Today

- data presentation Jennifer
- model selection – Cox & Donnelly
- in Boston
- semi-parametric regression
- HW 3: due March 21
- Final Exam: April 11 2:00 – 5:00 pm

... PSID – using `lme`

From last week: those annoying degrees of freedom

not reported if you use `lmer`

```
> mmod = lme(log(income) ~ cyear*sex + age + educ ,
random = ~ 1 +  cyear | person, data=psid)

Fixed effects: log(income) ~ cyear * sex + age + educ
                 Value Std.Error   DF   t-value p-value
(Intercept)  6.674204 0.5433252 1574 12.283995  0.0000
cyear        0.085312 0.0089996 1574  9.479521  0.0000
sexM         1.150313 0.1212925   81  9.483790  0.0000
age          0.010932 0.0135238   81  0.808342  0.4213
educ         0.104210 0.0214366   81  4.861287  0.0000
cyear:sexM  -0.026307 0.0122378 1574 -2.149607  0.0317


> mmod2 = lmer(log(income) ~ cyear*sex + age + educ +
+ (cyear | person), data=psid)

Fixed effects:
            Estimate Std. Error t value
(Intercept)  6.67420    0.54332  12.284
cyear        0.08531    0.00900   9.480
sexM         1.15031    0.12129   9.484
age          0.01093    0.01352   0.808
educ         0.10421    0.02144   4.861
cyear:sexM  -0.02631    0.01224  -2.150
```

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

... 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

... 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 adequate

... choice of a specific model

- 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
- it will typically be wise to measure the x_i from a meaningful origin near the centre of the data

- example:
  $E(Y) = \beta_{00} + \beta_{10}x_1 + \beta_{01}x_2 + \beta_{20}x_1^2 + \beta_{11}x_1x_2 + \beta_{02}x_2^2$
- 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

... model choice

- example: $E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$
- example: time series $AR(p)$
  $y_t = \mu + \rho_1(y_{t-1} - \mu) + \cdots + \rho_p(y_{t-p} - \mu) + \epsilon_t$
- 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

... choosing among explanatory variables

- response y, potential explanatory variables $x_1, \ldots, x_p$
- 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 these variables are conditionally independent of y, given x^* (and other variables in the model)
- OR unless they are conditionally independent of x^*, given other variables in the model
- variables intermediate between x^* and y are omitted in initial assessment of the effect of x^*
- but may be needed later to study the pathways of dependence

... choosing among explanatory variables

- 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

- A two-phase approach:
- First search among a large number of possibilities for a base for interpretation
- Second check the adequacy of that base

First phase: a broad strategy

- x^*, required explanatory variables; $\tilde{x}$ some potential further explanatory variables
- $\tilde{x}$ conceptually prior to x^*

- fit a reduced model with x^* only $\mathcal{M}_{\text{red}}$
- fit, if possible, a full model with x^* and $\tilde{x}$ $\mathcal{M}_{\text{full}}$
- compare the estimated standard errors of the coefficients for x^* under the two models

- if these are of the same order, then $\mathcal{M}_{\text{full}}$ is safer
- if precision improvement under $\mathcal{M}_{\text{red}}$ seems substantial, then explore eliminating some of $\tilde{x}$
- for example with backwards elimination

- with emphasis on the effect of x^*

Second phase: adequacy of the model

- add back selected components of the omitted variables $\tilde{x}$
- to check that conclusions are not changed
- or to report on the differences if they are
- if the model to date has been linear, may be important now to check some curvature terms, for continuous xs, and interaction terms for categorical xs
- these provide a 'warning system', but not usually direct interpretation

- interpretation of coefficients, especially in observational studies, needs care
- example: x includes several measurements of smoking behaviour: yes/no; years since quitting; no. of cigarettes smoked; pipe/cigar; etc.
- role of these depends on the goal of the study – confounder? primary exposure?

Semiparametric Regression §10.7

- model $y_j = g(x_j) + \epsilon_j, \quad j = 1, \ldots, n \quad x_j$ scalar
- mean function $g(\cdot)$ assumed to be "smooth"
- introduce a kernel function $w(u)$ and define a set of weights

$$w_j = \frac{1}{h} w\left(\frac{x_j - x_0}{h}\right)$$

- estimate of $g(x)$, at $x = x_0$:

$$\hat{g}(x_0) = \frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j}$$

- Nadaraya-Watson estimator (10.40) – local averaging

... kernel smoothing

- better estimates can be obtained using local regression at point x
- $$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & (x_1 - x_0) & \cdots & (x_1 - x_0)^k \\ \vdots & \vdots & & \vdots \\ 1 & (x_n - x_0) & \cdots & (x_n - x_0)^k \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$
- $$\hat{\beta} = (X^T W X)^{-1} X^T W y$$
- $$\hat{g}(x_0) = \hat{\beta}_0$$
- usually obtain estimates $\hat{g}(x_j), j = 1, \ldots, n$

... kernel smoothing

- odd-order polynomials work better than even; usually local linear fits are used
- kernel function is often a Gaussian density, or the tricube function (10.37)
- choice of bandwidth, h controls smoothness of function
- kernel estimators are biased
- larger bandwidth $=$ more smoothing – increases bias, decreases variance
- some smoothers allows variable bandwidth depending on density of observations near x_0
- `ksmooth` computes local averages; `loess` computes local linear regression (robustified)

Inference after fitting smooth functions

- $\hat{\beta} = (X^T W X)^{-1} X^T W y$
- $W = \text{diag}(w_1, \ldots, w_n)$
- $\hat{g}(x_0) = \hat{\beta}_0 = \sum_{j=1}^{n} S(x_0; x_j, h) y_j$
- $S(x_0; x_1, h), \ldots, S(x_0; x_n, h)$ first row of "hat" matrix $(X^T W X)^{-1} X^T W$
- $\text{E}\{\hat{g}(x_0)\} = \sum_{j=1}^{n} S(x_0; x_j, h) g(x_j)$
- $\text{var}\{\hat{g}(x_0)\} = \sigma^2 \sum_{j=1}^{n} S(x_0; x_j, h)^2$
- similarly $\hat{g} = (\hat{g}(x_1), \ldots, \hat{g}(x_n)) = \mathcal{S}_h y$
- $\nu_1 = \text{tr}(S_h)$, $\nu_2 = \text{tr}(S_h^T S_h)$ potential estimates of 'degrees of freedom'

Bias and MSE

- $\hat{g}(x)$ is biased: $\mathrm{E}\{\hat{g}(x)\} \doteq \frac{1}{2}h^2 g''(x)$
- $$\mathrm{var}\{\hat{g}(x)\} \doteq \frac{\sigma^2}{nhf(x)} \int w^2(u)du$$
- could choose h to minimize MSE $=$ bias^2 $+$ var, at x
- could choose h to minimize integrated MSE
- more usual to use cross-validation
- $$CV(h) = \sum_{j=1}^{n} \{y_j - \hat{g}_{-j}(x_j)\}^2$$

... bias and mse

- $$CV(h) = \sum_{j=1}^{n} \{y_j - \hat{g}_{-j}(x_j)\}^2$$
- $$CV(h) = \sum_{j=1}^{n} \left\{ \frac{y_j - \hat{g}(x_j)}{1 - S_{jj}(h)} \right\}^2$$
- $$GCV(h) = \sum_{j=1}^{n} \left\{ \frac{y_j - \hat{g}(x_j)}{1 - \mathrm{tr}(S_h)/n} \right\}^2$$
- $$\hat{g}(x_0) = \hat{\beta}_0 = \sum_{j=1}^{n} S(x_0; x_j, h) y_j$$
- $S(x_0; x_1, h), \ldots, S(x_0; x_n, h)$ is first row of $(X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W$

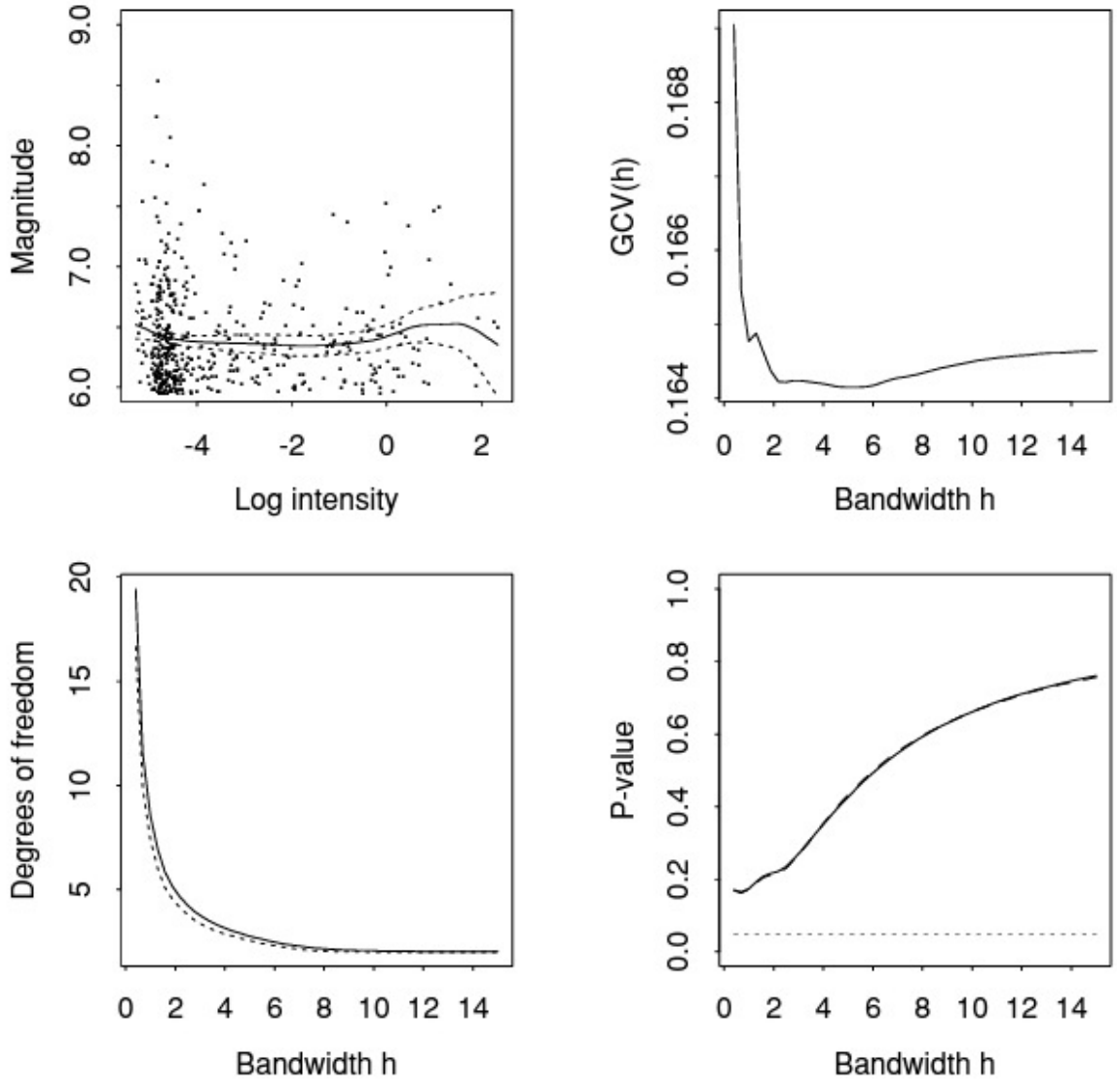

Figure 10.16 Smooth analysis of earthquake data. Upper left: local linear regression of magnitude on log intensity just before quake (solid), with 0.95 pointwise confidence bands (dots). Upper right: generalized cross-validation criterion GCV(h) as a function of bandwidth h. Lower left: relation between degrees of freedom ν_1 (solid), ν_2 (dots), and h. Lower right: significance traces for test of no relation between magnitude and log intensity, based on chi-squared approximation (dots) and saddlepoint approximation (solid). The horizontal line shows the conventional 0.05 significance level.

and has form $\lambda(v)$. This is the hazard function corresponding to the density of interval lengths, f. Statistical analysis for such a process is straightforward. Time series tools such as the correlogram and partial correlogram can be used to find serial dependence among successive intervals between events, though it may be clear from the context that these are independent. If independent and stationary, they can be treated as a random sample from f and inference performed in the usual way. ■

Example 6.37 (Birth process) In a birth process the intensity at time t depends on the number of previous events. Assuming that the number n of events up to t is finite, then $\lambda_{\mathcal{H}_t}(t) = \beta_0 + \beta_1 n$, where $\beta_0 > 0$, $\beta_1 \geq 0$. The complete intensity function is a step function which jumps β_1 at each event; if $\beta_1 = 0$ the process is a homogeneous Poisson process. ■

Before giving a numerical example, we briefly describe two functions useful for model checking and exploratory analysis of stationary processes.

The *variance-time curve* is defined as $V(t) = \text{var}\{N(t)\}$, for $t > 0$. A homogeneous Poisson process of intensity λ has $V(t) = \lambda t$, comparisons with which may be informative. Estimation of $V(t)$ is described in Problem 6.12.

The *conditional intensity function* is defined as

$$m_f(t) = \lim_{\delta x, \delta t \to 0} (\delta t)^{-1} \Pr\{N(t, t+\delta t) > 0 \mid N(-\delta x, 0) > 0\}, \quad t > 0,$$

which gives the intensity of events at t conditionally on there being an event at the origin. Evidently $m_f(t) = \lambda$ for a homogeneous Poisson process. An event at time t need not be the first event after that at the origin.

Example 6.38 (Japanese earthquake data) Figure 6.19 shows the times and magnitudes of earthquakes with epicentre less than 100km deep in an offshore region west of the main Japanese island of Honshū and south of the northern island of Hokkaidō. The figure shows all 483 earthquakes of magnitude 6 or more on the Richter scale in the period 1885–1980, about 5 tremors per year, in one of the most seismically active areas of Japan. A cumulative plot of the times rises fairly evenly and suggests that the data may be regarded as stationary; we shall assume this below. We take days as the units, giving $t_0 = 35{,}175$.

This is a *marked point process*, as in addition to the event times there is a *mark* — the magnitude — attached to each event. If we let the times be $0 < t_1 < \cdots < t_n < t_0$ and the associated magnitudes $m_1, \ldots, m_n$, their joint density may be written

$$\prod_{j=1}^{n} f(m_j \mid m_{(j-1)}, t_{(j)}) \prod_{j=1}^{n} f(t_j \mid m_{(j-1)}, t_{(j-1)}), \tag{6.42}$$

where $t_{(j-1)}$ and $m_{(j-1)}$ represent $t_1, \ldots, t_{j-1}$ and $m_1, \ldots, m_{j-1}$. Here we concentrate on inference for the times using the second term, leaving the magnitudes to Examples 10.7 and 10.31. The lower panels of Figure 6.19 show the estimated variance-time curve and conditional intensity function for the times, which are are clearly far from Poisson. The variance-time curve grows more quickly than for a Poisson process, indicating clustering of events, and this is confirmed by the

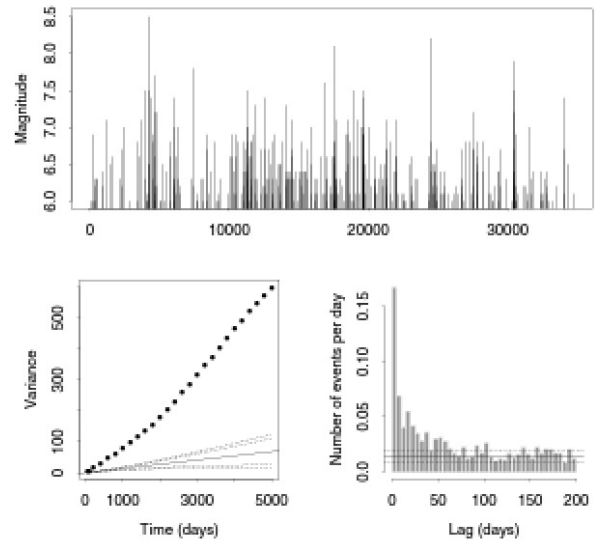

Figure 6.19 Japanese earthquake data (Ogata, 1988). The upper panel shows the times and magnitudes (Richter scale) of 483 shallow earthquakes. Lower left: estimated variance-time curve for earthquake times, with theoretical line for a Poisson process (solid) and two-sided 95% and 99% pointwise confidence limits (dots). Lower right: estimated conditional intensity, with baseline for Poisson process (solid) and two-sided 95% pointwise confidence limits (dots).

conditional intensity: for about 2–3 months after each shock the probability of another is increased.

One possible model for such data is a *self-exciting process* in which

$$\lambda_{\mathcal{H}_t}(t) = \mu + \sum_{j:t_j<t} w(t - t_j),$$

where μ is a positive constant and $w(u)$ is non-negative for $u > 0$ and otherwise zero. Here the intensity at any time is affected by the occurrence of previous events; often $w(u)$ is monotonic decreasing, so recent events affect the current intensity more than distant ones. This may be interpreted as asserting that events occur in clusters, whose centres occur as a Poisson process of rate μ. Subsidiary events are then spawned by the increase in intensity that occurs due to the superposition of the $w(t - t_j)$ for previous events. Seismological considerations suggest letting this function depend on m_j also, taking

$$w(t - t_j; m_j) = \frac{\kappa e^{\beta(m_j - 6)}}{(t - t_j + \gamma)^{\rho}}, \quad t > t_j,$$

where $\rho, \gamma, \kappa, \beta, \mu > 0$, with $\beta \doteq 2$. Under this formulation the increase in intensity depends not only on the time since an event but also on its magnitude.

IT'S OUR DAY!

Elizabeth Scott Gertrude Cox F N David

Mary Gray Lynne Billard Estela Dagum