

The biostatistics workhorse models are logistic regression and proportional hazards (“Cox model”).

Logistic regression. Papers by Cornfield et al on risk of heart disease. 1960s. Berkson (*JASA* 1944) on bioassays. Yule (*JRSS* 1925) on population growth, who credits Verhulst (*Royal Academy Brussels* 1845).

David Cox (1972). Regression models and lifetables. *Journal of the Royal Statistical Society Series B* 34: 187–220

MEDLINE: 50,000 hits on logistic regression, 25,000 hits on the Cox model (c. 12 million papers).

Progress? Causation can be inferred by modeling observational data—if you believe the modeling assumptions. Generally, the *if* is ignored. So, inferences are conditional on unexamined (and implausible?) assumptions. Other strategies are cruder, maybe more successful.

Epidemiology:

- Cholera. John Snow (1855). *On the Mode of Communication of Cholera*. Churchill, London.
- Puerperal fever. See Irvine Loudon (2000). *Tragedy of Childbed Fever*, Oxford.
- Pellagra. See Ken Carpenter (1981). *Pellagra*, Academic Press;  
also Milton Terris (1964). *Goldberger on Pellagra*, LSU Press.
- Cervical cancer.
- Smoking, Asbestos . . . .

Assumptions discussed: <http://www.stat.berkeley.edu/~census/651.pdf>

The book-length version (with exercises): <http://www.stat.berkeley.edu/~census/s151.pdf>

More technical; covers non-linear case: <http://www.stat.berkeley.edu/~census/601.pdf>

Examples discussed: DA Freedman (1999). *Statistical Science* 14: 243–58. Reprinted in John Panaretos (2003). *Stochastic Musings*. Lawrence Erlbaum.  
<http://www.stat.berkeley.edu/~census/521.pdf>

## An application of the Cox model

Pargament et al (2001). *Archives of Internal Medicine* 161: 1881–85.

Significant impact of negative religious feelings on risk of death.

“Physicians are now being asked to take a spiritual history . . . . Our findings suggest that patients who indicate religious struggle during a spiritual history may be at particularly high risk . . . . Referral of these patients to clergy to help them work through these issues may ultimately improve clinical outcomes; further research is needed . . . .

596 mainly Baptist & Methodist patients age 55+, hospitalized for serious illness at Duke Medical Center or Durham Veterans’ Affairs Medical Center

Two-year followup: 176 deaths, 152 lost to followup

Positive & negative religious feelings

Adjustment by Cox model for age, race, sex, severity of illness . . . and for missing data

Negative religious feelings increase risk of death by 6% (\*\*)

\*  $P < .10$    \*\*  $P < .05$    \*\*\*  $P < .01$

## The Cox model in brief

A log linear model for death rate. (More generally, hazard rate.) Effects are multiplicative, not additive, not synergistic. Constant across time on test, people. There is a background rate, and. . .

Old people die 1.39 times faster.† RR (Relative risk, Rate ratio) = 1.39

Men die 1.41 times faster.

*Therefore*, old men die  $1.39 \times 1.41$  times faster.

People with negative religious feelings die 1.06 times faster.

Income effects? Sicker? Missing data?

Fixed by the model—if you believe the model.

Relative Risks, i.e, Rate Ratios (RR), i.e, Hazard Ratios (HR)

Religious feelings –	1.06	**
Religious feelings +	0.98	
Age	1.39	*
Black	1.21	
Male	1.41	*
Hospital	1.14	
Education	0.98	
Physical Health		
Diagnoses	1.04	
ADL	0.98	
Patient	1.41	***
Anesthesiologist	1.54	***
Mental health		
MMSE	0.96	
Depression	0.95	
Quality of life	1.03	

Religious feelings—positive & negative: seven-item questionnaire, 0–3 points

“Decided the devil made this happen (RR = 1.19,  $P = .02$ )”

Mini-Mental State Examination

Depression—questionnaire

ADL—Activities of Daily Life—questionnaire

Physical Health

  Patient self rating, poor to excellent

  Anesthesiologist rating of patient, 0–5 points

“Quality of life” is observer-rated

Why the Cox model? “This robust semiparametric procedure was chosen for its flexibility in handling censored observations, time-dependent predictors, and late entry into the study.”

† Age at baseline; the background hazard rate is a function of time on test. “Old” is a dummy.

## HRT (Hormone Replacement Therapy) prevents heart disease (?)

50 observational studies say yes—by factor of 2.

Two experiments say, no effect, even harmful.

Diana B Petitti (1998). Hormone replacement therapy and heart disease prevention: Experimentation trumps observation. *Journal of the American Medical Association* 280: 650–51.

Diana B Petitti (2002). Hormone replacement therapy for prevention. *Journal of the American Medical Association* 288: 99–101.

Nurses' health study. Observational. Francine Grodstein, Meier Stampfer et al (1996). Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. *New England Journal of Medicine* 335: 453–61.

6,224 post-menopausal women on combined HRT vs 27,034 never-users. 0–16 years of followup (average is 12). Analysis by the Cox model. Treatment variable is HRT. 17 confounders, including age, age at menopause, height, weight, smoking, blood pressure, cholesterol, . . . , exercise. 11 get into main model.

“Proportional-hazards models were used to calculate relative risks and 95 percent confidence intervals, adjusted for confounding variables. We observed a marked decrease in the risk of major coronary heart disease among women who took estrogen with progestin, as compared with the risk among women who did not use hormones (multivariate adjusted relative risk 0.39; 95 percent confidence interval, 0.19 to 0.78). . . .”

“Women who take hormones are a self-selected group and usually have healthier lifestyles with fewer risk factors. [However,] participants in the Nurses' Health Study are relatively homogeneous. . . . Unknown confounders may have influenced our results, but to explain the apparent benefit on the basis of confounding variables, one must postulate unknown risk factors that are extremely strong predictors of disease and closely associated with hormone use.”

Women's Health Initiative. Experiment. Writing Group for the Women's Health Initiative Investigators (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the women's health initiative randomized controlled trial. *Journal of the American Medical Association* 288: 321–333.

16,608 post-menopausal women randomized to HRT or control. Trial stopped early.

Rate ratio for CHD (Coronary Heart Disease) is 1.29.

“Nominal” 95% confidence interval is 1.02 to 1.63.

“Adjusted” 95% confidence interval is 0.85 to 1.97.

Cox model. Covariates: clinical center, age, prior disease, assignment to diet.

“The adjusted 95% CIs presented herein use group sequential methods to correct for multiple analyses over time. A Bonferroni correction for 7 outcomes as specified in the monitoring plan . . . was applied to all clinical outcomes other than CHD and breast cancer . . .”

DAF

Deaths per 1000 women randomized, over 5 years of followup:

$$231/8506 = 27.2 \text{ vs } 218/8102 = 26.9$$

Their primary endpoint

CHD (fatal + non-fatal MI) per 1000 women randomized, over 5 years of followup:

164/8506 = 19.3 vs 122/8102 = 15.1, Rate ratio = 19.3/15.1 = 1.28

## Political science: JSTOR snapshots

About 1000 hits on logistic regression/logits.

Russell L Hanson (1983). The “content” of welfare policy: The states and Aid to Families with Dependent Children. *The Journal of Politics* 45: 771–785.

James M McCormick and Michael Black (1983). Ideology and senate voting on the Panama canal treaties. *Legislative Studies Quarterly* 8: 45–63.

Another 4000 hits on regression (including multiple regression, partial correlation; excluding logistic).

William F Ogburn and Inez Goltra (1919). How women vote. *Political Science Quarterly* 34: 413–433. (Partial correlations.)

Harold F Gosnell and Norman N Gill (1935). An analysis of the 1932 presidential vote in Chicago. *American Political Science Review* 29 967–984. (Partial correlations, ordinary regression.)

5,000 hits on “table 1” (excluding correlation and regression) and 50,000 on “political.”

## Anomie Take II

Emile Durkheim (1897). *Le suicide*. F. Alcan, Paris.

English translation by JA Spalding (1951). Free Press.

Bob Putnam (2000). *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster. “Telecommunications constitutes the third countertrend toward greater social connectedness considered in this chapter, and by all odds it is the most important.” [p. 166.]

DAF. Talk more on the phone, live longer. Cross national comparison, 84 countries. LHS variable is life expectancy at birth. Regression diagnostics suggest excluding least developed countries, due to causal heterogeneity. (Countries in sample have > .2 lines per person and income > US \$2500 per person per year.) Log-log.

	Coeff	SE	<i>t</i>
Constant	4.062	.060	67.1
Telephone lines per person	.058	.009	6.5
Income per person	.033	.006	5.8

$$R^2 = 0.50$$

The aim . . . is to provide a clear and rigorous basis for determining when a causal ordering can be said to hold between two variables or groups of variables in a model . . . . The concepts . . . all refer to a model—a system of equations—and not to the “real” world the model purports to describe.

—HA Simon (1957). *Models of Man*. Wiley. p. 32

## Response schedules and invariance; potential outcomes

There are two treatments (levels  $u$  and  $v$ ), and a response variable  $Y$ . Both treatments may be applied to subject  $i$ . There are three parameters,  $a$ ,  $b$ , and  $c$ . With no treatment at all, response level for subject  $i$  is  $a$ , up to random error. Each additional unit of treatment #1 adds  $b$  to the response. Likewise, each additional unit of treatment #2 adds  $c$  to the response. Constancy of parameters across subjects and levels of treatment is an assumption. If treatment #1 is applied at level  $u$  and #2 at level  $v$ , response is

$$Y_{i,u,v} = a + bu + cv + \epsilon_i.$$

Invariance of  $a$ ,  $b$ ,  $c$ ,  $\epsilon_i$ ? My response is unaffected by your treatment?? Manipulation???

## Statistical assumptions

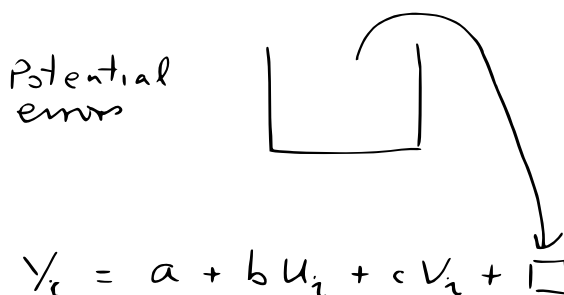
In order to make the transition from a hypothetical experiment to the actual observational study, and to justify OLS, we assume:

- (i)  $E(\epsilon_i) = 0$ .
- (ii)  $\epsilon_i$  are independent and identically distributed across subjects  $i$ .
- (iii) Exogeneity. Nature chooses  $U_i$ ,  $V_i$  independently of the random errors  $\epsilon_i$ , and determines the response  $Y_i$  from the response schedule:

$$Y_i = Y_{i,U_i,V_i} = a + bU_i + cV_i + \epsilon_i.$$

Nature shows us  $U_i$ ,  $V_i$ ,  $Y_i$ . We're good to go. a) OLS works. b) Causal inferences justified—built into the response schedule. (With small samples, need to assume errors are normal.)

The statistical assumptions in box model format



IID = Independent and Identically Distributed = Draws Made at Random with Replacement

The box of potential errors remains the same from one draw to another. The probability distribution of one draw is the same as any other. The distribution is identical. The outcome of one draw cannot affect the distribution of any another. That is independence.

The response-schedule model is due to Neyman, has been rediscovered several times, e.g., by Hodges & Lehmann. Sometimes called “Rubin’s model,” but this ignores the history. A few key cites:

