STA 442/2101 F Nov. 17/09

0. Teaching evaluations
1. Interpreting coefs from logistic regression → go back to probabilities of Cox & Snell
   also HW 2 Q 3
   + deviance test for fit of model
2. Simpson's paradox & reverse – use race example from 199
3. Redelmeier & Singh + ~~see~~ HO – Sylvestre & Hanley
4. Survival data text: 5.4 & 10.7, 8

2.

	death penalty yes	no	
w	18	141	11.88%
~~victim~~ defendant b	17	149	10.24%

	white victim dp yes	no			black victim dp yes	no	
w	19	132	12%	w	0	9	0%
b	11	52	17%	b	6	97	5.8%

12.6
17.5

1. Binomial regression $y_i \sim Bin(m_i, p_i(\beta))$ $i = 1, \dots, n$ indep

(residual) deviance $D = \sum_{i=1}^{n} \left\{ y_i \log \frac{y_i}{m_i p_i(\hat\beta)} + (m_i - y_i) \log\left(\frac{m_i - y_i}{m_i(1-p_i(\hat\beta))}\right) \right\}$

is a goodness-of-fit test for the model
$\hat p_i = p_i(\hat\beta)$

(relative to the model p_i unconstrained)

Example HW 2 $\quad \text{logit } p_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$

residual deviance $= 0$ $\qquad p_{ij}(\hat{\beta}) = y_{ij}/m_{ij}$

$i = 1, 2 \; ; \; j = 1, \dots, 6$

$\text{logit } p_{ij} = \mu + \alpha_i + \beta_j$ no $X^=$

resid deviance $= 20.204$ on 5 df

If $(\alpha\beta)_{ij} \equiv 0$, this $\dot\sim \chi^2_5$

$\text{pr}\left(\chi^2_5 > 20.204\right) = 0.001$ "reject H_0"

Example 10.15 where resid dev. $= 67.28 \dot\sim \chi^2_{31}$ poor fit
attributed to 2 outliers
(instead of $X^=$)
when those points omitted
$D = 28.02$ on 24 df ✓

N.B. This does not work for binary data Exercise 10.4.1 (a)

$$D = -2 \sum_{i=1}^{n} \hat{p}_i \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) + \log(1-\hat{p}_i)$$

doesn't ~~depend on~~ compare y_i to $\hat{p}_i$
(bec each $y_i = 0$ or 1)

3. Redelmeier & Singh ~~[illegible]~~ 'social determinants of health'
- motivation, selection of sample, nominees, controls, winners
education
- how do controls work? - same sex, in same age
- response : length of life
- covariates - born in USA, name change, ethnicity, genre,
birth yr , sex , age at 1st film, total films,
winner / loser or winner / nominee
- time zero
- survivor test selection bias lead time bias
- unmeasured confounding - "3 strategies"

Stat. Analysis primary - KM & log-rank test
regression Cox PH note: use quad. & cub. !

Results - baseline characteristics ✓
- career variables ..
- cause of death
note: none of these are 'primary' analyses

\# survival figure

- difference in ~~mean life~~ life expectancy ~~area under curve~~ 3.9 yrs $p = 0.003$
and sub-analyses
- 28% reduction in death rate ?? how computed
'Cox' model
- then compared to nominees

Discussion - see marked copy

4. Models & methods for survival data

- r.v. Y measures time ($Y \geq 0$)

density $f(y)$ on $\mathbb{R}^+$

cdf $F(y) = \Pr(Y \leq y)$

survivor f^{n} $1 - F(y) = \Pr(Y > y) = S(y)$

hazard f^{n} $h(y) = \dfrac{f(y)}{S(y)}$

cum. haz. $H(y) = \int_0^y h(t)\,dt = -\log S(y)$

i.e. $S(y) = \exp\{-H(y)\}$ $\quad f(y) = h(y)\exp\{-H(y)\}$

3 examples in §5.4.1.

Weibull ~~f~~ ~~th~~ $S(y) = \exp\{-(\frac{y}{\theta})^{\alpha}\}$, $\theta, \alpha > 0$

$$f(y) = \underbrace{\frac{\alpha}{\theta}\left(\frac{y}{\theta}\right)^{\alpha-1}}_{h(y)} \cdot \exp\left\{-\left(\frac{y}{\theta}\right)^{\alpha}\right\}$$

- censoring : often true failure time not observed

(right) random censoring

$$Y_j = \min(Y_j^0, C_j) \qquad \begin{matrix} Y_j^0 \sim F \\ C_j \sim G \end{matrix} \quad \text{indep't}$$

data (y_j, d_j) $j = 1, \ldots, n$ $\quad d_j = \begin{cases} 1 & \text{obs} \\ 0 & \text{censored} \end{cases}$

see Figure 5.8

- likelihood f^{n} $\prod_{d_j=1} f(y_j)\{1 - G(y_j)\} \cdot \prod_{d_j=0} S(y_j)\, g(y_j)$

often f, S dep. on θ, but not G $\Rightarrow$ $\ell(\theta) = \sum_{d_j=1} \log f(y_j; \theta) +$

Kaplan - Meier estimator of $S(\cdot)$:

$$\hat{S}(t) = \prod_{i: Y_{(i)} \le t} \left(\frac{n-i}{n-i+1} \right)^{\delta_{(i)}} , \quad t \le y_{(n)}$$

$$0 \text{ or undefined} \quad t > y_{(n)}$$

$y_{(1)} \le \dots \le y_{(n)}$ ordered failure times
$\delta_{(1)}, \dots, \delta_{(n)}$ associated censoring indicators

nb. in Davison p 197 $\hat{S}(\cancel{y}) = \prod_{j: y_j < \cancel{y}} \left(1 - \frac{1}{r_j}\right)^{d_j}$

$d_j = \delta_{(j)}$; r_j = # still alive at time y_j^- = "risk set for time y_j"

can be derived as max lik est., but more obviously is an extension of eddf $\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \le t\}$

Example Table 5.3

§10.8.1. Regression models for censored survival data

$(\underset{\sim}{x}_j, y_j, d_j)$ $\qquad f(y_j; x_j, \beta)$ density

d_j: 1, 0

x_j: cov.

y_j: failure or cens. time

$$\ell(\beta) = \sum_{j=1}^{n} \{ d_j \log h(y_j; x_j, \beta) - H(y_j; x_j, \beta) \}$$

PH model $h(y_j; x_j, \beta) = \underbrace{\xi(x_j^T \beta)}_{\text{increase due to covariates}} \underbrace{h_0(y_j)}_{\text{baseline hazard}}$

partial likelihood

$$L_p(\beta) = \prod_{\substack{j=1}}^{n} \left\{ \frac{\xi(x_j^T\beta)}{\sum_{i\in R_j} \xi(x_i^T\beta)} \right\}^{\delta_j} = \prod_j \left(\frac{\cancel{\exp} \, x_j^T\beta}{\sum_{i\in R_j} e^{x_i^T\beta}} \right)^{\delta_j}$$

usually $\xi(x_j^T\beta) = \exp(x_j^T\beta)$

$$l_p = \sum_{\substack{j=1 \\ \delta_j=1}}^{n} \left[x_j^T\beta - \log\left\{ \sum_{i\in R_j} \exp(x_i^T\beta) \right\} \right]$$

$$= \sum x_j^T\beta - A_j(\beta) \quad \text{say}$$

$$l_p' = \sum \left(x_j^T - \frac{B_j}{A_j} \right) \qquad l_p'' = \quad \cdots \qquad (10.62), (10.63)$$

Special case: $x_j = \begin{cases} 1 & \text{group } A \\ 0 & B \end{cases}$

$p=1$

$$l_p'(\beta) = \sum_{\delta_j=1} \left\{ x_j - \frac{\sum_{i\in R_j} x_j e^{x_j\beta}}{\sum_{i\in R_j} x_j e^{x_j\beta}} \right\}$$

$$\cancel{l_p'(\beta)}\Big|_{\beta=0} = \sum_{\delta_j=1} \left(x_j - \frac{\sum_{i\in R_j} x_j}{\sum_{i\in R_j} 1} \right) = \sum_{\delta_j=1} \left(x_j - \frac{m_{1j}}{m_{0j}+m_{1j}} \right)$$