Building a Predictive Biomarker

Noah Simon and Richard Simon

July 2016

Predictive Biomarker for a Continuous Measure

Randomize $n_T + n_C = n$ patients to treatment/control

On each patient measure

y_i - single continuous outcome

(eg. blood pressure, tumor growth)

$\mathbf{x}_i$ - p-vector of features

(eg. SNPs, gene expression values)

Want to determine who will benefit from treatment over control.

Predictive Models from Prognostic Models

Given "true prognostic models":

$$y_T = \beta_0 + x^\top \beta + \epsilon$$

$$y_C = \alpha_0 + x^\top \alpha + \epsilon$$

Would treat patients with

$$\mathrm{E}\left[y_T | x\right] - \mathrm{E}\left[y_C | x\right] > \delta$$

(with $\delta = 0$ if non-toxic)

Simplifies to patients with

$$\beta_0 - \alpha_0 + x_i^\top \left(\beta - \alpha\right) > \delta$$

Estimating Predictive Models

Don't know "true models" so we estimate:

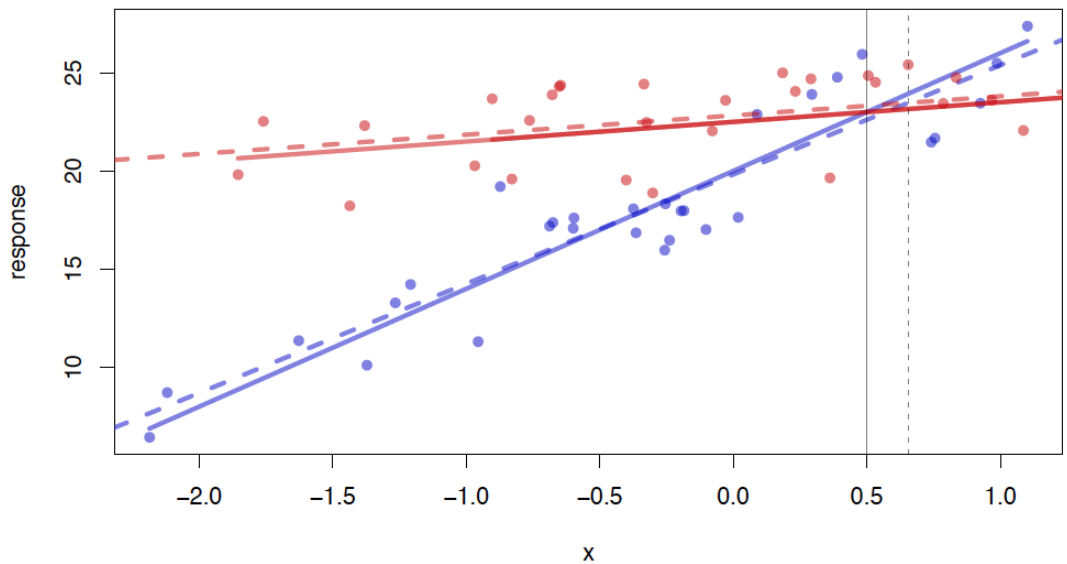
$$\mathsf{E}\left[y_T\right] = \hat{\beta}_0 + x^\top \hat{\beta}$$

$$\mathsf{E}\left[y_C\right] = \hat{\alpha}_0 + x^\top \hat{\alpha}$$

and classify new patients based on

$$\hat{\beta}_0 - \hat{\alpha}_0 + x_i^\top \left(\hat{\beta} - \hat{\alpha}\right)$$

Simple Example



Estimating Predictive Models

Can use complicated bells and whistles for prognostic models

$$\mathrm{E}\left[y_T\right] = \hat{f}_T\left(x\right)$$

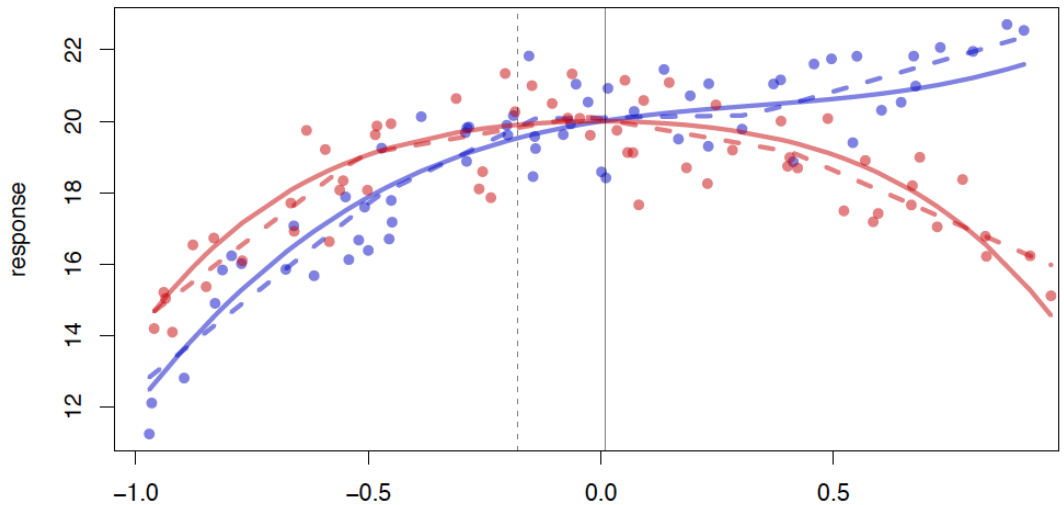
$$\mathrm{E}\left[y_C\right] = \hat{f}_C\left(x\right)$$

and classify new patients based on

$$\hat{f}_T\left(x\right) - \hat{f}_C\left(x\right)$$

Using *Lasso* and *Basis expansions* keeps that difference simple (closed form).

Nonlinear example



Separate Models vs Interaction Model

In the linear (*non-lasso*) case, separate models is equivalent to:

$$y_i = \alpha_0 + x_i^\top \alpha + \gamma_0 I_{\text{treatment}} + I_{\text{treatment}} x_i^\top \gamma + \epsilon_i$$

Useful for testing individual regression terms

Vanilla Lasso will give different solutions to this vs separate fit.

We recommend thinking in terms of separate models (though one can use other tools, penalties, priors etc. . . to share strength)

Interaction Vs Qualitative Interaction

Qualitative Interaction:

An *interaction* (shape change) between $f_T(\cdot)$ and $f_C(\cdot)$ that indicates a crossing of the two curves.

This will inform treatment decisions.

A usual interaction (shape change) does not necessarily indicate crossing.

Without crossing, an interaction will not inform treatment decisions.

Predictive Biomarker for a Binary Measure

Randomize $n_T + n_C = n$ patients to treatment/control

On each patient measure

y_i - single binary outcome

(eg. Progression after a year, pCR)

$\mathbf{x}_i$ - p-vector of features

(eg. SNPs, gene expression values)

Want to determine who will benefit from treatment over control.

Predictive Models from Prognostic Models

Given "true prognostic models":

$$\text{logit}\left(P_T\right) = \beta_0 + x^\top \beta$$

$$\text{logit}\left(P_C\right) = \alpha_0 + x^\top \alpha$$

Would treat patients with

$$\left(P_T|x\right) - \left(P_C|x\right) > \delta$$

For $\delta = 0$ still simplifies to patients with

$$\beta_0 - \alpha_0 + x^\top \left(\beta - \alpha\right) > 0$$

Predictive Biomarker for Survival Data

Randomize $n_T + n_C = n$ patients to treatment/control

On each patient measure

(t_i, s_i) - time, censoring-status

(eg. Disease free survival)

$\mathbf{x}_i$ - p-vector of features

(eg. SNPs, gene expression values)

Predictive and Prognostic Models

A bit trickier; to make things comparable assume hazard is:

$$\lambda_T(t) = h(t)e^{\beta_0 + x^\top \beta}$$

$$\lambda_C(t) = h(t)e^{x^\top \alpha}$$

Need to fit a joint model due to baseline hazard.

Make decisions based on

$$\beta_0 + x^\top (\beta - \alpha)$$

Remember lower hazard is better