Lecture 17 Math 390.03-02 4/11/16



We just did "simple"/"univariate" linear regression and logistic regression.

What did we do? Recap...

Recall norm-norm model...


What does this data look like?

$\{52.72, 48.91, \ldots\}$ i.e. "univariate data"

① Assume $\mathcal{F}$ if... (normal) $\Rightarrow$ $\theta, \sigma^2$ as params. Do we know either? No...

② $\Rightarrow$ Conjugate $G_\theta$ (normal) $\Rightarrow$ $\mu_0, m, a_0, \sigma_0^2$ as params which we set via

" $G_{\sigma^2}$ (Inv-Gamma)

$\Rightarrow$ nonconj. (?)

① objective $\Rightarrow$ $\mu_0 = 0, m \approx 0$

② Real prior subjective belief

③ Empirical Bayes

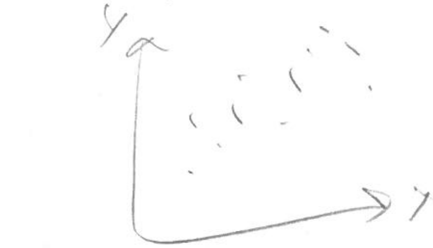
## Goals

(I) Figure out nature i.e. $\mathcal{F}$ i.e. $\langle \theta, \sigma^2 \rangle$

(II) Predict for the future (# or set of $X^*$'s) $\Rightarrow$ $\langle \theta, \sigma^2 \rangle$ nuisances to be marginalized out when considering...

$$P(X^* \mid X) = \int_\theta \int_{\sigma^2} \cdots \, d\theta \, d\sigma^2$$

In regression... the data is




$\{\langle 13.27, 15.97\rangle, \langle 9.14, 8.99\rangle, \dots\}$

the data ($Y$) now has other features tagged to it $\Rightarrow$ reason why $Y$ is the notation choice for data in most Bayesian textbooks $\rightarrow$ modern

① Assume $F|X$ (normal) $\Rightarrow$ params $\sigma^2, \beta_0, \beta_1$ of which we don't know any of them

all the $n$ $x$'s

Bayesian model? Hold on...

---

dim $= n$

let $\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, $\vec{\mathcal{E}} = \begin{bmatrix} \mathcal{E}_1 \\ \vdots \\ \mathcal{E}_n \end{bmatrix}$, $\vec{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

$\vec{a} = \vec{b}$

Def: equality of vectors $\Rightarrow \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \Rightarrow \forall i \; a_i = b_i$

$\Rightarrow \vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x} + \vec{\mathcal{E}}$

Def: addition of vectors $\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ \vdots \\ \vdots \end{pmatrix}$

let $\vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$

Def: scalar mult. of vectors $c\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} ca_1 \\ \vdots \end{pmatrix}$

$\Rightarrow \vec{y} = [\vec{1} : \vec{x}]\, \vec{\beta} + \vec{\mathcal{E}}$

the colon represents column-binding into a "matrix" := many col's put together

matrix times vector

$\begin{bmatrix} \beta_0 1 + \beta_1 x_1 \\ \beta_0 1 + \beta_1 x_2 \\ \vdots \\ \beta_0 1 + \beta_1 x_n \end{bmatrix}$

Note: same exact thing as we did before just in matrix notation

let $X := [\vec{1} : \vec{x}] = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ $\Rightarrow \vec{y} = X\vec{\beta} + \vec{\mathcal{E}}$

$\dim[X] = n \times 2$ matrix

Guess $\begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$ using l.s. Previously we minimized sqd. err.

$$\vec{\varepsilon} = \vec{y} - X\vec{\beta}$$

quantity: $\varepsilon_1^2 + \varepsilon_2^2 + \dots \varepsilon_n^2$

In vector form...

$\vec{\varepsilon}^T \vec{\varepsilon}$ what is this?

transpose turns a col into a row (and vice versa)

$$a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \Rightarrow a^T = [a_1 \ a_2 \ a_3]$$

dim A

$$a^T a = [a_1 \ a_2 \ a_3] \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = a_1^2 + a_2^2 + a_3^2 = \sum_{i=1}^{n} a_i^2$$

$$\Rightarrow \varepsilon^T \varepsilon = \sum \varepsilon_i^2 = (\vec{y} - X\vec{\beta})^T (\vec{y} - X\vec{\beta})$$

$$= (y^T - \beta^T X^T)(y - X\beta)$$

$$= y^T y - \underbrace{\beta^T X^T y}_{\substack{(1\times2)(2\times n)(n\times1) \\ 1\times1}} - \underbrace{y^T X \beta}_{\substack{(1\times n)(n\times2)(2\times1) \\ 1\times1}} + \beta^T X^T X \beta$$

$\Rightarrow (\beta^T y^T y)^T = \beta^T x^T y$ since transpose of a scalar is itself

$$\Rightarrow = y^T \beta X$$

$$= y^T y - 2\underset{(1\times2)(2\times n)(n\times1)}{\beta^T X^T y} + \beta^T \underset{(2\times n)(n\times2)}{X^T X} \beta$$

$\vec{\nabla}(\ \ ) \overset{\text{set}}{=} \vec{0}$ s.t. $\vec{\nabla} := \begin{bmatrix} \partial\beta_0 \\ \partial\beta_1 \end{bmatrix}$ "Del"/"Nabla" operator

$$\beta^T X^T y = \beta^T a = [\beta_0 \ \beta_1] \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = a_0\beta_0 + a_1 \beta_1 \Rightarrow \nabla \beta^T a = a$$ general form $\forall a$

$$\nabla \beta^T X^T y = X^T y$$

quadratic form

4A

$$\beta^T A \beta = \begin{bmatrix}\beta_0 & \beta_1\end{bmatrix}\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} = \begin{bmatrix}\beta_0 & \beta_1\end{bmatrix}\begin{bmatrix}a_{11}\beta_0 + a_{12}\beta_1\\ a_{21}\beta_0 + a_{22}\beta_1\end{bmatrix} = a_{11}\beta_0^2 + a_{12}\beta_0\beta_1 + a_{21}\beta_0\beta_1 + a_{22}\beta_1^2$$

$$\nabla \beta^T A \beta = \begin{bmatrix}2a_{11}\beta_0 + (a_{12}+a_{21})\beta_1\\ (a_{12}+a_{21})\beta_0 + 2a_{22}\beta_1\end{bmatrix} = \begin{bmatrix}2a_{11} & a_{12}+a_{21}\\ a_{21}+a_{12} & 2a_{22}\end{bmatrix}\beta$$

$$A = X^T X = \begin{bmatrix}1 & \cdots & 1\\ x_1 & \cdots & x_n\end{bmatrix}\begin{bmatrix}1 & x_1\\ \vdots & \vdots\\ 1 & x_n\end{bmatrix} = \begin{bmatrix}n & \Sigma x_i\\ \Sigma x_i & \Sigma x_i^2\end{bmatrix}$$

$$\begin{matrix}2n & 2\Sigma x_i\\ 2\Sigma x_i & 2\Sigma x_i^2\end{matrix} = 2X^T X$$

$$\Rightarrow \nabla \beta^T X^T X \beta = 2X^T X \beta$$ (genl law $\forall X$ matrices)

$$\Rightarrow -\cancel{2}X^T y + \cancel{2}X^T X\beta = 0 \Rightarrow X^T X \beta = X^T y \Rightarrow$$

$$Ay = x \Rightarrow A^{-1}Ay = A^{-1}x \Rightarrow Iy = A^{-1}x \Rightarrow y = A^{-1}x$$

$$I_n := \begin{bmatrix}1 & & 0\\ & \ddots & \\ 0 & & 1\end{bmatrix} \qquad I_n \vec{a} = \begin{pmatrix}1 & & 0\\ & \ddots & \\ 0 & & 1\end{pmatrix}\begin{pmatrix}a_1\\ \vdots\\ a_n\end{pmatrix} = \begin{pmatrix}a_1\\ a_2\\ \vdots\\ a_n\end{pmatrix}$$

$A^{-1}$ := matrix inverse($A$) := matrix s.t. $A^{-1}A = I$

inverting matrices ... 3 lectures in lin alg. 101

$$\Rightarrow \vec{b} = (X^T X)^{-1} X^T \vec{y}$$ ← compact form for L.S. est's!

$$\vec{b} = \begin{bmatrix}b_0\\ b_1\end{bmatrix}$$

$$\Rightarrow \beta = (X^T X)^{-1} X^T Y$$

Recall model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i \in \{1, \ldots, n\}$$

What if we had more than one feature "$x$", "$x_1, x_2, \ldots, x_p$"

$p$ features which combine linearly (with additive noise) to create $y$:

$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$ = multivariate linear regression

vector-matrix

$$\Rightarrow \vec{y} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{p1} \\ 1 & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & & x_{pn} \end{bmatrix}}_{\underbrace{[\vec{1} : \vec{x}_1 : \ldots : \vec{x}_p]}_{X}} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \varepsilon$$

$\Rightarrow \vec{y} = X\vec{\beta} + \vec{\varepsilon}$ Same as before!!

$$\Rightarrow \underbrace{b = \underset{(p \times p)}{(X^T X)^{-1}} \underset{(p \times n)}{X^T} \underset{(n \times 1)}{y}}_{} = p \times 1$$

$y_1 = x_1 . \beta + \varepsilon_1$

$y_2 = x_2 . \beta + \varepsilon_2$

$\vdots$

$y_n = x_n . \beta + \varepsilon_n$

rows of $X$

Vector notation

Super-powerful!

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

Now predict again $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$ (the three OLS assumptions)

$$\left.\begin{aligned} \Rightarrow Y_1 | X_{1\cdot} &\sim N(x_1 . \beta, \sigma^2) \\ Y_2 | X_{2\cdot} &\sim N(x_2 . \beta, \sigma^2) \\ &\vdots \\ Y_n | X_{n\cdot} &\sim N(x_n . \beta, \sigma^2) \end{aligned}\right\}$$ all independent (but not ident. distr.)

6

$$\Rightarrow \underset{n\times 1}{Y} \,|\, \underset{n\times 1}{X} \sim N_n\left(\overset{n\times p}{X}\overset{p\times 1}{\beta},\ \sigma^2 \overset{n\times n}{I_n}\right)$$

vector - generalized r.v. ($n$ dimensional case) **MVN**

seen expectation is $n$-dim vector

but variance is a $n\times n$ matrix

WAIT...

$$\dim(X) = n \Rightarrow \dim(E(X)) = n \Rightarrow \dim(Var(X)) = n\times n$$

---

let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$

SKIP...

$$E(X) := \begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix} = \overset{\mu}{\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}}$$ expectation is defined to be element-wise

scalar

$c \in \mathbb{R}$, $E(cX) = \begin{pmatrix} c\,E(X_1) \\ c\,E(X_2) \end{pmatrix} = \begin{bmatrix} c\mu_1 \\ c\mu_2 \end{bmatrix} = c\mu$

---

$a \in \mathbb{R}^n$,

$E[\underbrace{a^T}_{\text{row vector}} X] = E[a_1X_1 + \dots + a_nX_n] = \sum_{i=1}^n a_i\mu_i = a^T\mu$

SKIP

$$A \in \mathbb{R}^{p\times n},\ E[AX] = E\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pn} \end{bmatrix}\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = E\begin{bmatrix} \sum_{i=1}^n a_{1i}X_i \\ \vdots \\ \sum_{i=1}^n a_{pi}X_i \end{bmatrix} = \begin{bmatrix} E[\ ] \\ \vdots \\ E[\ ] \end{bmatrix} = \begin{bmatrix} a_1^T\mu \\ \vdots \\ a_p^T\mu \end{bmatrix} = \begin{bmatrix} a_1^T \\ \vdots \\ a_p^T \end{bmatrix}\mu = A\mu$$

$\Rightarrow \beta = (X^TX)^{-1}X^TY$

$E[\beta|X] = E[(X^TX)^{-1}X^TY|X]$

const

$= (X^TX)^{-1}X^T(X\beta) = \boxed{\beta}$ LS/MLE unbiased

$\dim[X] = n$ ..

$$Var[X] := \begin{bmatrix} Var[X_1] & Cov[X_1, X_2] & & \\ Cov[X_2, X_1] & Var[X_2] & & \\ & & \ddots & \\ & & & Var[X_n] \end{bmatrix} = \left[\{Cov[X_i, X_j]\}\right]$$

If $Cov > 0 \Rightarrow$ pos lin. assoc

$Cov < 0 \Rightarrow$ neg lin assoc

$= 0 \Rightarrow$ no lin. association

Recall $Cov[X_1, X_2] := E(X_1 - \mu_1)(X_2 - \mu_2) = E[X_1 X_2] - \mu_1 \mu_2$

Special case

$Cov[X_1, X_1] = E(X_1 - \mu_1)(X_1 - \mu_1) = E(X_1 - \mu)^2 = Var[X_1]$

Special case:

If $X_1, X_2$ indep

$p(X_1, X_2) = p(X_1) p(X_2) \Rightarrow E[X_1 X_2] = \mu_1 \mu_2$

$\Rightarrow Cov[X_1, X_2] = 0$

Special case

If all $X_i, X_j$ indep,

$$\Sigma := Var(X) = \begin{bmatrix} \sigma_1^2 & & O \\ & \ddots & \\ O & & \sigma_n^2 \end{bmatrix} = I \overset{\text{vec}}{\sigma^2} \quad \text{if } \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2 \Rightarrow \underset{\text{scalar}}{\Sigma = \sigma^2} I_n$$

$a \in \mathbb{R}^n$ col vec

SKIP

---

$$\underbrace{Var[\overset{1\times n}{a^T} \overset{n\times 1}{X}]}_{1\times 1} = Var[a_1 X_1 + \ldots a_n X_n] = a_1^2 Var[X_1] + \ldots + a_n^2 Var[X_n] \quad \in \mathbb{R}$$

$+ a_i a_j Cov[X_i, X_j] \quad \forall i \neq j$ i.e. dim 1

why?

$= a_i a_j \underbrace{Cov[X_i, X_j]}_{\Sigma_{ij}} \quad \forall i \forall j$

$\dim[E(a^T X)] = 1$

quadratic form

$$\underbrace{\overset{}{\underset{1\times n}{b^T}} \underset{n\times n}{A} \underset{n\times 1}{b}}_{1\times 1} = b^T \begin{bmatrix} a_1 \cdot b \\ a_2 \cdot b \\ \vdots \\ a_n \cdot b \end{bmatrix} = b^T \begin{bmatrix} \Sigma a_{1i} b_i \\ \Sigma a_{2i} b_i \\ \vdots \\ \Sigma a_{ni} b_i \end{bmatrix} = \sum_{j=1}^n \sum_{i=1}^n a_{ij} b_i b_j$$

$\Rightarrow Var[a^T X] = a^T \Sigma a$

SKIP

$$\underbrace{\mathrm{Var}\left[\overset{\parallel Y}{AX}\right]}_{\substack{p\times n \quad n\times 1 \\ p\times p}} \quad \mathrm{dim}: p\times p$$

let $Y =$

$$\mathrm{Var}\left(\begin{bmatrix} a_{1\cdot} \\ a_{2\cdot} \\ \vdots \\ a_{p\cdot} \end{bmatrix} \begin{bmatrix} X \end{bmatrix}\right) = \mathrm{Var}\begin{bmatrix} a_{1\cdot}X \\ a_{2\cdot}X \\ \vdots \\ a_{p\cdot}X \end{bmatrix} = \begin{bmatrix} a_{1\cdot}\Sigma a_{1\cdot}^T & & & \\ & a_{2\cdot}\Sigma a_{2\cdot}^T & & \\ & & \ddots & \\ & & & a_{p\cdot}\Sigma a_{p\cdot}^T \end{bmatrix}$$

$$\mathrm{Cov}\left[a_{i\cdot}X, a_{j\cdot}X\right]$$

$$:= E\left[a_{i\cdot}X \, a_{j\cdot}X\right] - E\left[a_{i\cdot}X\right]E\left[a_{j\cdot}X\right]$$

$$= E\left[\sum_{k=1}^{n}(a_{ik}X_k)\sum_{k=1}^{n}(a_{jk}X_k)\right] - \sum a_{ik}\mu_k \sum a_{jk}\mu_k$$

foil it...

$$= \sum_{l=1}^{n}\sum_{k=1}^{n} \mathrm{Cov}\left[a_{ik}X_k, a_{jl}X_l\right]$$

$$= \sum_{l=1}^{n}\sum_{k=1}^{n} \underbrace{a_{ik}}_{b_k}\underbrace{a_{jl}}_{c_k}\Sigma_{kl} = a_{i\cdot}\Sigma a_{j\cdot}$$

$$\underset{1\times n \quad n\times n \quad n\times 1}{b\,A\,c} = b\begin{bmatrix} a_{1\cdot}c \\ a_{2\cdot}c \\ \vdots \\ a_{n\cdot}c \end{bmatrix} = b\begin{bmatrix} \sum a_{1i}c_i \\ \sum a_{2i}c_i \\ \vdots \\ \sum a_{ni}c_i \end{bmatrix} = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ji}c_ib_j$$

(80)

SKIP

$$A \Sigma A^T$$

$p \times n \quad n \times n \quad n \times p = p \times p$

$$\begin{bmatrix} — a_{1\cdot} — \\ — a_{2\cdot} — \\ \vdots \\ — a_{p\cdot} — \end{bmatrix} \begin{bmatrix} \\ \Sigma \\ \\ \end{bmatrix} \begin{bmatrix} | & | & & | \\ a_{1\cdot}^T & a_{2\cdot}^T & \cdots & a_{p\cdot}^T \\ | & | & & | \end{bmatrix}$$

$$= A \begin{bmatrix} \Sigma a_{1}^T \\ \Sigma a_{2}^T \\ \vdots \\ \Sigma a_p^T \end{bmatrix} = \begin{bmatrix} a_{1\cdot} \Sigma a_{1}^T & a_{2\cdot} \Sigma a_{1}^T & \cdots & a_{p\cdot} \Sigma a_{1}^T \\ & & & \\ & & & \end{bmatrix} = Var(AX)$$

$\forall$ dim $n$ r.v. $X$ and $A$ dim $p \times n$ scalar matrices,

nice formula!!

$$Var[AX] = A \Sigma A^T \quad \text{s.t.} \quad \Sigma := Var(X)$$

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

$$Var[B \mid X, \sigma^2] = Var[\underbrace{(X^T X)^{-1} X^T}_{\text{const}} Y \mid X, \sigma^2] = (X^T X)^{-1} X^T \underset{\sigma^2 I}{\Sigma} \left((X^T X)^{-1} X^T\right)^T$$

$$= \sigma^2 \cancel{(X^T X)^{-1}} \cancel{(X^T X)} (X^T X)^{-1})^T$$

proof?

$B \mid X \sim$ ? get expectation/var but not distr.

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

go back to explain this to p7

9

Def $X \sim N_n(\mu, \Sigma)$ or $X \sim MVN_n(\mu, \Sigma)$

is a n-dim multivariate normal r.v with mean vector $\mu$

and var-cov matrix $\Sigma := Var[X]$ with PDF

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

special case:
if $n=1$
$\Sigma = [\sigma^2] \Rightarrow \Sigma^{-1} = \frac{1}{\sigma^2}$
$|\Sigma| = \sigma^2$, $(X-\mu)^T(X-\mu) = (X-\mu)^2$
$\Rightarrow X \sim N(\mu, \sigma^2)$

n dim free variable

$$M_X(t) := E\left[e^{\overset{1\times n}{t^T} \overset{n\times 1}{X}}\right] = \ldots = e^{t^T\mu + \frac{1}{2}\underbrace{t^T\Sigma t}_{\text{quadratic form again!}}}$$

$A \in \mathbb{R}^{p \times n}$

$$\underbrace{\overset{p\times n}{A}\overset{n\times n}{X}}_{p\times n} \sim ?$$

$$E\left[e^{\overset{1\times p}{t^T} \overset{p\times n}{A} \overset{n\times 1}{X}}\right] = E\left[e^{\overset{1\times n}{s^T} \overset{n\times 1}{X}}\right] = M_X(s) = e^{s^T\mu + \frac{1}{2}s^T\Sigma s}$$

$s^T := t^T A \Rightarrow s = A^T t$

$$= e^{t^T A\mu + \frac{1}{2} t^T A\Sigma A^T t}$$

$$= e^{t^T\mu' + \frac{1}{2}t^T\Sigma' t}$$

$cX$

$M_{cX}(t) = E[e^{tcX}] = E[e^{t'X}] = M_X(t') = M_X(ct)$

$$\Rightarrow AX \sim N(A\mu, A\Sigma A^T)$$

generally, $E[AX] = A\mu$, $Var[AX] = A\Sigma A^T$

(prove this on homework using def & properties of MGF)

Ex. $\beta | X \sim ?$ $\quad \beta|X = (X^TX)^{-1}X^T \underbrace{Y|X}_{N_n(X\beta, \sigma^2 I_n)}$

const