STA 22101S: Applied Statistics II
Wednesday, 2-5 pm, BA B024 Spring, 2015

Course description: This course focuses on generalized linear models and related methods for applications, with an emphasis on planning of studies, analysis of categorical data, generalizations of linear regression models, and interpretation. The topics covered will include: planning of studies, categorical data, generalized linear models, random effects and mixed linear models, and semiparametric and nonparametric regression.

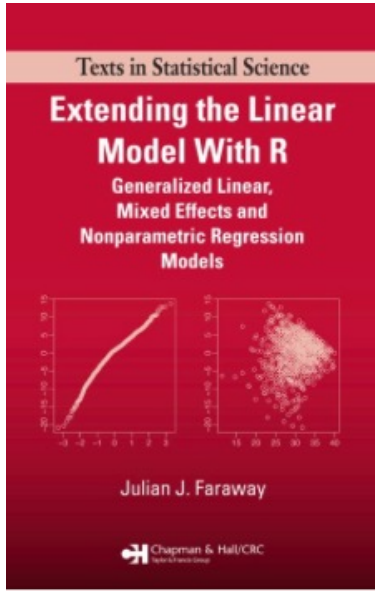
Grading: The grade in the course will be based on three homework sets (60%) and a final project (40%). Late homework will not be accepted, but the homework with the worst grade will count for just 10%, with the remaining three homework sets counting equally. The final project will consist of analysis of a data set, which you will find. Each student must find a unique set of data. You will submit a report on the analysis of the data, along with executable R code that reproduces the analysis. Homework and project will be submitted electronically. Tentative due dates for homework sets are 11.59 pm on: Feb 4, Mar 4, Apr 1.
Project due April 15.

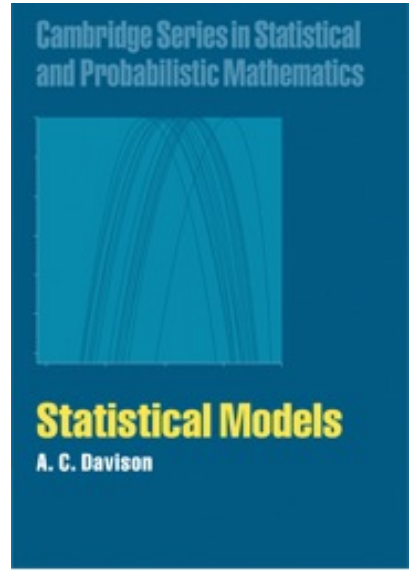
Text: The course text is *Extending the Linear Model with* `R` by J.J. Faraway (Chapman & Hall). Highly recommended is *Principles of Applied Statistics* by D.R. Cox and C.A. Donnelly (CUP).
I will often refer to Chapters 8 - 10 of *Statistical Models* by A.C. Davison (Cambridge University Press).
I will also refer to *Advanced Data Analysis from an Elementary Point of View*, by C. Shalizi. The current version is available (and often updated) at Shalizi's web page.
Additional useful resources include the 4th addition of *Modern Applied Statistics with* `S` by W.N. Venables and B.D. Ripley (Springer), *Applied Statistics* by D.R. Cox and E.J. Snell (Chapman & Hall) and *Elements of Statistical Learning*, by T. Hastie, R. Tibshirani, and J. Friedman (Springer).

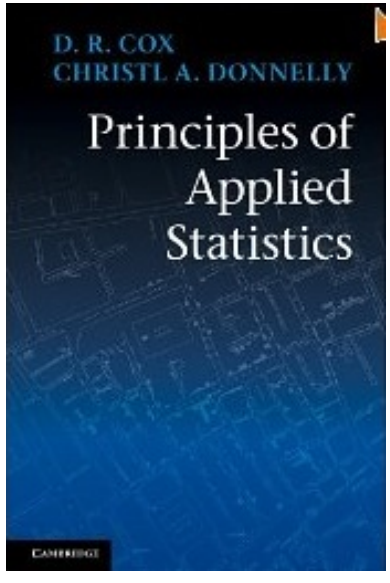
Course web page(s): I am using Blackboard to manage the course list and grading, but the course information is all on the web page
`http://www.utstat.utoronto.ca/reid/2201S15.html`. The Blackboard page for STA2201S will lead you to this page via the first announcement.

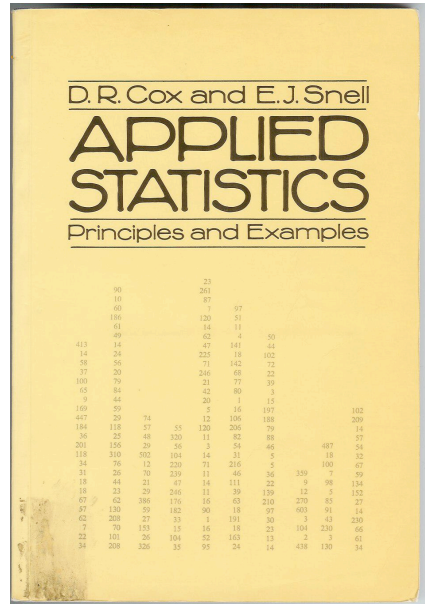
Computing: The course is built around the `R` computing package. There are some `R` resources listed on the course webpage.

Contact: Nancy Reid: SS 6002A, `reid@utstat.utoronto.ca`.
Office Hours: Tuesday 1 to 3, or by appointment. Please check with me by email during workshop weeks at the Fields Institute – Jan 13, 20 27, Feb 10, 24, Mar 24.
TA: Bo Chen, `broad.chen@mail.utoronto.ca`.


Texts in Statistical Science
Extending the Linear Model With R
Generalized Linear, Mixed Effects and Nonparametric Regression Models
Julian J. Faraway
Chapman & Hall/CRC


Cambridge Series in Statistical and Probabilistic Mathematics
Statistical Models
A. C. Davison


D. R. COX
CHRISTL A. DONNELLY
Principles of Applied Statistics
CAMBRIDGE


D. R. Cox and E. J. Snell
APPLIED STATISTICS
Principles and Examples

Link

36–402, Undergraduate Advanced Data Analysis (2013)

www.stat.cmu.edu/~cshalizi/uADA/13/

Reader

Topics in I... Data – Blog Weebly – W...n Made Easy Apple Yahoo! Google Maps Wikipedia News Popular

Topics in I... | Weebly – W... | (40) Twitter | 11 Ways T... | 36–402, U... | Blackboard... | Create and... | yau confer... | Take a scr...

Cosma Shalizi

36-402, Undergraduate Advanced Data Analysis

Spring 2013

Tuesdays and Thursdays, 10:30--11:50 Porter Hall 100

The goal of this class is to train you in using statistical models to analyze data — as data summaries, as predictive instruments, and as tools for scientific inference. We will build on the theory and applications of the linear model, introduced in 36-401, extending it to more general functional forms, and more general kinds of data, emphasizing the computation-intensive methods introduced since the 1980s. After taking the class, when you're faced with a new data-analysis problem, you should be able to (1) select appropriate methods, (2) use statistical software to implement them, (3) critically evaluate the resulting statistical models, and (4) communicate the results of their analyses to collaborators and to non-statisticians.

During the class, you will do data analyses with existing software, and write your own simple programs to implement and extend key techniques. You will also have to write reports about your analyses.

Graduate students from other departments wishing to take this course should register for it under the number "36-608". Enrollment for 36-608 is very limited, and by permission of the professor only.

Prerequisites

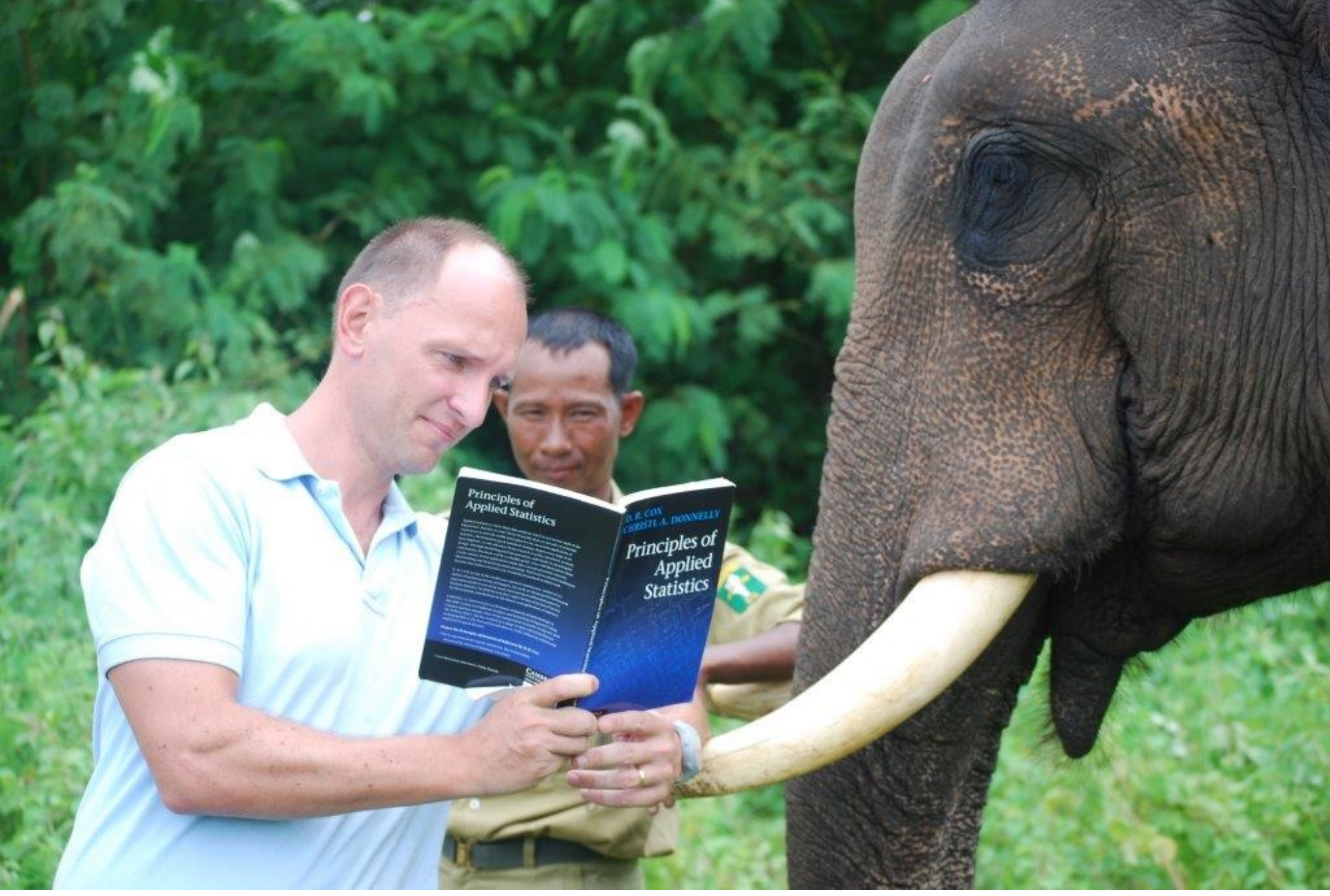
36-401, or consent of the instructor. The latter is only granted under *very* unusual circumstances.

Instructors

Professor	Cosma Shalizi	cshalizi [at] cmu.edu 229 C Baker Hall 268-7826
Teaching assistants	Mr. Beau Dabbs	
	Ms. Francesca Matano	
	Mr. Mingyu Tang	
	Ms. Xiaolin Yang	

Topics, Notes, Readings

Model evaluation: statistical inference, prediction, and scientific inference; in-sample and out-of-sample errors, generalization and over-fitting, cross-validation; evaluating by


Principles of
Applied Statistics
D. R. COX
CHRISTL A. DONNELLY
Principles of
Applied Statistics

Topics

This course focuses on generalized linear models and related methods for applications, with an emphasis on planning of studies, analysis of categorical data, generalizations of linear regression models, and interpretation. The topics covered will include:

- planning of studies,
- categorical data,
- generalized linear models,
- random effects and mixed linear models,
- semiparametric and nonparametric regression

STA2201H Methods of Applied Statistics II

The course will focus on generalized linear models (GLM) and related methods, such as generalized additive model involving nonparametric regression, generalized estimating equations (GEE) and generalized linear mixed models (GLMM) for longitudinal data. This course is designed for Master and PhD students in Statistics, and is REQUIRED for the Applied paper of the PhD Comprehensive Exams in Statistics. We deal with a class of statistical models that generalizes classical linear models to include many other models that have been found useful in statistical analysis, especially in biomedical applications. The course is a mixture of theory and applications and includes computer projects featuring R (S+) or/and SAS programming.

Topics: Brief review of likelihood theory, fundamental theory of generalized linear models, iterated weighted least squares, binary data and logistic regression, epidemiological study designs, counts data and log-linear models, models with constant coefficient of variation, quasi-likelihood, generalized additive models involving nonparametric smoothing, generalized estimating equations (GEE) and generalized linear mixed models (GLMM) for longitudinal data.

... topics

This course focuses on generalized linear models and related methods for applications, with an emphasis on planning of studies, analysis of categorical data, generalizations of linear regression models, and interpretation. The topics covered will include:

- planning of studies,
- categorical data,
- generalized linear models,
- random effects and mixed linear models,
- semiparametric and nonparametric regression

- principles of applied statistics

- where do models come from?
- how are they to be interpreted?
- where is the data? how was it collected?

... topics

This course focuses on generalized linear models and related methods for applications, with an emphasis on planning of studies, analysis of categorical data, generalizations of linear regression models, and interpretation. The topics covered will include:

- planning of studies,
- categorical data,
- generalized linear models,
- random effects and mixed linear models,
- semiparametric and nonparametric regression

- how does this advance science

- case studies, examples, items from the news, topics in the workshops

Topics in more detail

- Categorical Data: binomial, Poisson, multivariate ELM Ch. 2, 3, 4, 5
- Generalized Linear Models: theory, fitting, analysis of deviance, quasi-likelihood ELM Ch. 6, 7
- Mixed and Random Effects ELM Ch. 8, 9, 10
- Nonparametric Regression, Smoothing ELM Ch. 11, 12
- Survival Data; Proportional Hazards Regression SM 5.4, 10.8

- Principles of Applied Statistics Cox & Donnelly

  We are writing ... for students of statistics concerned with the relationship between the detailed methods and theory they are studying and the effective application of these ideas

Preliminaries CD Ch. 1

- we need statistics when we have “unexplained and haphazard variation”
- distinguish between natural variability and measurement error
- one is of interest, the other needs to be accommodated
- example: blood pressure varies over time scales of minutes, hours, days, even in healthy individuals.

  Measurements of blood pressure are also imprecise, but this variability is not of especial interest, although we need to be aware of it

Components of investigation CD Ch. 1

- "the ideal sequence"
  - formulation of research questions
  - search for relevant data
  - design and implementation of investigations to obtain data
  - analysis of data
  - interpretation of the results
- "The essence of our discussion will be on the achievement of individually secure investigations. These are studies which lead to unambiguous conclusions ... Yet virtually all subject-matter issues are tackled sequentially ... typically the important and challenging issue of synthesizing information of very different kinds, so crucial for understanding, has to be carried out informally"
- Examples: Northern Hemisphere temperature time series; investigations of bovine tuberculosis; evidence for HIV as the cause of AIDS

... components of investigation CD Ch. 1

- very focused research question – ideal
- research questions emerge as the study develops – “consequent reformulation of the detailed statistical model used for analysis... usually causes no conceptual problem ... Major changes of focus... ideally need confirmation in supplementary investigations”
- “An extreme case of departure from the ideal sequence ... a large body of administrative data become available, and there is a perception that it must contain interesting information about something... the term ‘data mining’ is often used in such context... how much effort should be spent on such issues beyond the simple tabulation of frequencies and pairwise dependencies must depend in part on the quality of the data ... any conclusions are in most cases likely to be tentative and in need of independent confirmation”
- “ A large amount of data is in no way synonymous with a large amount of information”

In the news

Buzzfeed, Dec 22

Is Drinking Wine Better Than Going To The Gym? According To Scientists, Yes!

Like Share Comment Email More

By Natalie Roterman | Sep 15 2014, 04:51PM EDT

A glass of red wine per day is as beneficial as going to the gym. Shutterstock

BMJ, Dec 10

Cancer cure

The headline: “Long-used drug shows new promise for cancer”

Globe & Mail, Jan 17, 2007

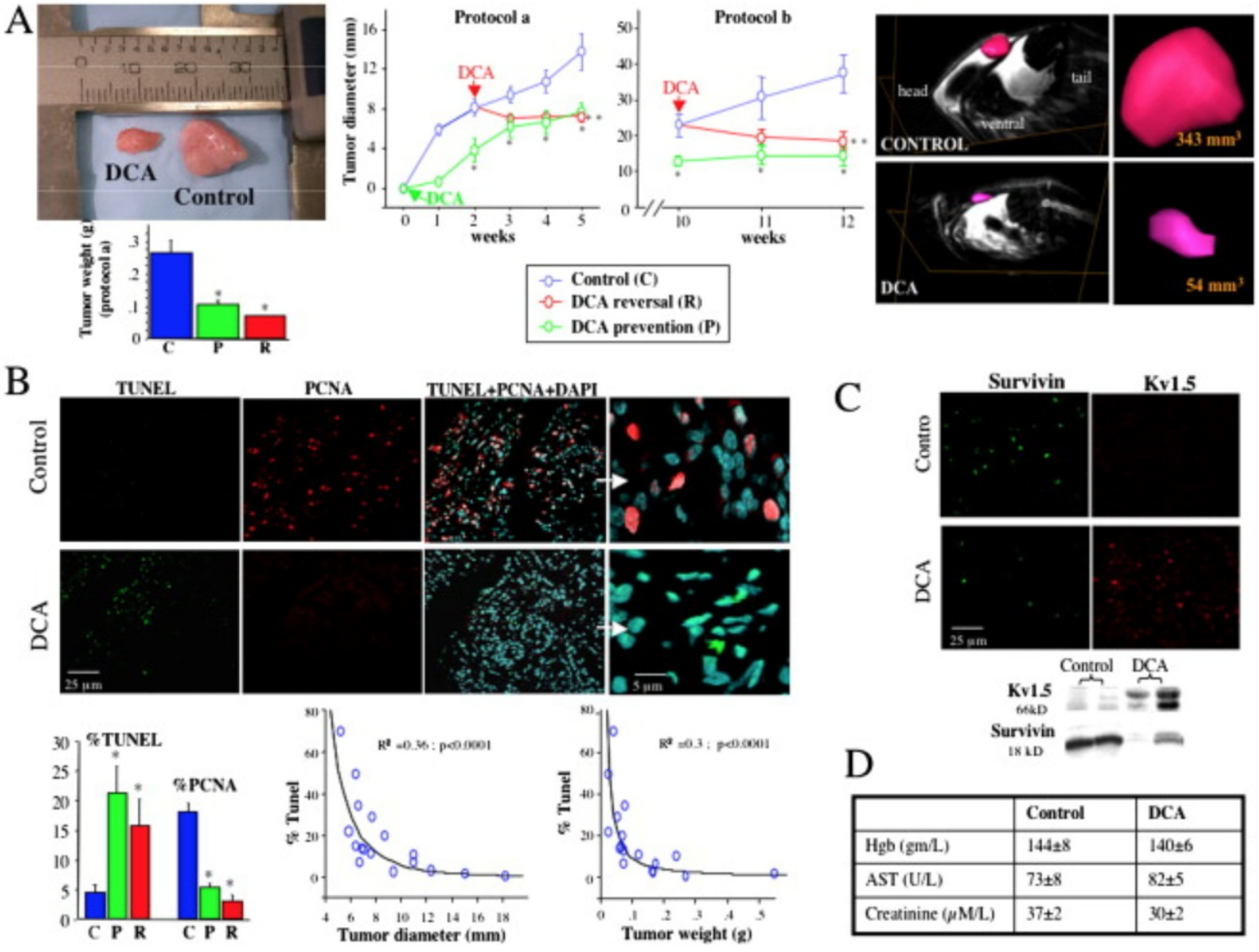
The news article:

Imagine, if you will, a drug that shrinks cancer cells and can even make tumours disappear. A couple of spoonfuls a day of powder in a glass of water is all you need. There are no nasty side effects like nausea and hair loss, and no damage to internal organs such as with traditional chemotherapy. And it costs only about $2 a dose. Too good to be true? Not according to a Canadian researcher who stumbled upon the potentially new anti-cancer agent called dichloroacetate, or DCA, a drug long used to treat rare metabolic disorders. ‘This is one of the most exciting results I’ve ever had,’ said Evangelos Michelakis, an associate professor of medicine at the University of Alberta in Edmonton. ‘But I can’t be overenthusiastic until it works in a human being.’

Cancer cure

The journal article: "A Mitochondria-K+ Channel Axis Is Suppressed in Cancer and Its Normalization Promotes Apoptosis and Inhibits Cancer Growth," Bonnet et al., 2007 Cancer Cell 11, 37–51.

In the first set of experiments (protocol a), 21 animals were divided into three groups: untreated controls (n = 5), DCA-prevention rats (n = 8), which received DCA just after cell injection for 5 weeks, and DCA-reversal rats (n = 8), which received DCA 2 weeks post-cell injection for 3 more weeks. ... In a second set of experiments (protocol b), we studied whether the effects of DCA were sustained for longer periods of time and whether DCA would have a similar effect in more advanced tumors. We followed three groups of rats (n = 6/group) for 12 weeks



	Control	DCA
Hgb (gm/L)	144±8	140±6
AST (U/L)	73±8	82±5
Creatinine (μM/L)	37±2	30±2

Some stories never die

Globe & Mail, August 2006

... chocolate

Easter, 2010

An ounce of (dark chocolate) prevention

Or less: Researchers find those who eat 7.5 grams a day have a lower risk of heart disease

BY KATE KELLAND

Easter eggs may be good for you, but only if you eat small ones made from cocoa-rich dark chocolate, according to the latest in a string of scientific studies to show potential health benefits of chocolate.

German researchers studied more than 19,300 people over a decade and found those who ate the most chocolate – an average of 7.5 grams a day – had lower blood pressure and a 39 per cent lower risk of having a heart attack or stroke than those who ate the least amount of chocolate – an average of 1.7 grams a day.

But the difference between the two groups was just less than six grams of chocolate a day, less than one small square of an average 100-gram bar, they wrote in a study in the European Heart Journal to be published today.

Brian Buijsse of the German Institute of Human Nutrition in Nuthetal, who led the study, said people should not use his work as an excuse to stuff themselves with chocolate.

"Small amounts of chocolate may help to prevent heart disease, but only if it replaces other energy-dense food, such as snacks, in order to keep body weight stable," he said.

Although they said more work needed to be done to be sure, the researchers think the flavanols in cocoa may be the reason why chocolate seems to be good for blood pressure and heart health – and since there is more cocoa in dark chocolate, dark chocolate may have a greater effect.

Flavanols are a class of the antioxidant flavonoids that are sponsible for improving the bioavailability of nitric oxide from the cells that line the inner wall of blood vessels," Dr. Buijsse said.

Nitric oxide is a gas that, once released, causes the smooth muscle cells of the blood vessels to relax and widen, he said, adding that this may contribute to lower blood pressure.

researchers used data from participants of a larger study called European Prospective

sure, height and weight were measured and details of their diet, lifestyle and health were recorded.

Put in terms of absolute risk, he said, the findings showed that if people in the group eating the least amount of chocolate increased their chocolate intake by six grams a day, 85 fewer heart attacks and strokes

be expected to occur over a period of about 10 years.

Hospital Zurich said basic science had now demonstrated "quite convincingly" that dark chocolate with a cocoa content of at least 70 per cent reduces some kinds of stress and can improve blood flow and blood pressure.

But he said: "Before you rush to add dark chocolate to your diet, be aware that 100 grams

ries.

"You may want to subtract an

Researchers think the flavanols in cocoa may be the reason why chocolate seems to be good for heart health. ELISE AMENDOLA/ASSOCIATED PRESS

> ***Before you rush to add dark chocolate to your diet, be aware that 100 grams ... contains roughly 500 calories.***
>
> Frank Ruschitzka,
> University Hospital Zurich

Link

U.S.

To Improve a Memory, Consider Chocolate

By PAM BELLUCK OCT. 26, 2014

Link

CBCnews | Health

Home | World | Canada | Politics | Business | Health | Arts & Entertainment | Technology &

Health | Rate My Hospital

Chocolate health myth dissolves

Health-enhancing flavanols that end up on the shelf will likely appear in form other than chocol

By Kelly Crowe, CBC News Posted: Jan 05, 2015 10:00 AM ET | Last Updated: Jan 05, 2015 10:01 PM ET

Binomial Data

ELM, Ch. 2, SM, Ch.1; §4.4,5;

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).

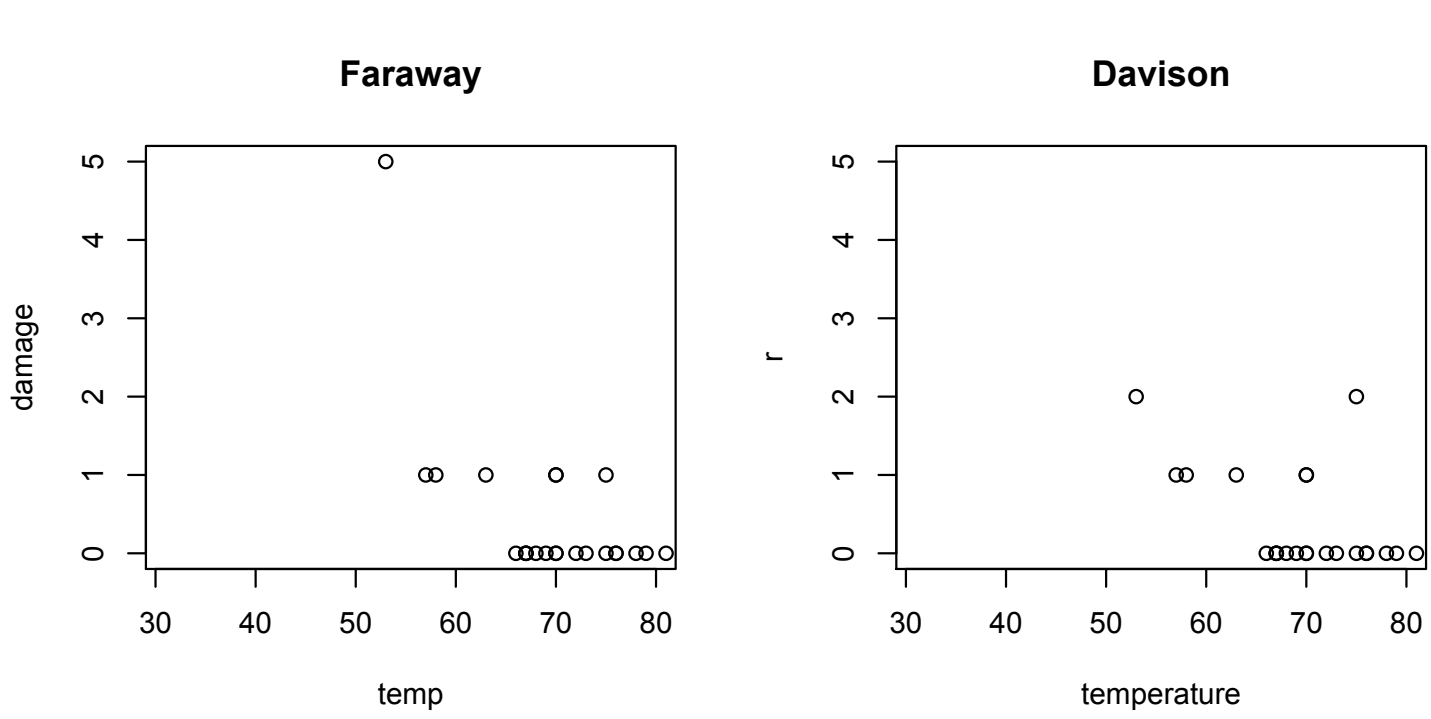
Flight	Date	Number of O-rings with thermal distress, r	Temperature (°F) x_1	Pressure (psi) x_2
1	21/4/81	0	66	50
2	12/11/81	1	70	50
3	22/3/82	0	69	50
5	11/11/82	0	68	50
6	4/4/83	0	67	50
7	18/6/83	0	72	50
8	30/8/83	0	73	100
9	28/11/83	0	70	100
41-B	3/2/84	1	57	200
41-C	6/4/84	1	63	200
41-D	30/8/84	1	70	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
61-I	28/1/86	—	31	200

Challenger data: Faraway

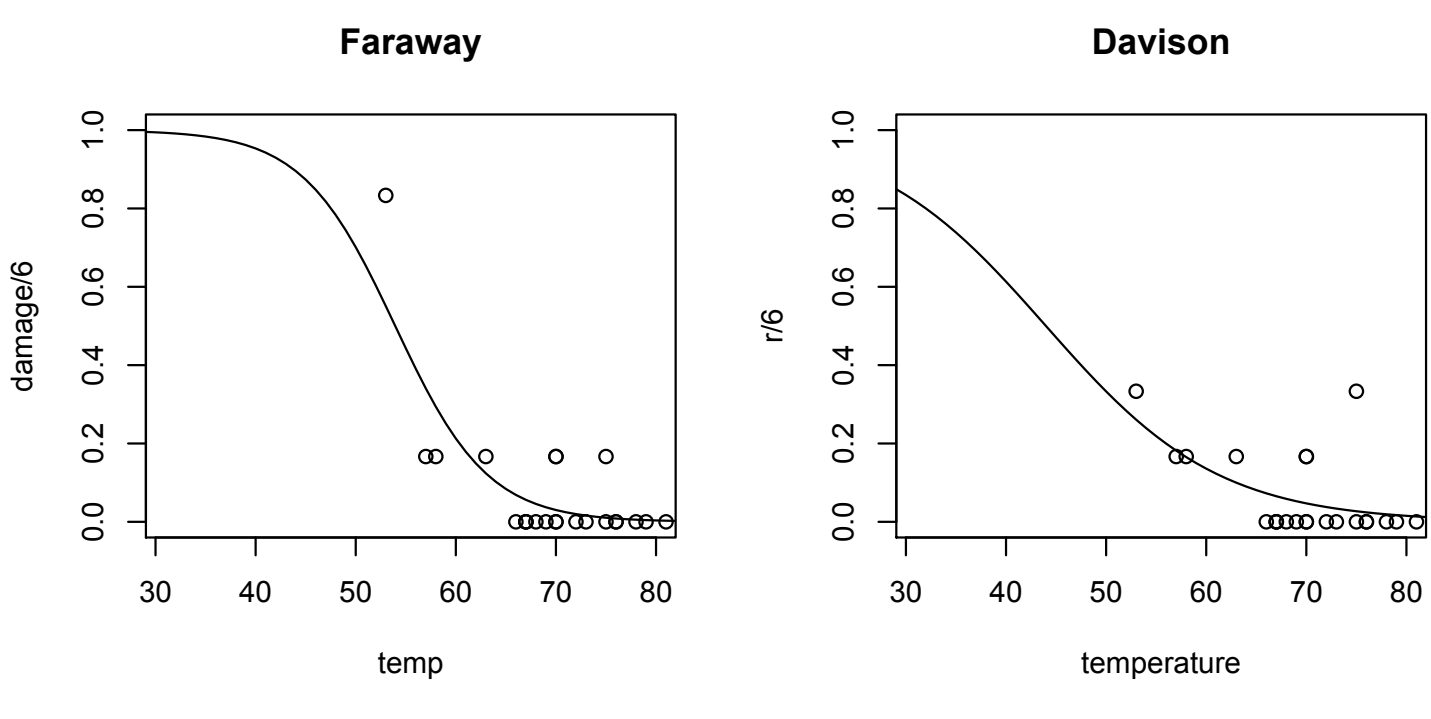
```
> library(faraway); data(orings); head(orings)
  temp damage
1   53      5
2   57      1
3   58      1
4   63      1
5   66      0
6   67      0
> logitmod <- glm(cbind(damage,6-damage) ~ temp, family = binomial, data = orings)
> summary(logitmod)
Call:
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial,
    data = orings)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9529  -0.7345  -0.4393  -0.2079   1.9565

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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(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
AIC: 33.675

Number of Fisher Scoring iterations: 6
```

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 avo
                                    #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)))
```


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

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


Faraway
damage/6
temp
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
80

Davison
r/6
temperature
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
80

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

Dalal et al (1989) *Journal of the American Statistical Association*

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 covariate classes, p.26

Estimation for binomial

- observations $y_1, \ldots, y_n$ independent $Bin(n_i, p_i)$
- likelihood function

$$L(\underline{p}; \underline{y}) = \prod_{i=1}^{n} f(y_i; p_i) = \prod_{i=1}^{n} \binom{n_i}{p_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

- log-likelihood function

$$\ell(\underline{p}; \underline{y}) = \sum_{i=1}^{n} \{y_i \log(p_i) + (n_i - y_i) \log(1 - p_i) + \log \binom{n_i}{p_i}\}$$

- maximum likelihood estimator

$$\hat{p}_i = y_i / n_i$$

$$\partial \ell / \partial p_i = 0, i = 1, \ldots, n$$

Regression modelling with binomial

- $\hat{p}_i = y_i/n_i$ $\qquad \partial\ell/\partial p_i = 0, i = 1, \ldots, n$
- saturated model – one parameter for each observation y_i

- regression: link the p_i's through x_i: $p_i = \textit{function}(\beta; x_i)$
- 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

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

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

all vectors are column vectors

... regression modelling with binomial

- $p_i = \dfrac{\exp(x_i^{\mathrm{T}}\beta)}{1+\exp(x_i^{\mathrm{T}}\beta)}$
- $0 < p_i < 1$
- cumulative distribution function for the logistic distribution
- the inverse is called the logit link function:

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

- using other cdf's gives different link functions
- normal cdf: $p_i = \Phi(x_i^{\mathrm{T}}\beta)$, probit link: $\Phi^{-1}(p_i) = x_i^{\mathrm{T}}\beta$
- Gumbel cdf: $p_i = \exp\{-\exp(-x_i^{\mathrm{T}}\beta)\}$,
  c-log-log link:$\log\{1-\log(1-p_i)\} = x_i^{\mathrm{T}}\beta$
- $x_i^{\mathrm{T}}\beta = \eta_i$ called the linear predictor

Inference

```
> summary(logitmodcorrect)

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

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.95227  -0.78299  -0.54117  -0.04379   2.65152

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: $\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i$
- log-likelihood function

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

... inference

```
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 *
```

- $\ell(\beta; y) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - n_i \log\{1 + \exp(\beta_0 + \beta_1 x_i)\}]$
- maximum likelihood estimate $\hat{\beta}_0$, $\hat{\beta}_1$ $\partial\ell(\beta; y)/\partial\beta = 0$
-

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

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

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

... inference

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

... 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 = shu
> 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 $1 - \Phi\{\frac{\hat{\beta}_2}{\widehat{\text{s.e.}}(\hat{\beta}_2)}\} = 1 - \Phi(1.105) = 0.27$

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 $\ell_p(\beta_1)$, details to follow
- `confint(logitmodcorrect):`
  `( -0.2122262, -0.0244701 )`

ELM p. 31

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 p.29

```
> summary(logitmodcorrect)

...
    Null deviance: 24.230  on 22  degrees of freedom
Residual deviance: 18.086  on 21  degrees of freedom
AIC: 35.647
Number of Fisher Scoring iterations: 5
```

Actually, null model ($\beta_1 = 0$) also fits: `pchisq(24.23,22, lower.tail = F) = 0.33`, but improvement is statistically significant

Logistic regression

- read §2.4 for one motivation of logistic regression model
- read §2.5 (and AS I) for interpretation of parameters in terms of log odds
- see Example `mdl` in §2.5 for logistic regression with qualitative covariates
- what is the algebraic form of the model? how are the dummy covariates coded?
- in other words, what is x_i^{T}?

Example:

SM Example 10.18

aggregated data presented in textbook

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
1	0	0	0	0	1	1

... example 10.18

- `library(SMPracticals); data(nodal); head(nodal)` all covariates 0/1
- several patients have the same value of the covariates

  covariate classes: ELM
- these can be added up to make a binomial observation
- ```
  > nodal2[1:4,]
    m r age stage grade xray acid
  1 6 5   0     1     1    1    1
  2 6 1   0     0     0    0    1
  3 4 0   1     1     1    0    0
  4 4 2   1     1     0    0    1
  ```
- ```
  > ex1018binom = glm(cbind(r,m-r) ~ ., data = nodal2, family = binomial)
  > summary(ex1018binom)  # stuff omitted
  Coefficients:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -3.0794     0.9868  -3.121  0.00180 **
  age          -0.2917     0.7540  -0.387  0.69881
  stage         1.3729     0.7838   1.752  0.07986 .
  grade         0.8720     0.8156   1.069  0.28500
  xray          1.8008     0.8104   2.222  0.02628 *
  acid          1.6839     0.7915   2.128  0.03337 *
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

  (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: 41.693
  ```

... example 10.18

```
> 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:      40.71
Residual Deviance: 19.64  AIC: 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

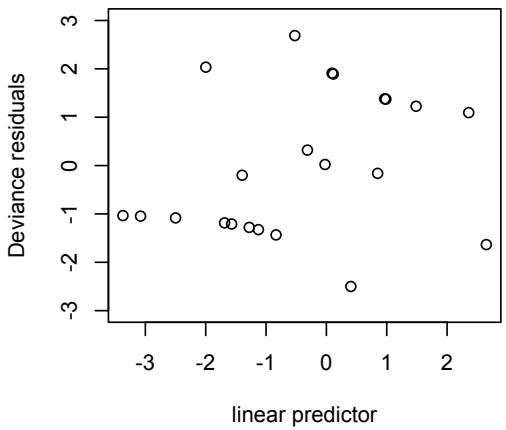
– several mistakes in text on pp. 491,2;

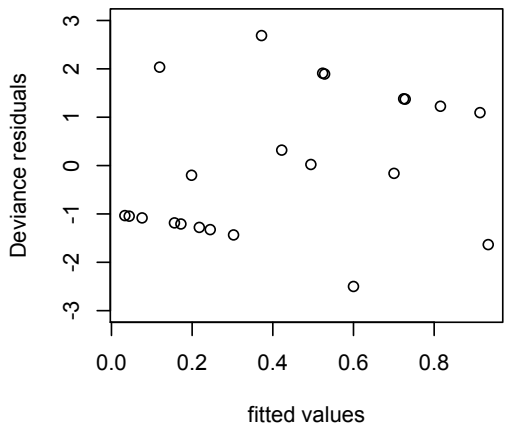
– deviances in Table 10.9 are incorrect as well

`http://statwww.epfl.ch/davison/SM/` has corrected version

... example 10.18

```
> 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
```





... inference

- Residual Deviance is log-likelihood ratio statistic for the fitted model compared to the saturated model
- saturated model maximized at $\tilde{p}_i = y_i/n_i$

$$\ell(\tilde{p}) = \sum_{i=1}^{n} \{y_i \log(y_i/n_i) + (n_i - y_i) \log(1 - y_i/n_i)\}$$

- fitted model maximized at $\hat{\beta}$

$$\ell(\hat{\beta}) = \sum_{i=1}^{n} \{y_i \log p_i(\hat{\beta}) + (n_i - y_i) \log(1 - p_i(\hat{\beta}))\}$$

- twice the difference:

$$2 \sum_{i=1}^{n} [y_i \log\{y_i/n_i p_i(\hat{\beta})\} + (n_i - y_i) \log\{(n_i - y_i)/(n_i - n_i p_i(\hat{\beta}))\}]$$

- see p.29, after (2.1), where $n_i p_i(\hat{\beta}) = \hat{y}_i$

Deviance residuals

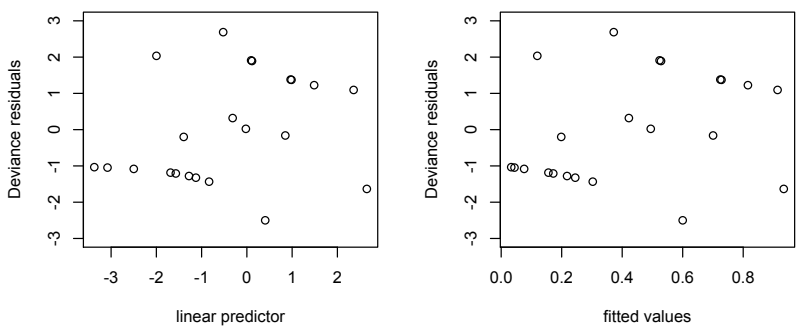
```
> summary(ex1018binom)

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
```

Deviance:

$$2\sum_{i=1}^{n}[y_i \log\{y_i/n_i p_i(\hat{\beta})\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i p_i(\hat{\beta}))\}]$$

approximately χ^2_{n-q}

$$r_{Di} = \pm\sqrt{(2[y_i \log\{y_i/n_i\hat{p}_i\} + (n_i - y_i)\log\{(n_i - y_i)/(n_i - n_i\hat{p}_i)\}])}$$



Design and Analysis CD Ch. 1

- choice of material/individuals to study – "units of analysis"
- "For studies of a new phenomenon it will usually be best to examine situations in which the phenomenon is likely to appear in the most striking form, even if this is in some sense artificial"
- statistical analysis needs to take account of the design (even if statistician enters the project at the analysis stage)
- need to be clear at the design stage about broad features of the statistical analysis – more publicly convincing **and** "reduces the possibility that the data cannot be satisfactorily analysed"
- example: Female faculty salary survey
- "it is unrealistic and indeed potentially dangerous to follow an initial plan unswervingly ... it may be a crucial part of the analysis to clarify the research objectives"

Experimental and observational studies CD Ch. 1

- experiment is a study in which all key elements are under the control of the investigator
- in an observational study key elements cannot be manipulated by the investigator.
- “It often, however, aids the interpretation of an observation study to consider the question: what would have been done in a comparable experiment?”
- Example: hormone replacement therapy and heart disease
- observational study – strong and statistically significant reduction in heart disease among women taking hormone replacement therapy
- women’s health study (JAMA, 2002, p.321) – statistically significant increase in risk among women randomized to hormone replacement therapy

Principles of measurement CD Ch. 1

- “construct validity – measurements do actually record the features of concern”
- “record a number of different features sufficient to capture concisely the important aspects”
- reliable – i.e. reasonably reproducible
- “cost of the measurements is commensurate with their importance”
- “measurement process does not appreciably distort the system under study”

Types and phases of analysis CD Ch. 1

- "A general principle, sounding superficial but difficult to implement, is that analyses should be as simple as possible, but no simpler."
- the method of analysis should be transparent
- main phases of analysis
  - data auditing and screening;
  - preliminary analysis;
  - formal analysis;
  - presentation of conclusions

“What are the principles of applied statistics?” CD Ch. 1

- “formulation and clarification of focused research questions of subject-matter importance
- design of individual investigations and sequences of investigations that produce secure answers and open up new possibilities
- production of effective and reliable measurement procedures
- development of simple, and where appropriate, not-so-simple methods of analysis, with suitable software, that address the primary research questions, often through a skilful choice of statistical model, and give some assessment of uncertainty
- effective presentation of conclusions
- structuring of analyses to facilitate their interpretation in subject matter terms and their relationship to the knowledge base of the field.”