BHF (1988) JASA

Prediction of county crop areas using survey and satellite data

- areas under corn and soybeans in 12 counties
- USDA Statistical Reporting Service
  - "exact" area of corn/soybeans in 37 segments in the 12 counties
  - segment ~ 250 hectares
- Satellite data available for all segments in the 12 counties
  - pixels are classified as corn/soybean
  - pixel ~ 0.45 hectare
- Data

y_{ij} = number of hectares of corn/soybean in segment j of county i

$i = 1, \ldots, T = 12$, $j = 1, \ldots, n_i$

x_{1ij} = number of pixels classified as corn in segment j in county i

x_{2ij} = ... soybeans

- Model

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + v_i + e_{ij}$$

$$\nu_i \sim \text{iid } N(0, \sigma_\nu^2)$$ random county specific effect

$$e_{ij} \sim \text{iid } N(0, \sigma_e^2)$$

- Matrix Formulation

$$Y = X\beta + Z\nu + e$$

$$Y = (Y_1', Y_2', \ldots, Y_T')'$$

$$X = [\underline{1}, \underset{\sim}{x}_1, \underset{\sim}{x}_2]$$

$$\beta = (\beta_0, \beta_1, \beta_2)'$$

$$Z = [Z_1', Z_2', \ldots, Z_T']'$$

where Z_i is $n_i \times 12$ and has rows equal to indicator vectors with 1 in position i

$$\underset{\sim}{\nu} = (\nu_1, \ldots, \nu_{12})'$$

$$\text{var}(Y) = V = \text{blkdiag}\left(\sigma_e^2 I_{n_i} + \sigma_\nu^2 J_{n_i}\right)$$

- Mean crop area per segment in county i conditional on ν_i is

$$\theta_i = \bar{\underset{\sim}{x}}_{i(p)}' \underset{\sim}{\beta} + \nu_i$$

where $$\bar{\underset{\sim}{x}}_{i(p)} = \frac{1}{N_i} \sum_{j=1}^{N_i} (1, x_{1ij}, x_{2ij})'$$

and N_i = total number of segments in county i

But $E(v_i \mid y_i) = g_i(\bar{y}_i - \bar{\underset{\sim}{x}}_i' \beta)$

where $g_i = \dfrac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2/n_i}$

BLUP for θ_i is

$$\tilde{\theta}_i = \bar{\underset{\sim}{x}}_{i(p)}' \tilde{\underset{\sim}{\beta}} + \tilde{v}_i$$

where $\tilde{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}Y$

and $\tilde{v}_i = g_i(\bar{y}_i - \bar{x}_i' \tilde{\beta})$

- Naive Prediction Variance

$$E\{(\tilde{\theta}_i - \theta_i)^2\} = \sigma_v^2(1-g_i) + c_i'(X'V^{-1}X)^{-1}c_i$$

where $c_i = \bar{\underset{\sim}{x}}_{i(p)} - g_i \bar{\underset{\sim}{x}}_i$

- In practice σ_v^2 and σ_e^2 also have to be estimated. The EBLUP for θ_i is

$$\hat{\theta}_i = \bar{\underset{\sim}{x}}_{i(p)}' \hat{\beta} + \hat{v}_i$$

- Survey Regression Estimate
  - fixed county specific effects.

$$\hat{\theta}_i = \bar{y}_i - (\bar{\underset{\sim}{x}}_i - \bar{\underset{\sim}{x}}_{i(p)})' \hat{\underset{\sim}{\beta}}$$

- BHF appendix – complicated formula for prediction variance that accounts for estimation of σ_v^2 and σ_e^2

Parametric Bootstrap Prediction Variances

- Simulate B "resamples" from the fitted model
- $\nu_i^* \sim N(0, \hat{\sigma}_\nu^2)$
- $e_{ij}^* \sim N(0, \hat{\sigma}_e^2)$
- $y_{ij}^* = \underset{\sim}{x}_{ij}' \hat{\beta} + \nu_i^* + e_{ij}^*$
- for each resample calculate predictions for the 12 counties; ie

$$\hat{\theta}_{i,b}^* \quad i = 1, \ldots, 12, \quad b = 1, \ldots, B$$

- Notice that the true values for each resample are known

$$\theta_{i,b}^* = \bar{x}_{i(p)}' \hat{\beta} + \nu_i^*$$

$$\hat{\theta}_{i,b}^* = \bar{x}_{i(p)}' \hat{\beta}^* + \hat{\nu}_i^*$$

- Bootstrap prediction variance for County i is

$$\frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}_{i,b}^* - \theta_{i,b}^* \right)^2$$

The BLUP and Penalized Least Squares

$$Y = X\beta + ZU + E$$

$u \sim N(0, D)$, $E \sim N(0, \sigma_0^2 I)$

- (Y, u) jointly normal with log density

$$-\frac{1}{2}\log|2\pi\sigma_0^2 I| - \frac{1}{2}\log|2\pi D|$$
$$-\frac{1}{2}\left[\frac{1}{\sigma_0^2}\|Y - X\beta - Zu\|^2 + u'D^{-1}u\right]$$

- Maximizing jointly with respect to β and u is equivalent to minimizing

$$\frac{1}{\sigma_0^2}\|Y - X\beta - Zu\|^2 + u'D^{-1}u$$

This results in

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y \qquad \text{BLUE}$$

$$\hat{u} = C'\Sigma^{-1}(Y - X\hat{\beta}) \qquad \text{BLUP}$$

where $\Sigma = \sigma_0^2 I + ZDZ'$ and $C = ZD$

(Henderson, 1950)

Penalized Spline Regression

(Reference: Ruppert, Wand, Carroll, 2003 "Semiparametric Regression")

- Want to fit a model of the form

$$y_i = f(x_i) + e_i$$

to data (x_i, y_i) $i = 1, \dots, n$

- linear regression

$$f(x) = \beta_0 + \beta_1 x$$

- least squares estimation – minimize

$$\| Y - X\beta \|^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

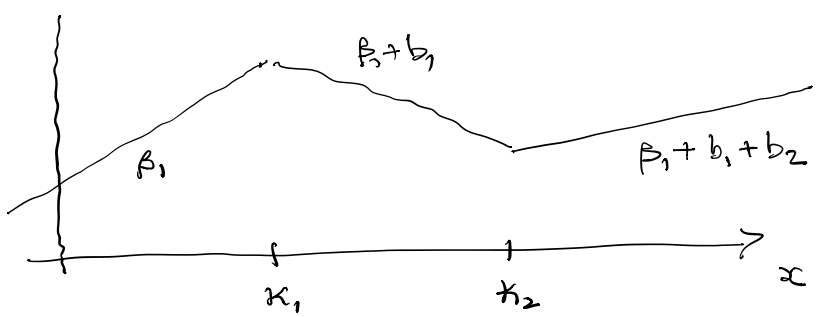
- linear spline model

$$f(x) = \beta_0 + \beta_1 x + \sum_{j=1}^{K} b_j (x - x_j)_+$$

where $\min(x_i) < x_1 < x_2 < \dots < x_K < \max(x_i)$ are "knots" and

$$(x - x_j)_+ = \max(0, x - x_j)$$

eg. $f(x) = \beta_0 + \beta_1 x + b_1 (x - x_1)_+ + b_2 (x - x_2)_+$



least squares fit with

$$X = \left(\mathbf{1}, \underset{\sim}{x}, (\underset{\sim}{x} - \mathbf{1} x_1)_+, \dots, (\underset{\sim}{x} - \mathbf{1} x_K)_+ \right)$$

$\hat{f}(x)$ has erratic changes in slope

- Penalized Least Squares

minimize $\|Y - X\beta\|^2$ subject to $\beta' A \beta \leq C$ for some psd A and $C > 0$. This is equivalent to minimizing

$$\|Y - X\beta\|^2 + \lambda \underbrace{\beta' A \beta}_{\text{penalty term}}$$

for some $\lambda > 0$

Taking derivatives

$$\frac{\partial}{\partial \beta}\left[(Y - X\beta)'(Y - X\beta) + \lambda \beta' A \beta \right]$$

$$= -2X'(Y - X\beta) + 2\lambda A \beta$$

The solution is

$$\hat{\beta} = (X'X + \lambda A)^{-1} X'Y$$