Methods of Applied Statistics I

STA2101H F LEC9101

Week 6

October 15 2020

Today

Start Recording

1. A note on polling
2. A note on factorial designs
3. Preliminary Analysis – CD Ch. 5 Left over from last week
4. In the News
5. Introduction logistic regression
6. Summary linear regression (2-3) FELM Ch. 1

- October 19 5.45 – 4.30 Gillian Hadfield
- https://canssiontario.utoronto.ca/?mec-events=dsss_gillian_

"Why Machine Learning has us all talking about bias, privacy, and the end of democracy"

Why were the 2016 polls wrong?

knowable MAGAZINE
FROM ANNUAL REVIEWS

HEALTH & DISEASE | LIVING WORLD | PHYSICAL WORLD | SOCIETY | FOOD & ENVIRONMENT | TECHNOLOGY | THE MIND

SOCIETY

Election polls aren't broken, but they still can't predict the future

People who decide last-minute who they will vote for are one of the reasons elections polls may not accurately predict a winner.

Why were the 2016 polls wrong?

- “post-election analyses of the 2016 US election suggest that national election polling was about as accurate as it has always been

  but not state polls

- “Clinton… won the popular vote by 2 percent, not far from the 3 percent average that the polls found, and within the range of errors seen in previous elections

- “unusual circumstances that magnified typically small errors

- “the issue may be one of expectations. Polls aren’t clairvoyant –especially if an election is close

- “College graduates are more likely to take surveys than people with less education … in 2016 people’s education levels were pretty correlated with how they voted ”

Preliminary Analysis

- Four topics: data auditing, data screening, data cleaning, preliminary analysis
- CD – "data cleaning and screening procedures used should be reported in publications and testimony" ASA Statement on Professional Ethics, 1999
- ASA – Reports the sources and assessed adequacy of the data, accounts for all data considered in a study, and explains the sample(s) actually used. Clearly and fully reports the steps taken to preserve data integrity and valid results.

  updated April 2018.

- Data auditing
  - large studies will often require a quality assurance audit
  - see, e.g. Health Effects Institute statement on quality assurance
  - like financial auditing, follows a prescribed set of checks on the validity and accuracy of the data, ideally going back to the source for independent collection

... Preliminary Analysis: Data screening/cleaning

CD Ch. 5.3

- "pedigree of the data": how was the database prepared; how many people involved; were guidelines set out in advance; how were dates coded; how were missing values coded; are units clear
- sanity check: sample means, standard deviations, minima and maxima
- spreadsheet errors – see Reinhart & Rogoff 2010; Herndon 2013
  - original paper argued that high public debt leads to slow growth
  - attempt by H to reproduce the work discovered 5 rows had been left off the spreadsheet
    Australia, Austria, Belgium, Canada and Denmark
  - there were other disagreements about the correct way to analyse the data

BBC News magazine

- identifying outliers, through plots or diagnostics, and verifying their accuracy

"outliers may arise from rare but accurate observations ... important insight"

- missing data – exploring patterns of missing-ness; deciding whether to eliminate variables with high percentage of missing-ness, et.

ozone layer

- distributions of variables – histograms, boxplots, density estimates
- correlation among pairs of variables – scatterplot matrix
- display of observations in time, or in space e.g. maps
- some principles of visualization
    - axes clearly labelled
    - related graphs on same scale
    - false origins marked by scale break
    - distinct points should be of roughly equal precision
    - distinct points should have independent errors
    - large numbers of confidence intervals can be misleading
    - plots with substantial noise should not have prominent smooth curves on them
    - legend self-explanatory
- *Fundamentals of Data Visualization*, C. Wilke, 2019 O'Reilly

"large numbers of confidence intervals can be misleading"
" confidence intervals are appropriate for assessing uncertainty in a single estimate but are less so for comparative purposes"

5.4 Preliminary graphical analysis 85

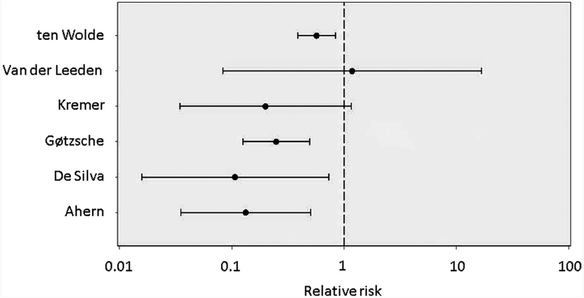


Figure 5.3 A forest plot comparing the results of six randomized controlled trials (Ahern *et al.*, 1984; De Silva and Hazleman, 1981; Gøtzsche *et al.*, 1996; Kremer *et al.*, 1987; ten Wolde *et al.*, 1996; Van der Leeden *et al.*, 1986). Four trials showed that patients who remained on their initial dosage of disease

A nugget (p.85)

"One implication of these points is that there is at most a limited and specialized role for smoothing methods other than simple binning. The reason is that smoothness which is artefactually generated is virtually impossible to distinguish from real smoothness implicit in the data. Put differently, undersmoothed data are easily smoothed by eye, or more formally, but oversmoothed data are vulnerable to misinterpretation and can be unsmoothed only, if at all, by delicate analysis."

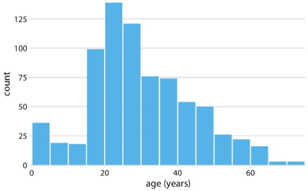


Figure 7.1: Histogram of the ages of Titanic passengers.

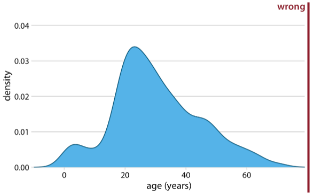


Figure 7.5: Kernel density estimates can extend the tails of the distribution into areas where no data exist and no data are even possible. Here, the density estimate has been allowed to extend into the negative age range. This is clearly nonsensical and should be avoided.

Wilke, Ch.7

… Preliminary Analysis: tables

- "simple descriptive tables of count and/or means
- "sometimes considered uninteresting, … can play important roles
- "… can demonstrate the extent to which … conclusions were well grounded in the observed data"

Illustration: Tabular analysis demonstrates some results were off support
Messer *et al.* (2010) explored the effects of neighbourhood characteristics (economic deprivation and racial segregation) on the risk of preterm birth. Their tabulation of women and preterm births by every combination of level of economic deprivation and racial segregation and by county, race and maternal education level demonstrated how unevenly the data were distributed, some combinations of economic deprivation and racial segregation being entirely absent. On the basis of these tables, the authors concluded that their logistic regression results were 'off support' (Manski, 1993), in that they involved extrapolation to predict the risk of preterm births for groups of women for whom no data were available. Although not necessarily to be avoided, the interpretation of off-support results should be particularly cautious.

The primary reporting of the conclusions of statistical analysis is by de-

… Preliminary Analysis: tables

Messer et al 2010 (on website)

Confounding associated with social stratification or other selection processes has been called structural confounding. In the presence of structural confounding, certain covariate strata will contain only subjects who could never be exposed, a violation of the positivity or experimental treatment effect assumption. Thus, structural confounding can prohibit the exchangeability necessary for meaningful causal contrasts across levels of exposure. The authors explored the presence and magnitude of structural confounding by estimating the independent effects of neighborhood deprivation and neighborhood racial composition (segregation) on rates of preterm birth in Wake and Durham counties, North Carolina (1999–2001). Tabular analyses and random-intercept fixed-slope multilevel logistic models portrayed different structural realities in these counties. The multilevel modeling results suggested some nonsignificant effect of residence in tracts with high levels of socioeconomic deprivation or racial residential segregation on adjusted odds of preterm birth for white and black women living in these counties, and the confidence limit ratios indicated fairly consistent levels of precision around the estimates. The results of the tabular analysis, however, suggested that many of these regression modeling findings were off-support and based on no actual data. The implications for statistical and public health inference, in the presence of no data, are considered.

confounding factors (epidemiology); multilevel analysis; premature birth; residence characteristics; social class; social environment

Effects of Socioeconomic and Racial Residential Segregation on Preterm Birth: A Cautionary Tale of Structural Confounding

Lynne C. Messer*, J. Michael Oakes, and Susan Mason

* Correspondence to Dr. Lynne C. Messer, Center for Health Policy, Duke Global Health Institute, Duke University, CB 90392, 2812

Another nugget (p. 87)”

“Units should always be stated and chosen so that, as far as is feasible, estimates are neither very large numerically nor very small. …

Standard errors should be given to two or at most three working digits and primary estimates should have a rounding error of less than one-tenth of a standard error. For example, a mass might be quoted as 0.453 kg with a standard error of 0.026 kg”

Explanation

- Note: Chapter uploaded to Quercus page, under "Modules"
- §5.1: interpretation of $\hat{\beta}_j$: a unit change increase in covariate x_j (will produce / is associated with) a change of $\hat{\beta}_j$ in the response y, all other variables held fixed — my wording
- we often cannot hold other variables fixed; correlation is not causation
- §5.2: "the causal effect of an action is the difference between the outcomes where the action was or was not taken"
- potential outcomes $y_i(1)$ and $y_i(0)$; the response under T (treatment) or C (control)
- causal effect defined as $y_i(1) - y_i(0)$
- note that variables like gender cannot be changed; not usually considered a cause

  exceptions: employment equity CD Ch 7

Explanation

- §5.3 designed experiments
- “Although we would like to know the individual causal effects $(y_i(1) - y_i(0))$. this is not possible because only (T or C) can be assigned to a given subject at a given moment in time. However, we can aspire to estimate the average value over the group.”
- “randomization ensures that the groups will be balanced on the average”
- analysis can be based on permutation test
- if experimental units differ in identifiable ways, we might incorporate this
- for example, dividing into male and female, and randomizing within gender
- this is called blocking

- Please read Chapter 5 of FLM-2

Sections 5.4 – 5.7 to be discussed next week

In the News

- Five national surveys – reflecting national quotas for age and gender – were conducted this year to evaluate susceptibility to coronavirus-related misinformation
- The study found the most consistent predictor of decreased susceptibility to misinformation about Covid-19 was numerical literacy
- study author Dr Sander van der Linden: “ I was surprised to see numeracy playing such a strong role here – it was one of the single most important predictors.”

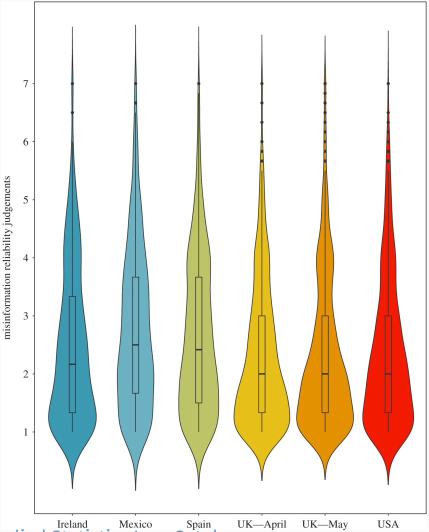
Poor numerical literacy linked to greater susceptibility to Covid-19 fake news

Cambridge University study also suggests older people less likely to believe coronavirus misinformation

- Coronavirus - latest updates
- See all our coronavirus coverage

▲ A man in a mask passes 5G conspiracy graffiti in London in April 2020. Photograph: Neil Hall/EPA

- Title: Susceptibility to misinformation about Covid-19 around the world
- Data: large national surveys in Ireland (n = 700), the USA (n = 700), Spain (n = 700) and Mexico (n = 700), conducted between mid-April and early May of 2020, and two separate surveys in the UK (n = 1050 and n = 1150).
- samples were balanced on national quotas for age and gender and obtained from Respondi, an ISO-certified panel provider of digital online data for public opinion research. Sampling continued until the quotas were filled
- Measures: general predictors – age, gender, education level, political ideology, trust in gov/science/journalists, numeracy score
- Response: participants were asked to rate the reliability of each of these statements on a 1–7 Likert scale
- Analysis: we estimated an ordinary least squares (OLS) linear regression to predict susceptibility to misinformation
- we conducted two logistic regressions ...



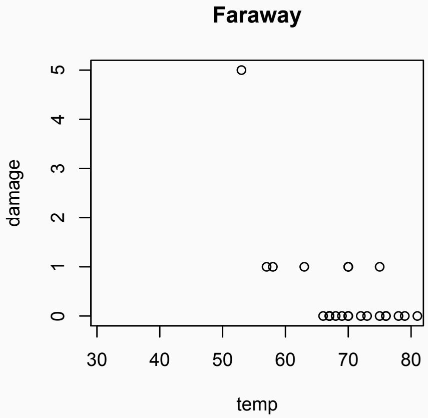
- Violin plot – modern replacement of a boxplot Wilke, Ch. 9
- Density estimate turned sideways and mirrored
- shows bimodality but boxplot cannot
- "violins begin and end at the minimum and maximum values" not here
- see `ggplot2::violin`
- ```
  violinplotall <- ggplot(data=covidall
   geom_violin(trim=FALSE, show.legend=
   scale_x_discrete(labels=c("Ireland",
   ylab("Misinformation reliability jud
  ```
- https://osf.io/jnu74/
  ```

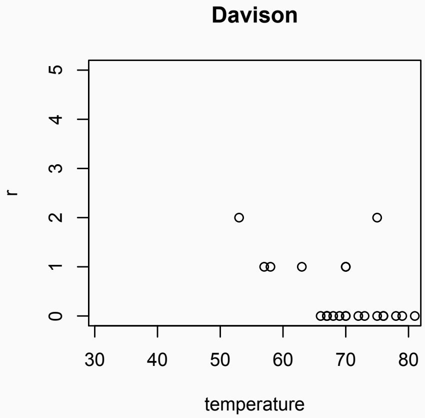
Pause

Binomial Data

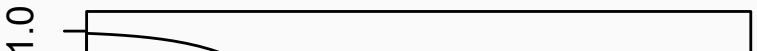
Table 1.3 O-ring thermal distress data. r is the number of field-joint O-rings showing thermal distress out of 6, for a launch at the given temperature (°F) and pressure (pounds per square inch) (Dalal *et al.*, 1989).

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





Faraway



Davison

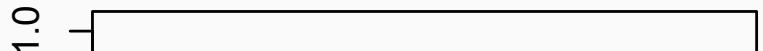


Table 1. O-Ring Thermal-Distress Data

		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.

Link

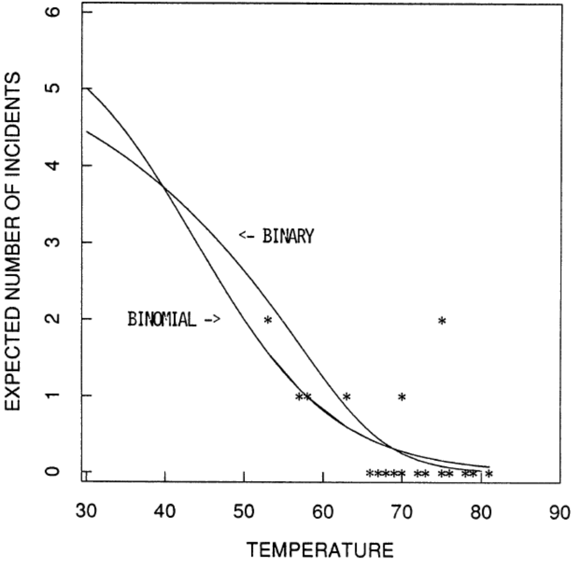


Figure 4. O-Ring Thermal-Distress Data: Field-Joint Primary O-Rings, Binomial-Logit Model, and Binary-Logit Model.

Modelling numbers/proportions of events

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

FELM 'covariate classes', p.26

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)

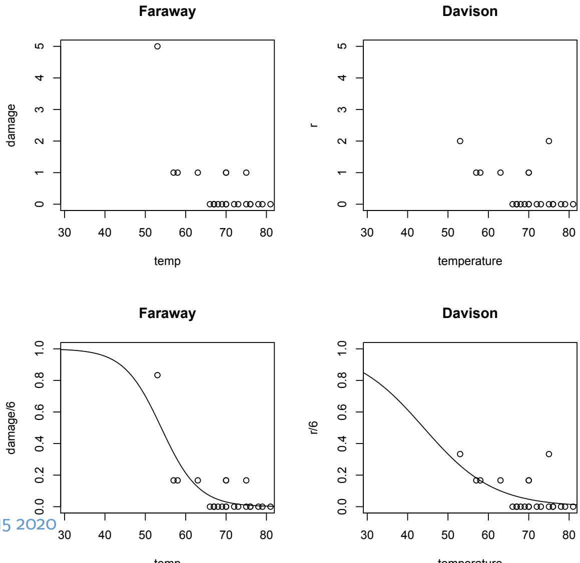
    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
```

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



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:

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

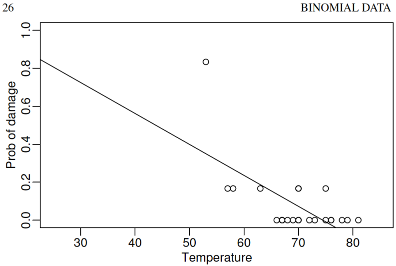
- Linear on the logit scale:

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

- linear predictor:

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

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



Inference

```
> 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:

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

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

... inference

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

- maximum likelihood estimate $\hat\beta_0, \hat\beta_1$ $\quad \partial\ell(\beta; y)/\partial\beta = 0$

- $$\hat\beta_0 = 5.08498, \quad \hat\beta_1 = -0.11560 \qquad j(\beta) \equiv -\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}$$

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

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

- 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
- $\nu = dim(A) - dim(B)$

... inference

```
 > 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
> 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
> 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
```

... inference

- 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 p.30

… 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 $\ell_{\text{p}}(\beta_1)$, details to follow
- `confint(logitmodcorrect):`
  (-0.212~~2262~~, -0.0244~~701~~)

ELM p. 31