Lecture 18

Recall that if $\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$ & we have OLS assumptions (meaning $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$ & $Y|X \sim N_p(X\beta, \sigma^2 I_n)$, then

① $b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} = (X^T X)^{-1} X^T y$

② $B = (X^T X)^{-1} X^T Y$

③ $B | X \sim N_p(\vec{\beta}, \sigma^2 (X^T X)^{-1})$

$$:= \frac{1}{\sqrt{(2\pi)^p |\sigma^2 (X^T X)^{-1}|}} \cdot e^{\frac{1}{2}(B-\beta)^T (\sigma^2 (X^T X)^{-1})^{-1} (B-\beta)}$$

free variable parameter

And the above density is proportional to

$$p(\beta|X) \propto e^{-\frac{1}{2\sigma^2}(Y^T X (X^T X)^{-1} - \beta^T)(Y^T X)(X^T X)^{-1} X^T Y - \beta^T X^T Y - Y^T X \beta + \beta^T X^T X \beta}$$

Now recall $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta)$

For us,

$$p(B|Y, \sigma^2, X) \propto P(Y|\beta, X, \sigma^2) \cdot p(\beta|X, \sigma^2)$$

fixed

Let us use an uninformative prior: $p(\beta_0) \propto 1$, $p(\beta_1) \propto 1$, $\dots$, $p(\beta_{p-1}) \propto 1$ $\Big\}$ $p(\vec{\beta}) \propto 1$

So In the uninformative case, $p(\vec{\beta}|X, \sigma^2) \propto 1$

Then $p(\beta|Y, \sigma^2, X) \propto p(Y|\beta, X, \sigma^2)$ (1)

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

$$\vdots$$

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

parameter free variable

$$\propto P(B|X) = N(\beta, \sigma^2 (X^T X)^{-1})$$

This is similar to the following: recall if $X|\theta, \sigma^2 \sim N(\theta, \sigma^2)$ then $p(\theta|X, \sigma^2) \propto p(X|\theta, \sigma^2)\, p(\theta, \sigma^2) \propto p(X|\theta, \sigma^2) \propto N(X, \sigma^2)$

uninformative prior so $p(\theta, \sigma^2) \propto 1$

So $P(\beta | X, Y, \sigma^2) = P(B|X)$ under an uninformative prior.
Your free variable & parameter switched!

What if we have an informative prior? Let
$$P(\beta | X, \sigma^2) = N_p(\mu_0, \tfrac{\sigma^2}{m} I)$$
$$= N_p(\vec{0}, \tfrac{\sigma^2}{m} I)$$ 32:00

Then $p(\beta | X, Y, \sigma^2) \propto p(Y | \beta, X, \sigma^2)\, p(\beta | X, \sigma^2)$

$$\propto N_p(B_R, \sigma^2 (X^T X + mI)^{-1})$$

where
$$B_R := (X^T X + mI)^{-1} X^T y \neq B = (X^T X)^{-1} X^T Y$$

& $\hat{\beta}_{mmse} = \hat{\beta}_{MAE} = \hat{\beta}_{MAP} = B_R$

We call B_R the "RIDGE ESTIMATOR"

As $m \to 0$ $B_R = (X^T X + mI)^{-1} X^T y \longrightarrow (X^T X)^{-1} X^T y = B$. Like uninformative prior.

As $m \to \infty$, $B_R = 0 X^T y = 0$ b/c $\frac{\sigma^2}{m} I \to 0$ so you are locked around a tight $\vec{\mu}$ & here, $\vec{\mu} = 0$.

Thm (NOT on exam): If $\theta \in \mathbb{R}^p$, $X_1, \ldots, X_n \overset{\text{each}}{\sim} N_p(\theta, \Sigma)$ & $\theta \sim N_p(\mu_0, \Sigma_0)$ then

$$\theta | X, \Sigma \sim N\left((\Sigma_0^{-1} + \Sigma^{-1})^{-1} (\Sigma_0^{-1} \mu_0 + n \Sigma^{-1} \bar{x}),\ (\Sigma_0^{-1} + n\Sigma^{-1})^{-1} \right)$$

For us, $p = n$ & $n = 1$.

Also not on our exam: If $b := \operatorname{argmin}\{\varepsilon^T \varepsilon\}$
then $\beta = (X^T y + mI)^{-1} X^T y$

shouldn't that be b?
What is he trying to show here?