Lecture 18 Math 390.03-03 4/13/16 Multivariate linear regression

$$\Rightarrow B|X \sim N_p\left(\left((X^TX)^{-1}X^T\right)(X\beta),\ (X^TX)^{-1}X^T\sigma^2 I\left((X^TX)^{-1}X^T\right)^T\right)$$

$$\Sigma = \sigma^2 \cdot (X^TX)^{-1}X^TX(X^TX)^{-1}$$

$$= N_p\left(\beta, \sigma^2(X^TX)^{-1}\right) := \frac{1}{\sqrt{(2\pi)^p|\sigma^2(X^TX)^{-1}|}} e^{-\frac{1}{2}(B-\beta)^T(\sigma^2(X^TX)^{-1})^{-1}(B-\beta)}$$

$$\propto e^{-\frac{1}{2}((X^TX)^{-1}X^TY-\beta)^T \frac{1}{\sigma^2}(X^TX)((X^TX)^{-1}X^TY-\beta)}$$

see p2

FU

Why??

One of the crown achievements of statistics ... the OLS estimator is normal, unbiased, + a whole bunch of other properties which I would go into if I was teaching a course on this...

But we can start something different..

$$P(\beta|X,Y) \propto P(Y|\beta,X)\,P(\beta|X)$$

Bayesian estimation of linear model with the 5 OLS assumptions...

Assume σ^2 known (nuisance parameter anyway)

$$P(\beta|X,y,\sigma^2) \propto P(Y|\beta,X,\sigma^2)\,P(\beta|X,\sigma^2)$$

let $P(\beta|X,\sigma^2) \propto 1$

no information

$$\frac{1}{\sqrt{(2\pi)^n|\sigma^2 I|}} e^{-\frac{1}{2}(Y-X\beta)^T(\sigma^2 I)^{-1}(Y-X\beta)}\,(1)$$

$$\propto e^{-\frac{1}{2\sigma^2}(Y^T-\underbrace{(X\beta)^T}_{\beta^TX^T})(Y-X\beta)}$$

$$= e^{-\frac{1}{2\sigma^2}} \left(Y^T X (X^T X)^{-1} - \beta^T\right)\left(X^T X\right)\left((X^T X)^{-1} X^T Y - \beta\right)$$

$$= e^{-\frac{1}{2\sigma^2}} \left(Y^T X - \beta^T X^T X\right)\left((X^T X)^{-1} X^T Y - \beta\right)$$

FOIL THIS NOW

$$= e^{-\frac{1}{2\sigma^2}} \left(Y^T X (X^T X)^{-1} X^T Y - \beta^T X^T Y - Y^T X \beta + \beta^T X^T X \beta\right)$$

$\cancel{1 \times p \quad p \times n \quad n \times 1 \qquad 1 \times n \quad n \times p \quad p \times 1}$

$\cancel{1 \times 1 \qquad\qquad 1 \times 1}$

$\cancel{(\beta^T X^T Y)^T = Y^T X \beta}$

~~since~~ $\cancel{(17.3)^T = 17.3}$

~~i.e. the transpose of a scalar equals itself . . .~~

~~these two terms are equal~~

$$\cancel{= e^{-\frac{1}{2\sigma^2}} \left(Y^T X (X^T X)^{-1} X^T Y - 2 Y^T X \beta + \beta^T X^T X \beta\right)}$$

go back to place . . .

$$\propto e^{-\frac{1}{2\sigma^2}\left(y^Ty \underbrace{\underbrace{-\beta^TX^Ty}_{1\times p\ p\times n\ n\times 1}}_{1\times 1} \underbrace{\underbrace{-y^TX\beta}_{1\times n\ n\times p\ p\times 1}}_{1\times 1} + \beta^TX^TX\beta\right)} \propto e^{-\frac{1}{2\sigma^2}\left(-2y^TX\beta + \beta^TX^TX\beta\right)}$$

Proportional since y is given

~~$(\beta^TX^Ty)^T = y^TX\beta$~~

~~$\Rightarrow \beta^TX^Ty = y^TX\beta$ since transpose of a scalar is itself~~

~~SAME GAME~~

$$\propto e^{-\frac{1}{2\sigma^2}\left(-\beta^TX^Ty - y^TX\beta + \beta^TX^TX\beta\right)}$$

$$\underbrace{e^{-\frac{1}{2\sigma^2}y^TX(X^TX)^{-1}X^Ty}}$$

Since ols is a constant

$\propto$ Kernel of $N_p\left(\cdot, \sigma^2(X^TX)^{-1}\right)$

What's the free variable now? $B \to \beta \Rightarrow$ mean is $(X^TX)^{-1}X^Ty$

So

$P(\beta \mid X, y, \sigma^2) = P(B \mid X)$ the OLS dist. under non-informative prior

Same as $\theta \mid X, \sigma^2 \sim N\left(\bar{X}, \frac{\sigma^2}{n}\right)$ under $P(\theta) \propto 1$

$P(X \mid \theta, \sigma^2) \propto e^{-\frac{1}{2\sigma^2}(X-\theta)^2} \propto N(\theta, \sigma^2)$

$P(\theta \mid X, \sigma^2) \propto e^{-\frac{1}{2\sigma^2}(X-\theta)^2} \propto N(X, \sigma^2)$

Same game

"L2 best"

Both est's from above

Derivation of Ridge Estimator

Try to demonstrate pdr of

no info above ave

$\beta \perp X, \sigma^2$ a priori

no dep of β_i, β_j

$$P(\beta \mid X, \sigma^2) = P(\beta) = N_p\left(0, \frac{\sigma^2}{m} I_p\right)$$

Same variance for all β_j's which is var of $\mathcal{E}$ modified

then we get conjugacy $\Rightarrow P(\beta_j) \sim N(0, \infty)\ \forall j$

$$\text{if } m \to 0 \Rightarrow P(\beta) \propto 1$$

just as before in σ known case

$$P(\beta) = \frac{1}{\sqrt{(2\pi)^p |\sigma_0^2 I_p|}} e^{-\frac{1}{2}(\beta - 0)^T \left(\frac{\sigma^2}{m} I_p\right)^{-1} (\beta - 0)} \quad \text{Note } I^{-1} = I$$

$$\propto e^{-\frac{1}{2} \beta^T \left(\frac{m}{\sigma^2} I\right) \beta} = e^{-\frac{1}{2\sigma^2} \beta^T (m I) \beta}$$

$$P(\beta \mid X, y, \sigma^2) \propto P(y \mid \beta, X, \sigma^2)\, \underbrace{P(\beta \mid X, \sigma^2)}_{= P(\beta)}$$

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

Note: $a^T A a + a^T B a = a^T (Aa + Ba) = a^T (A + B) a$

$$= e^{-\frac{1}{2\sigma^2}\left(-\beta^T X^T y - y^T X \beta + \beta^T (X^T X + m I) \beta\right)}$$

$$\propto e^{-\frac{1}{2\sigma^2}\left(y^T X (X^T X + mI)^{-1} X^T y - \beta^T X^T y - y^T X \beta + \beta^T (X^T X + mI) \beta\right)}$$

$$= e^{-\frac{1}{2\sigma^2}} \left(y^T X - \beta^T (X^T X + mI)\right)\left((X^T X + mI)^{-1} X^T y - \beta\right)$$

$$= e^{-\frac{1}{2\sigma^2}} \left(y^T X - \beta^T (X^T X + mI)\right)(X^T X + mI)^{-1}(X^T X + mI)\left((X^T X + mI)^{-1} X^T y - \beta\right)$$

$$= e^{-\frac{1}{2\sigma^2}} \left(\overbrace{y^T X (X^T X + mI)^{-1}}^{B_R^T} - \beta^T\right)(X^T X + mI)\left(\underbrace{(X^T X + mI)^{-1} X^T y}_{B_R} - \beta\right)$$

$$= e^{-\frac{1}{2\sigma^2}} (B_R - \beta)^T (X^T X + mI)(B_R - \beta)$$

the "ridge estimator"

$$\propto N\left(B_R, \sigma^2 (X^T X + mI)^{-1}\right) \Rightarrow \hat{\theta}_{MMSE} = \hat{\theta}_{MAP} = \hat{\theta}_{MAE} = B_R = (X^T X + mI)^{-1} X^T y$$

$\theta \in \mathbb{R}^p \quad X_1, \ldots, X_n \sim N_p(\theta, \Sigma), \quad \theta \sim N_p(\mu_0, \Sigma_0)$

$$P(\theta|X,\Sigma) \propto P(X|\theta,\Sigma)\, P(\theta|X)$$

$$= \prod_{i=1}^{n} N(\theta,\Sigma)\, N_p(\mu_0,\Sigma_0)$$

$$\propto N_p\left(\left(\Sigma_0^{-1} + n\Sigma^{-1}\right)^{-1}\left(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{X}\right), \left(\Sigma_0^{-1} + n\Sigma^{-1}\right)^{-1}\right)$$

really hard matrix algebra

if $n=1$

$$= N_p\left(\left(\Sigma_0^{-1} + \Sigma^{-1}\right)^{-1}\left(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}X\right), \left(\Sigma_0^{-1} + \Sigma^{-1}\right)^{-1}\right)$$

if $\Sigma = \sigma^2 I_p$

$$= N_p\left(\left(\Sigma_0^{-1} + \frac{1}{\sigma^2}I\right)^{-1}\left(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}X\right), \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}I\right)^{-1}\right)$$

if $\mu_0 = 0$

$$= N_p\left(\left(\Sigma_0^{-1} + \frac{1}{\sigma^2}I\right)^{-1}\left(\frac{1}{\sigma^2}X\right), \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}I\right)^{-1}\right)$$

if $\Sigma_0 = \frac{\sigma^2}{m} I_p$

$$= N_p\left(\left(\frac{m}{\sigma^2}I + \frac{1}{\sigma^2}I\right)^{-1}\left(\frac{1}{\sigma^2}X\right), \left(\frac{m}{\sigma^2}I + \frac{1}{\sigma^2}I\right)^{-1}\right) = N\left(\frac{\cancel{\sigma^2}}{m+1}\frac{1}{\cancel{\sigma^2}}X, \frac{\sigma^2}{m+1}I\right)$$

if $\mu_0 = 0$, $\Sigma_0 = \frac{\sigma^2}{m} I_p$

$$= N_p\left(\left(\frac{m}{\sigma^2} I + \Sigma^{-1}\right)^{-1}\left(\Sigma^{-1} x\right), \left(\frac{m}{\sigma^2} I + \Sigma^{-1}\right)^{-1}\right)$$

if $x = (X^T X)^{-1} X^T y$, $\Sigma = \sigma^2 (X^T X)^{-1} \Rightarrow \Sigma^{-1} = \frac{1}{\sigma^2}(X^T X)$

$$= N_p\left(\left(\frac{m}{\sigma^2} I + \frac{1}{\sigma^2}(X^T X)\right)^{-1} \frac{1}{\sigma^2}(X^T X)\left((X^T X)^{-1} X^T y\right), \left(\frac{m}{\sigma^2} I + \frac{1}{\sigma^2}(X^T X)\right)^{-1}\right)$$

$$= N_p\left(\underbrace{\cancel{\sigma^2}\left(X^T X + mI\right)^{-1} \frac{1}{\cancel{\sigma^2}} \cancel{(X^T X)(X^T X)^{-1}} X^T y}, \sigma^2 \left(X^T X + mI\right)^{-1}\right)$$

Ridge estimator
derived as
a special
case
of the
general
conjugacy
formula

What does Ridge do?

$m \to 0$ $\quad \beta \sim N(0, \infty I) \Rightarrow \beta_j \sim N(0,\infty)$ i.e. uninformative

$$B_R \to B = (X^TX)^{-1}X^Ty$$

$m \to \infty$ super informative

$$B_R \to 0 \qquad (X^TX + mI) \to \begin{bmatrix} \text{Big} & & \text{Insignificant} \\ & \text{Big} & \\ \text{Insignificant} & \ddots & \text{Big} \end{bmatrix}^{-1} X^Ty$$

$$= \begin{bmatrix} 0 & & \text{Insignificant} \\ & 0 & \\ \text{Insignificant} & \ddots & 0 \end{bmatrix} X^Ty \approx \vec{0}$$

So $b_R \in [\vec{0}, \vec{B}]$ with $m\uparrow \Rightarrow$ more shrinkage to 0.

Aboard... if two matrices are full rank

$$(A+B)^{-1} = A^{-1} - \frac{1}{1+t} A^{-1}BA^{-1} \quad \text{s.t. } t := \text{trace}(BA^{-1})$$

$$(\underbrace{X^TX}_{A} + \underbrace{mI}_{B})^{-1} = (X^TX)^{-1} - \frac{1}{1+mT}(X^TX)^{-1}(mI)(X^TX)^{-1}$$

$$\text{trace}(mI(X^TX)^{-1}) = m\,\underbrace{\text{trace}((X^TX)^{-1})}_{T}$$

$$B_R = \left((X^TX)^{-1} - \frac{m}{1+mT}(X^TX)^{-1}(X^TX)^{-1}\right)X^Ty = B - \frac{m}{1+mT}(X^TX)^{-1}B = B\left(I - \frac{m}{1+mT}(X^TX)^{-1}\right)$$

Before we had $\to$ sqd error loss

$$b := \operatorname{argmin} \overset{l(\varepsilon)}{\overset{\shortparallel}{\varepsilon^T \varepsilon}} = \operatorname{argmin} (y - Xb)^T (y - Xb) = \operatorname{argmin} y^T y - 2\beta^T X^T y + \beta^T X^T X \beta$$

What if we have ... $\|\varepsilon\|_2^2 + m\|\beta\|_2^2$ $\leftarrow$ norm form $\Leftrightarrow$ min $\varepsilon^T\varepsilon$ s.t. $\beta^T\beta < c$

$l(\varepsilon) + m\, l(\beta)$ $\longrightarrow$ "Ridge loss function"

$$b = \operatorname{argmin} \varepsilon^T \varepsilon + m \beta^T \beta$$

$$= \operatorname{argmin} y^T y - 2\beta^T X^T y + \beta^T X^T X \beta + m \beta^T \beta$$

$$= y^T y - 2\beta^T X^T y + \beta^T (X^T X + mI) \beta$$

$$\nabla(\ ) = -2X^T y + 2(X^T X + mI)\beta \overset{set}{=} 0$$

$$\Rightarrow (X^T X + mI)\beta = X^T y \Rightarrow b_R := (X^T X + mI)^{-1} X^T y$$

Ridge derived as a minimum of the ridge loss function which is similar to the sqd. err. loss

---

Ridge loss is due to: $\boxed{\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \text{ NORM}}$

$\Leftrightarrow$ min $\|\varepsilon\|_2^2$ s.t. $\|\beta\|_2^2 < c$ "L2-penalized regression"

Setting up a Lagrange... (1811)

$$\min f(\beta; X, y) \quad \text{s.t.} \quad \underbrace{g(\beta) < c}_{g(\beta) - c = 0} \qquad \nabla := \begin{bmatrix} \frac{\partial}{\partial \beta_0} \\ \vdots \\ \frac{\partial}{\partial \beta_p} \end{bmatrix}$$

$$\mathcal{L}(\beta, X, y, \lambda) = f(\beta; X, y) + \lambda (g(\beta) - c)$$

$$0 \overset{set}{=} \nabla \mathcal{L} = \nabla \|\varepsilon\|_2^2 + \lambda \nabla \|\beta\|_2^2 = \nabla (\|\varepsilon\|^2 + \lambda\|\beta\|^2)$$

$\frac{\partial \mathcal{L}}{\partial \lambda} = 0$ to solve for λ in terms of c.

let $m := \lambda$ "the Lagrange multiplier"

MIDTERM ?

FINAL