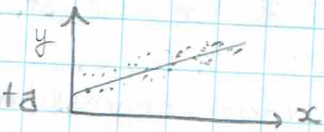
LECTURE 17

Recap: Before we were doing the normal model.
We got our data X_1 which looked normal 
so we assumed $\mathcal{F} = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\theta)^2} \,\middle|\, \theta \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+ \right\}$ & attempted
to estimate our parameters θ, σ^2. We then got our
conjugate family: Normal InvGamma.
We had 2 goals
1) INFERENCE (i.e. figure out posterior)
2) Predict $X^* \mid X$

data is univariate, i.e. $\{49.7, 50.1, \text{etc}\}$

We then looked at bivarate data 
meaning we had $\{ \langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \ldots, \langle x_n, y_n \rangle \}$
We assumed $\mathcal{F} = \{ \beta_0 + \beta_1 X \mid \beta_0, \beta_1 \in \mathbb{R} \}$ & our parameters
were $\sigma^2, \beta_0, \beta_1$. We then predict $y^* \mid y, x, x^*$ where

$$y^* \mid y, x, x^* = \int_{\vec{\theta}} P(y^* \mid x^*, x, y, \vec{\theta}) \, P(\vec{\theta} \mid x, y) \, d\vec{\theta}.$$

Now our "data" is y here.

Let $\vec{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $\vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$, $\vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$, $\vec{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$, $\vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$

Before we had

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$$
$$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$$
$$\cdots$$
$$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n$$

In vector format, we write this as

$$\vec{Y} = \beta_0 \vec{1} + \beta_1 \vec{X} + \vec{\varepsilon}$$

If we define $X := \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$ then $X\vec{\beta} = \underset{n \times 2}{\begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}} \underset{2 \times 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}} = \underset{n \times 1}{[\beta_0 + \beta_1 X_1, \ldots, \beta_0 + \beta_1 X_n]}$

so $\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$

Let us guess $\begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$ using the least squares penalty.

We do this by minimizing $\sum \varepsilon_i^2$ = minimizing $\vec{\varepsilon}^T \vec{\varepsilon}$
Which is done by solving $\nabla(\vec{\varepsilon}^T \vec{\varepsilon}) = 0$.
Solving this equation yields

$$\boxed{b = \hat{\beta} = (X^T X)^{-1} X^T Y}$$

Rewriting the last equation is

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

and we let

$$B = (X^T X)^{-1} X^T Y$$

Note b depends on the realization Y, whereas the r.v. B depends on the r.v. Y

We call b the least squared estimates & we call B the least squared estimator

Recall univariate regression: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ← 1 regressor, x

Now $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$ ← p regressors

So blood pressure depends on weight, age, glucose, etc

Define

$$y = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$$

Now you want to minimize ε so you solve $\nabla(\varepsilon^T \varepsilon) = 0$. Doing so yields

$$b = (X^T X)^{-1} X^T y \quad \& \quad B = (X^T X)^{-1} X^T Y$$

which is THE SAME SOLUTION AS BEFORE. So the solution in the univariate case equals the solution in the multivariate case!

Note in the multivariate case $b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}$ where b_i approximates β_i

Moving on, let us again consider our OLS assumptions.

Assumptions: $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$

This implies

$$\left.\begin{aligned} Y_1 \mid \vec{X}_1 &\sim N(\vec{X}_1 \vec{\beta}, \sigma^2) \\ Y_2 \mid \vec{X}_2 &\sim N(\vec{X}_2 \vec{\beta}, \sigma^2) \\ &\vdots \\ Y_n \mid \vec{X}_n &\sim N(\vec{X}_n \vec{\beta}, \sigma^2) \end{aligned}\right\} \text{independent but NOT identically distributed.}$$

We will now take a big leap... we are going stick all these normal r.v.'s into 1 single r.v.

- $Y \mid X \sim N_n(X\vec{\beta}, \sigma^2 I_n)$ which is called a multivariate normal r.v.

* All we are doing is sticking a bunch of r.v.'s into a vector. *

As an example let $S = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$ where S_i is a univariate r.v.

Define $E[S] = \begin{bmatrix} E(S_1) \\ E(S_2) \end{bmatrix}$, define $E(cS) = cE(S)$,

define $E[\vec{a}^T S] = E\left[[a_1 \ a_2] \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \right] = a^T E[S]$

$Var[S_1] = E[S_1 - E[S_1]]^2$

$Var[S] := \begin{bmatrix} var[S_1] & cov(S_1, S_2) \\ cov(S_1, S_2) & var[S_2] \end{bmatrix}$ where $cov(S_1, S_2) = E[S_1 - \mu_1] E[S_2 - \mu_2]$

For simplicity we write

$Var[S] = [\{cov[S_i, S_j]\}]$

- $Cov(S_i, S_i) = S_i$

If S_i & S_j are independent then $Cov[S_i, S_j] = 0$

If S_i are all independent then

$Var[S] = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix}$ if identical varience then $\sigma^2 I_n = var[S]$

Now,

$$Y \mid X \sim N_n(X\vec{\beta}, \sigma^2 I_n) = MVN(\vec{\mu}, \Sigma)$$

$$:= \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot e^{-\frac{1}{2}(X - \vec{\mu})^T \Sigma^{-1} (X - \vec{\mu})}$$

If $\Sigma = \sigma^2 I$ then $N_n(\mu, \sigma^2 I) \propto e^{-\frac{1}{2\sigma^2}(X - \vec{\mu})^T (X - \vec{\mu})}$

If $\dim[\mu] = 1$ then you get back the univariate normal.

Now if $X \sim N_n(\mu, \Sigma)$ what is the distribution of AX for $A \in \mathbb{R}^{p \times n}$ constants.

Recall for a r.v. X, $M_X(t) := E(e^{tX})$

Now If X is a vector r.v., $M_{\vec{X}}(\vec{t}) = E(e^{\vec{t}^T \vec{X}})$

and $M_X(t) = M_Y(t) \iff X \overset{d}{=} Y$

For $X \sim N_n(\mu, \Sigma)$, $M_X(t) = e^{t^T \mu + \frac{1}{2} t^T \Sigma t}$

It follows $AX \sim N_n(A\mu, A\Sigma A^T)$. Also,

① $E[AX] = A\, E[X]$

② $Var[AX] = A \Sigma A^T$

How does this help us? Why would we need this? We will use this

Recall $B = \underbrace{(X^T X)^{-1} X}_{A} Y$ ← really this is $B|X$

$= AY$

So $B|X = AY = A \cdot Y|X$

Now $Y|X \sim N_n(X\vec{\beta}, \sigma^2 I_n)$ so

$B|X = AY = (AX\vec{\beta}, A\sigma^2 I_n A^T)$

$= ((X^T X)^{-1} X X \vec{\beta}, (X^T X^{-1}) X \sigma^2 I_n ((X^T X^{-1}) X)^T)$

$= N_n(\beta, \sigma^2 (X^T X^{-1}))$