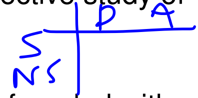
From last class

grimreaper.R

- contingency tables – Example 10.19; prospective study of survival, with 2 covariates



- age; smoking status
- interest in effect of smoking on survival; confounded with age
- note that results are invariant to order
- how to interpret coefficients?

- estimated odds of survival among smokers, adjusted for age, $\exp(-0.4274) = 0.65 = 65\%$, relative to non-smokers
- 95% confidence interval $\exp\{-0.4274 - 2(0.1770)\}, \exp\{-0.4274 + 2(0.1770)\} = (0.46, 0.93)$

- HW 2 Q4 could be done this way... or, could follow analysis of Ex 10.24

... Example 10.19

	sm	non-sm	sm	non-sm	sm	non-sm	
d	2	1	3	5	14	7	
a	53	61	121	152	95	114	···
	55	62	124	157	109	121	
Age	18-24		25-34		35-44		···

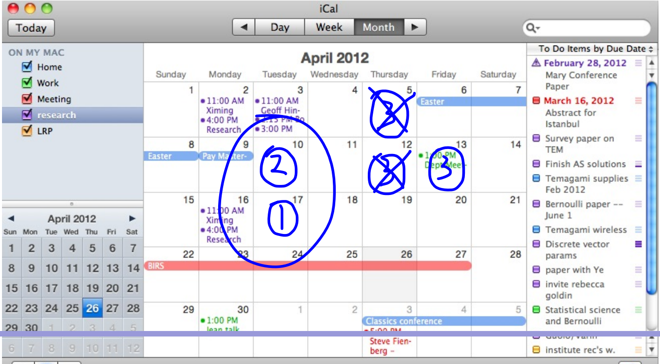
```
> summary(glm(cbind(alive,dead) ~ smoker + factor(age), data = smoking, family = binomial))

Call:
glm(formula = cbind(alive, dead) ~ smoker + factor(age), family = binomial,
    data = smoking)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.68162  -0.19146  -0.00005   0.22836   0.72545

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.8601     0.5939   6.500 8.05e-11 ***
smoker            -0.4274     0.1770  -2.414 0.015762 *
factor(age)25-34  -0.1201     0.6865  -0.175 0.861178
factor(age)35-44  -1.3411     0.6286  -2.134 0.032874 *
factor(age)45-54  -2.1134     0.6121  -3.453 0.000555 ***
factor(age)55-64  -3.1808     0.6006  -5.296 1.18e-07 ***
factor(age)65-74  -5.0880     0.6195  -8.213  < 2e-16 ***
factor(age)75+   -27.8073 11293.1437  -0.002 0.998035
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

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

e^{β}: odds of surv. among sm relative to non sm adj for age

The next weeks

March 2	§10.5 Count data and log-linear models
March 9	§10.6 Overdispersion and quasi-likelihood, GEEs
March 16	§10.7 Semiparametric models
March 23	Generalized additive models and lasso OR
	§10.8 Survival data
March 30	Finishing pieces, + review

10



STA 2201S Assignment 2. *due Friday, March 2 at the beginning of class*

When answering questions requiring numerical work, the results are to be reported in a narrative summary, in your own words. Tables and Figures may be included, but must be formatted along with the text. **Do not include in this summary printouts of computer code.** Analysis of variance/deviance tables, tables of coefficients and their estimated standard errors, and other output should be formatted separately and reported only to the relevant number of significant digits. All computer code used to obtain the results summarized in the response should be provided as an appendix.

1. Exercise 10.2.5, Davison (p. 479)
2. Exercise 10.3.8, Davison (p. 487)
3. Exercise 10.4.1, Davison (p. 497)
4. The data in Table 1 below is taken from *Applied Statistics* by Cox & Snell (p.176). This shows the numbers of subjects reporting "breathlessness" and "wheeze", categorized by age group. The subjects are a sample of 18,282 coalminers known to be smokers, but with no Xray indication of lung disease.

Table 1: Set 11 from Cox & Snell (1981). Numbers of coalminers responding to breathlessness and wheeze according to age group.

Breathlessness		Yes		No		Total
Wheeze		Yes	No	Yes	No	
	20–24	9	7	95	1841	1952
	25–29	23	9	105	1654	1791
	30–34	54	19	177	1863	2113
	35–39	121	48	257	2357	2783
Age	40–44	169	54	273	1778	2274
Group	45–49	269	88	324	1712	2393
	50–54	404	117	245	1324	2090
	55–59	406	152	225	967	1750
	60–64	372	106	132	526	1136
Total		1872	600	1833	14022	18282

(a) Consider first the incidence of wheeze among the group "breathlessness = yes".

Cox & Donnelly: Model Choice (Ch. 7)

- Mostly, we aim to summarize the aspects of interest by parameters, preferably small in number and formally defined as properties of the probability model
- parameters of interest, directly addressing the questions of concern; often concerning systematic variation
- nuisance parameters necessary to complete the statistical model; often concerning haphazard variation
- the choice of parameters involves their interpretability

built in → step

stepAIC ⟷ library(MASS)

... parameters of interest §7.1.2

- it is essential that subject-matter interpretation is clear and measured in appropriate units, which should always be stated
- it is preferable that the units chosen give numerical answers that are neither inconveniently large or small
- example: assessment of risk factors often/usually expressed as a ratio or percentage effect
- but for public health we'd like to know how many individuals could be affected – this is a difference of probabilities, not a ratio
- `http://understandinguncertainty.org/spinning`
- while we're at it: `http://www.statisticsblog.com/`, `http://projecteuclid.org.myaccess.library.utoronto.ca/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aoas/1267453942`, `http://biostatisticsryangosling.tumblr.com/`

... choice of a specific model §7.3

- often this will involve at least two levels of choice, first between distinct separate families and then between specific models within a chosen family
- of course all choices are to some extent provisional
- example: survival data – gamma or weibull model both extend the exponential
- example: linear regression $E(Y) = \beta_0 + \beta_1 x$, or $E(Y) = \gamma_0/(1 + \gamma_1 x)$
- neither, one, or both may be adeuqate
- comparisons between models are sometimes made using Bayes factors, ... however, misleading if neither model is adequate
- for dependencies of y on x that are curved, a low-degree polynomial might be adequate
- but subject-matter may suggest an asymptote, in which case $E(Y) = \alpha + \gamma e^{-\delta x}$ may be preferred

... model choice with a natural hierarchy

- polynomials provide a flexible family of smooth relationships, although poor for extrapolation
- examples: $E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$, $E(Y) = \beta_{00} + \beta_{10} x_1 + \beta_{01} x_2 + \beta_{20} x_1^2 + \beta_{11} x_1 x_2 + \beta_{02} x_2^2$
- it will typically be wise the measure the x_i from a meaningful origin near the centre of the data
- example: time series $AR(p)$
- for a single set of data choose the smallest order compatible with the data, using standard tests
- for several sets of data, usually would choose the same order for each set
- it would not normally be sensible to include β_{11} and not β_{20}, β_{02}
- with qualitative (categorical) x's, this means models with interaction terms should include the corresponding main effects
- there are rare exceptions, see p.133

... lots of x's, which to use?

- response y, potential explanatory variables $x_1, \dots, x_p$
- suppose interest focusses on the role of a particular variable or set of variables, x^*
  - the value, standard error, and interpretation of the coefficient of x^* depends on which other variables are included
  - variables prior to x^* in the generating process should be included in the model unless...
  - unless conditional independent of y given x^* and other vars in model OR conditionally independent of x^* given other vars in model
  - variables intermediate between x^* and y omitted in initial assessment
  - relatively mechanical methods of choosing may be helpful in preliminary exploration, but are insecure as a basis for final interpretation
  - explanatory variables not of direct interest but known to have a substantial effect should be included
  - several different models may be equally effective
  - if there are several potential explanatory variables on an equal footing, interpretation is particularly difficult

... model choice

-