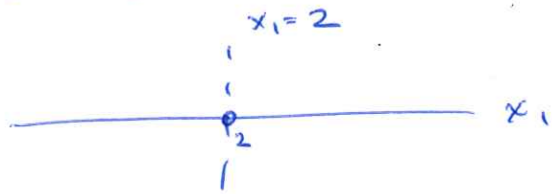
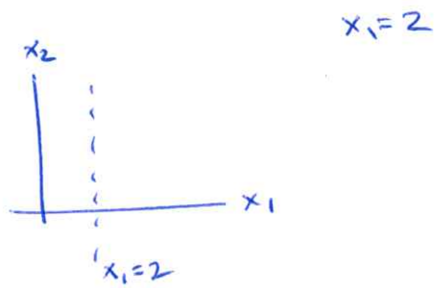
Classical linear regression may be viewed as a search for a hyperplane.

Ex: 1D (1 measurement):



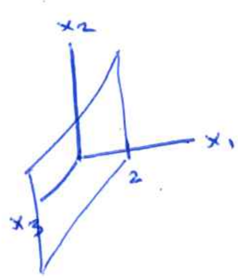
A linear equation in 1D defines a <u>pt</u> (a 0-dimensional object).

Ex: 2D (2 measurements):



A linear equation in 2D space defines a <u>line</u> (a 1-dim object).

Ex: 3D (3 measurements):



A linear eqn. in 3D space defines a <u>plane</u> (a 2D object).

Ex: p-D (p dimensional).

A linear eqn. in p-dimensional space defines a p-1 dimensional object, called a <u>hyperplane</u>.

In general, regression is modeling a conditional expectation.

Ex #1: $\mathbb{E}\,\underline{Y}|\underline{\underline{X}} = \underline{\underline{X}}\,\underline{\beta}$, where $\underline{Y} \in \mathbb{R}^n$

$\underline{\underline{X}} \in \mathbb{R}^{n \times p}$

$\underline{\beta} \in \mathbb{R}^p$

Common model:

$$\underline{Y}|\underline{\underline{X}}, \underline{\beta}, \sigma^2 \sim N_n(\underline{\underline{X}}\,\underline{\beta}, \sigma^2 \underline{\underline{I}})$$

Notice: $\underline{\beta}$ & σ^2 are unknown. In the Bayesian regression we need priors: $p(\underline{\beta}, \sigma^2)$

Ex #2: Another example to conceptualize "regression" as "modeling a cond'l expectation":

Poisson regression: $\underline{Y}|\underline{\underline{X}}, \underline{\beta} \sim \text{Poisson}(\theta(\underline{\underline{X}}, \underline{\beta}))$, i.e.

$$\mathbb{E}\,\underline{Y}|\underline{\underline{X}}, \underline{\beta} = \theta(\underline{\underline{X}}, \underline{\beta})$$

Typically, $\log \mathbb{E}[Y_i | \underline{x}_i, \underline{\beta}] = \underline{x}_i^T \underline{\beta}$

Exercise: Matching up ridge regression w/ a posterior expectation:

Let $\underline{Y}|\underline{\underline{X}}, \underline{\beta}, \sigma^2 \sim N_n(\underline{\underline{X}}\,\underline{\beta}, \sigma^2 \underline{\underline{I}})$ * typical model

$\underline{\beta}|\sigma^2, \lambda \sim N_p(\underline{0}, \frac{\sigma^2}{\lambda} \underline{\underline{I}})$ * specific prior formulation for $\underline{\beta}$.

Let's show that $\mathbb{E}[\underline{\beta}|\cdot] = \hat{\underline{\beta}}_{ridge}$

full conditional posterior expectation

ridge estimator

Exercise cont.

Recall: $$p(\underline{\beta} \mid \cdot) \propto \overbrace{p(\underline{y} \mid \underline{\underline{X}}, \underline{\beta}, \sigma^2, X)\, p(\underline{\beta} \mid \sigma^2, \lambda)\, p(\sigma^2, \lambda)}^{\text{joint distr.}}$$ by rules of probability

"everything else"

notice only these terms depend on $\underline{\beta}$

$$\Rightarrow p(\beta \mid \cdot) \propto \exp\left\{\frac{-1}{2\sigma^2}(\underline{y} - \underline{\underline{X}}\underline{\beta})^T(\underline{y} - \underline{\underline{X}}\underline{\beta})\right\} \exp\left\{\frac{-1}{2}\frac{\lambda}{\sigma^2}\underline{\beta}^T\underline{\beta}\right\}$$

$$\propto \exp\left\{\frac{-1}{2\sigma^2}\left(-2\underline{\beta}^T\underline{\underline{X}}^T\underline{y} + \underline{\beta}^T\underline{\underline{X}}^T\underline{\underline{X}}\underline{\beta}\right) - \frac{1}{2\sigma^2}\underline{\beta}^T(\lambda\underline{\underline{I}})\underline{\beta}\right\}$$

$$\propto \exp\left\{\frac{-1}{2}\underline{\beta}^T\underbrace{\left[\frac{1}{\sigma^2}\underline{\underline{X}}^T\underline{\underline{X}} + \lambda\underline{\underline{I}}\right]}_{\text{"A"}}\underline{\beta} + \underbrace{\frac{1}{\sigma^2}\underline{\beta}^T\underline{\underline{X}}^T\underline{y}}_{\text{"b"}}\right\}$$

$$\Rightarrow \underline{\beta} \mid \cdot \sim N_p\left(A^{-1}b, A^{-1}\right)$$

$$\Rightarrow \underline{\beta} \mid \cdot \sim N_p\left(\left(\underline{\underline{X}}^T\underline{\underline{X}} + \lambda\underline{\underline{I}}\right)^{-1}\underline{\underline{X}}^T\underline{y},\ \sigma^2\left(\underline{\underline{X}}^T\underline{\underline{X}} + \lambda\underline{\underline{I}}\right)^{-1}\right)$$

$$\Rightarrow \boxed{\mathbb{E}[\underline{\beta} \mid \cdot] = \left(\underline{\underline{X}}^T\underline{\underline{X}} + \lambda\underline{\underline{I}}\right)^{-1}\underline{\underline{X}}^T\underline{y} = \hat{\underline{\beta}}_{ridge}}$$ $\square$

You can check $\hat{\underline{\beta}}_{ridge} = \underset{\underline{\beta}}{\operatorname{argmin}} \|\underline{y} - \underline{\underline{X}}\underline{\beta}\|_2^2 + \lambda\underline{\beta}^T\underline{\beta}$

by differentiating w.r.t. $\underline{\beta}$ & set equal to $\underline{0}$.

It turns out, we can view many common regularization procedures as reflecting Bayesian estimators under various prior beliefs.

Consider the following common prior setup:

$$p(\underline{\beta}) = \prod_{j=1}^{p} p(\beta_j)$$

$$p(\beta_j) \propto \exp\left\{-\left|\frac{\beta_j}{\tau}\right|^{\alpha}\right\}$$

independent exp. power priors on entries of $\underline{\beta}$.

$\alpha = 2 \Rightarrow$ normal

$\alpha = 1 \Rightarrow$ Laplace

$\alpha \in (0,1) \Rightarrow$ "Bridge"

For simpler notation, consider σ^2 fixed in what follows.

$$p(\underline{\beta} \mid \underline{y}, \underline{\underline{X}}) \propto p(\underline{y} \mid \underline{\beta}, \underline{\underline{X}})\, p(\underline{\beta})$$

& $\boxed{\hat{\underline{\beta}}_{MAP} \equiv \underset{\underline{\beta}}{\operatorname{argmax}}\ p(\underline{\beta} \mid \underline{y}, \underline{\underline{X}})}$ corresponds to many well known regularization procedures.

$$\hat{\underline{\beta}}_{MAP} = \underset{\underline{\beta}}{\operatorname{argmax}}\ \log p(\underline{\beta} \mid \underline{y}, \underline{\underline{X}})$$

$$= \underset{\underline{\beta}}{\operatorname{argmax}}\ -\frac{1}{2}\|\underline{y} - \underline{\underline{X}}\underline{\beta}\|_2^2 - \sum_j \left|\frac{\beta_j}{\tau}\right|^{\alpha}$$

$$= \underset{\underline{\beta}}{\operatorname{argmin}}\ \frac{1}{2}\|\underline{y} - \underline{\underline{X}}\underline{\beta}\|_2^2 + \lambda \sum_j |\beta_j|^{\alpha} \quad ; \quad \lambda = \left|1/\tau\right|^{\alpha}$$

$\alpha = 2 \Rightarrow$ ridge reg.

$\alpha = 1 \Rightarrow$ lasso

$\alpha = 0 \Rightarrow$ best subset selection

2

Mike West (1987): exp. power prior can be represented as scale mixture of normals.

$$p(\theta) \propto e^{-|\theta|^\alpha} = \int_0^\infty e^{-s\theta^2/2} g(s)\, ds$$

for some mixing distr. $g(s)$.

often we parameterize:

$$\beta_j \mid \lambda_j, \tau \sim N(0, \lambda_j^2 \tau^2)$$

λ_j: local shrinkage parameter; τ: global shrinkage parameter

$\lambda_j \sim g(\lambda)$

so $p(\beta_j \mid \tau) \propto \exp\left\{-\left|\frac{\beta_j}{\tau}\right|^\alpha\right\} = \int \underbrace{p(\beta_j \mid \lambda_j, \tau)}_{\text{normal density}} g(\lambda)\, d\lambda$

when $\lambda_j \sim C^+(0,1)$ (1/2 Cauchy) $\Rightarrow$ horseshoe prior

when $\lambda_j \sim \exp(2) \Rightarrow$ Laplace prior

$\lambda_j \sim$ inverse-Gamma $\Rightarrow$ student-t

see Carvalho, Polson & Scott (2009) "Handling sparsity via the Horseshoe" for further reading.