i(\beta) = \text{expit}(x_i^{\mathrm{T}}\beta), \quad \hat{p}_i = p_i(\hat{\beta})$
- this model is nested in the saturated model $\tilde{p}_i = y_i/n_i$

- residual deviance compares fitted model to saturated model

Special to the binomial

- likelihood ratio test for logistic model $p_i = p_i(\beta) = \text{expit}(x_i^{\mathrm{T}}\beta), \quad \hat{p}_i = p_i(\hat{\beta})$
- this model is nested in the saturated model $\tilde{p}_i = y_i/n_i$

- residual deviance compares fitted model to saturated model

- under the fitted model, approximately distributed as χ^2_{n-q}

Special to the binomial

and Poisson

- likelihood ratio test for logistic model $p_i = p_i(\beta) = \text{expit}(x_i^{\mathrm{T}}\beta), \quad \hat{p}_i = p_i(\hat{\beta})$
- this model is nested in the saturated model $\tilde{p}_i = y_i/n_i$

- residual deviance compares fitted model to saturated model

- under the fitted model, approximately distributed as χ^2_{n-q}
  if each n_i "large"

ELM-1 p.29

```
> summary(Ex1018.glm)

(Dispersion parameter for binomial family taken to be 1)

     Null deviance: 40.710  on 22  degrees of freedom
 Residual deviance: 18.069  on 17  degrees of freedom
 AIC: 41.69
```

... example 10.18 variable selection

```
> step(ex1018binom)

Coefficients:
(Intercept)        stage         xray         acid
     -3.052        1.645        1.912        1.638

Degrees of Freedom: 22 Total (i.e. Null);  19 Residual
Null Deviance:^^I    40.71
Residual Deviance: 19.64 ^^IAIC: 39.26
```

– we can drop `age` and `grade` without affecting quality of the fit

– in other words the model can be simplified by setting two regression coefficients to zero

– several mistakes in text on pp. 491,2;

– deviances in Table 10.9 are incorrect as well `http://statwww.epfl.ch/davison/SM/` has corrected version

... example 10.18: variable selection

- `step` implements stepwise regression
- evaluates each fit using AIC = $-2\ell(\hat{\beta}; y) + 2p$
- penalizes models with larger number of parameters

- we can also compare fits by comparing deviances

... example 10.18: variable selection

- `step` implements stepwise regression
- evaluates each fit using AIC = $-2\ell(\hat{\beta}; y) + 2p$
- penalizes models with larger number of parameters

- we can also compare fits by comparing deviances
-
```
> update(ex1018binom, . ~ . - aged - stage)

Call:  glm(formula = cbind(r, m - r) ~ grade + xray + acid, family = binomial,
    data = nodal2)

Coefficients:
(Intercept)        grade         xray         acid
     -2.734        1.420        1.750        1.797

Degrees of Freedom: 22 Total (i.e. Null);  19 Residual
Null Deviance:      40.71
Residual Deviance:  21.28      AIC: 40.9

> deviance(ex1018binom)
[1] 18.06869
> pchisq(21.28-18.07,df=2,lower=F)
[1] 0.2008896
```

AIC

- as terms are added to the model, deviance always decreases
- because log-likelihood function always increases
- similar to residual sum of squares

AIC

- as terms are added to the model, deviance always decreases
- because log-likelihood function always increases
- similar to residual sum of squares

- Akaike Information Criterion penalizes models with more parameters
- $$AIC = 2\{-\ell(\hat{\beta}; y) + p\}$$

SM (4.57)

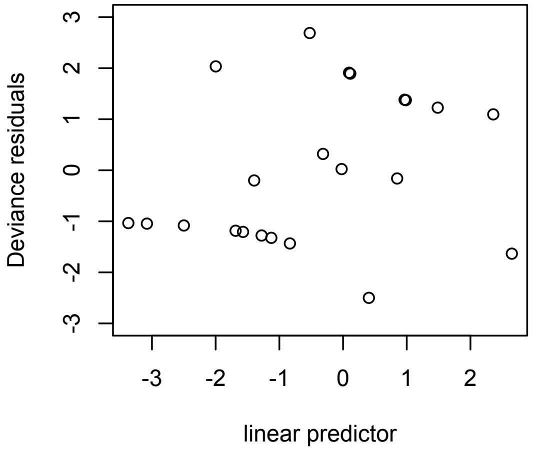
- comparison of two model fits by difference in *AIC*

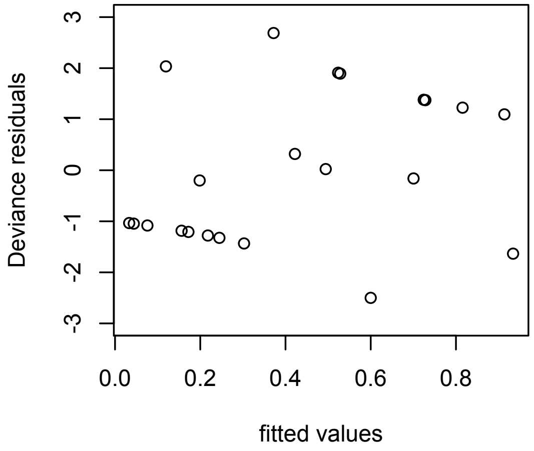
Deviance residuals


```
> summary(ex1018binom)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```

... example 10.18: residuals

```
> summary(ex1018binom)

Call:
glm(formula = cbind(r, m - r) ~ ., family = binomial, data = nodal2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4989  -0.7726  -0.1265   0.7997   1.4351
```





Generalized linear models

`glm` has several options for `family`

```
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")
```

Each of these is a member of the class of generalized linear models

Generalized: distribution of response is not assumed to be normal

Linear: some transformation of $E(y_i)$ is of the form $x_i^T\beta$ link function