Bayesian inference in simple conjurgate families.

Preview: All you need to know

① Bayes' rule: $p(\theta \mid x) = \dfrac{P(x \mid \theta) P(\theta)}{P(x)}$ $\boxed{\propto P(x \mid \theta) P(\theta)}$

referred to as Bayes' rule*

② Gaussian kernel: $\exp\{-\frac{W}{2}(x-\mu)^2\}$

③ Beta kernel: $x^{\alpha-1}(1-x)^{\beta-1}$

④ Gamma kernel: $x^{\alpha-1}\exp\{-\beta x\}$

⑤ Two-step method for finding posteria:

A. Write the kernel of $p(\theta \mid x)$ using tip 1;

B. match this kernel with an exist distribution

to derive normalization constant.

(A) $p(w \mid x_1, \cdots, x_N) \propto P(x_1, \cdots, x_N \mid w) \cdot P(w)$ (tip 1)

Beta-Bernoulli $= \left[\prod_{i=1}^{N} P(x_i \mid w)\right] P(w)$ ($x_1, \cdots, x_N$ ind.)

$$\propto \left[\prod_{i=1}^{N} w^{x_i}(1-w)^{(1-x_i)}\right] w^{a-1}(1-w)^{b-1}$$

$$= w \wedge \{\textstyle\sum_{i=1}^{N} x_i + a - 1\} \, (1-w) \wedge \{N - \textstyle\sum_{i=1}^{N} x_i + b - 1\}$$

which is the kernel of Beta $(\sum_{i=1}^{N} x_i + a,\ N + b - \sum_{i=1}^{N} x_i)$

(B)
$$\begin{cases} y_1 = x_1/(x_1+x_2) \\ y_2 = x_1 + x_2 \end{cases} \Rightarrow \begin{cases} x_1 = y_1 y_2 \\ x_2 = y_2 - y_1 y_2 \end{cases}$$

$$J = \begin{vmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 \end{vmatrix} = \begin{vmatrix} y_2 & y_1 \\ -y_2 & 1-y_1 \end{vmatrix} = y_2$$

Tip for Jacobian: consider the integrators are arranged in a column vector,

originally, we're integraling on $(dx_1, dx_2)'$, since

$$\begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix} = \begin{bmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 \end{bmatrix} \begin{bmatrix} dy_1 \\ dy_2 \end{bmatrix}$$

and the new integration is taken on $(dy_1, dy_2)'$, the Jacobian is sort of "amount of scaling", which is the determinant of the transformation mat

Back to problem (B). With Jacobian J, we have

$f_Y(y_1, y_2) = f_X(y_1 y_2, y_2 - y_1 y_2) \cdot |J|$.

Calculations omitted, finally, we have

$f_Y(y_1) = \int f_Y(y_1, y_2) dy_2 \sim Ga(a_1 + a_2, 1)$

$f_Y(y_2) = \int f_Y(y_1, y_2) dy_1 \sim Beta(a_1, a_2)$

So we learn from this problem:

① $Ga(a_1, 1) / [Ga(a_1, 1) + Ga(a_2, 1)] \sim Ga(a_1 + a_2, 1)$

② Gamma $\mapsto$ Beta: $Ga(a_1, 1) + Ga(a_2, 1) \sim Beta(a_1, a_2)$

(C) Normal - Normal

$$p(\theta) \propto \exp\left\{-\frac{(\theta - m)^2}{2V}\right\},$$

$$p(\underset{\sim}{x}|\theta) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right\} = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n [(x_i - \bar{x})^2 + (\bar{x} - \theta)^2]\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}[S_x + n(\bar{x} - \theta)^2]\right\}.$$

Here $\underset{\sim}{x}$ denotes vector $(x_1, x_2, \cdots, x_n)'$, $S_x = \sum_{i=1}^n (x_i - \bar{x})^2$ and $\bar{x}$ is the mean value of elements of $\underset{\sim}{x}$. By Bayes' rule*,

$$p(\theta|\underset{\sim}{x}) \propto \exp\left\{-\frac{(\theta - m)^2}{2V} - \frac{S_x}{2\sigma^2} - \frac{n(\bar{x} - \theta)^2}{2\sigma^2}\right\}$$

$$= \exp\left\{\theta^2\left(-\frac{1}{2V} - \frac{n}{2\sigma^2}\right) + \theta\left(\frac{m}{V} + \frac{n\bar{x}}{\sigma^2}\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\cdot\left[\frac{1}{V} + \frac{n}{\sigma^2}\right]\cdot\left(\theta - \frac{m/V + n\bar{x}/\sigma^2}{1/V + n/\sigma^2}\right)^2\right\}$$

$$\sim N\left(\frac{m/V + n\bar{x}/\sigma^2}{1/V + n/\sigma^2}, \text{ precision} = \frac{1}{V} + \frac{n}{\sigma^2}\right)$$

The normal posteria is a combination of prior and likelihood, with 1. additive precision, i.e., precision of posteria equals the sum of precisions of prior and likelihood; 2. convex combined mean, i.e., the mean of output distribution is a convex combination of the means of component distributions, with each weight proportional to precision.

(D) Inverse Gamma – Normal.

$p(w) \propto w^{a-1} e^{-bw}$

$p(\underline{x}|w) \propto \exp\{-\frac{w}{2}[S_x + n(\bar{x}-\theta)^2]\} \cdot w^{n/2}$

$\Rightarrow p(w|\underline{x}) \propto w^{\frac{n}{2}+a-1} \exp\{-[\frac{S_x+n(\bar{x}-\theta)^2}{2}+b]w\}$

$\sim Ga(n/2+a, \frac{S_x+n(\bar{x}-\theta)^2}{2}+b)$

(E) Normal – Normal, generalized sample sd.

Following steps in problem (C), we have

$p(\theta|\underline{x}) \propto \exp\{-\frac{1}{2}[\frac{1}{v}+\sum_{i=1}^{n}\frac{1}{\sigma_i^2}]\cdot(\theta - \frac{m/v + \sum_{i=1}^{n} x_i/\sigma_i^2}{1/v + \sum_{i=1}^{n} 1/\sigma_i^2})^2\}$

$\sim N(\frac{m/v + \sum_{i=1}^{n} x_i/\sigma_i^2}{1/v + \sum_{i=1}^{n} 1/\sigma_i^2}, \text{ precision} = \frac{1}{v} + \sum_{i=1}^{n}\frac{1}{\sigma_i^2})$.

The result gives a more clear demonstration of the conclusion we reached at (C).

(F) $p(x,w) = p(x|w)p(w)$

$\propto w^{1/2} \exp\{-\frac{w}{2}(x-m)^2\} w^{a/2-1} \exp\{-\frac{b}{2}w\}$

$= w^{\frac{a+1}{2}-1} \exp\{-\frac{(x-m)^2+b}{2} w\}$,

which is a Gamma function w.r.t. w, so when we integral out w, what is left will be

$p(x) = \int_0^{+\infty} p(x,w)\,dw \propto [\frac{b}{2} + \frac{(x-m)^2}{2}]^{-\frac{a+1}{2}}$

$\propto [1 + \frac{1}{a}\cdot\frac{(x-m)^2}{b/a}]^{-\frac{a+1}{2}}$, which is the kernel of t distribution,

i.e., $\nu = a$, $m = m$, $s^2 = b/a$ ← scale

center

degrees of freedom

t distribution as a scale mixture of Gaussian:

$p(x) = \int p(x|w)p(w)dw \triangleq \sum \underbrace{Pr(x|w\in\Delta W_i)}_{\text{Mixture component}} \cdot \underbrace{Pr\{w\in\Delta W_i\}}_{\text{weight}}$

We use this property to deal with outliers.

The multivariate normal distribution. - Basics

(A) $\mathrm{cov}(\underline{x}) = \mathbb{E}(\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})' = \mathbb{E}(\underline{x}\underline{x}') - \mathbb{E}(\underline{x})\underline{\mu}' - \underline{\mu}\,\mathbb{E}(\underline{x})' + \underline{\mu}\underline{\mu}'$

$= \mathbb{E}(\underline{x}\underline{x}') - 2\underline{\mu}\underline{\mu}' + \underline{\mu}\underline{\mu}' = \mathbb{E}(\underline{x}\underline{x}') - \underline{\mu}\underline{\mu}'$

(B) $f(\underline{z}) = f(z_1)f(z_2)\cdots f(z_p)$ (idependence of $z_1 \sim z_p$)

$= (2\pi)^{-p/2}\exp\{-\frac{1}{2}\sum_{i=1}^{p} z_i^2\} = (2\pi)^{-p/2}\exp\{-\frac{1}{2}\underline{z}'\underline{z}\}$ (PDF)

$M_{\underline{z}}(\underline{t}) = \mathbb{E}[\exp\{\underline{t}'\underline{z}\}] = \mathbb{E}[\exp\{t_1z_1\}\exp\{t_2z_2\}\cdots\exp\{t_pz_p\}]$

$= \mathbb{E}[\exp\{t_1z_1\}]\cdot\mathbb{E}[\exp\{t_2z_2\}]\cdots\mathbb{E}[\exp\{t_pz_p\}]$ (MGF)

(C) $\Rightarrow$: $\underline{x} \sim N(\underline{\mu}, \Sigma)$, then $\forall\, \underline{a} \in \mathbb{R}^p/\{0\}$, $\underline{a}'\underline{x}$ is univariate normal,

i.e., MGF of $z = \underline{a}'\underline{x}$ is $M_z(t) = \exp\{\mathbb{E}(z)t + \frac{1}{2}\mathrm{var}(z)t^2\}$, where

$\mathbb{E}(z) = \mathbb{E}(\underline{a}'\underline{x}) = \underline{a}'\underline{\mu}$, $\mathrm{var}(z) = \underline{a}'\Sigma\underline{a}$. so

$M_z(t) = \mathbb{E}(\exp(zt)) = \exp\{\underline{a}'\underline{\mu}t + \frac{1}{2}\underline{a}'\Sigma\underline{a}t^2\}$.

On the other hand,

$M_z(\lambda) = \mathbb{E}[\exp((\underline{a}\lambda)'\underline{x})] = \exp\{(\underline{a}\lambda)'\underline{\mu} + \frac{1}{2}(\underline{a}\lambda)'\Sigma(\underline{a}\lambda)\}$

$= \exp\{\underline{t}'\underline{\mu} + \frac{1}{2}\underline{t}'\Sigma\underline{t}\}$, $\underline{t} = \underline{a}\lambda$,

which means

$M_{\underline{x}}(\underline{t}) = \mathbb{E}[\exp(\underline{t}'\underline{x})] = \exp\{\underline{t}'\underline{\mu} + \frac{1}{2}\underline{t}'\Sigma\underline{t}\}$

$\Leftarrow$: If $\underline{x}$ has MGF $M_{\underline{x}}(\underline{t}) = \exp\{\underline{t}'\underline{\mu} + \frac{1}{2}\underline{t}'\Sigma\underline{t}\}$, then for $\forall\, \underline{a} \in \mathbb{R}^p/\{0\}$

and $z = \underline{a}'\underline{x}$,

$M_z(\lambda) = \mathbb{E}[\exp\{\lambda z\}] = \mathbb{E}[\exp\{(\lambda\underline{a})'\underline{x}]$

$= M_{\underline{x}}(\underline{t} = \lambda\underline{a}) = \exp\{\lambda\underline{a}'\underline{\mu} + \frac{1}{2}\lambda^2\cdot\underline{a}'\Sigma\underline{a}\}$

Since z has mean $\underline{a}'\underline{\mu}$ and variance $\underline{a}'\Sigma\underline{a}$, the MGF of z can be rewritten by $M_z(t) = \exp\{\lambda\mathbb{E}(z) + \frac{1}{2}\lambda^2\mathrm{var}(z)\}$, which indicates z is a univariate Gaussian variable.

Above proves $\{$"any linear combinition of $\underline{x}$'s components is univar normal"

$=$ "$\underline{x}$ is multivar normal"$\}$

$\Longleftrightarrow$ $\{$"$\underline{x} \sim N(\underline{\mu}, \Sigma)$" $=$ "MGF of $\underline{x}$ is $\exp(\underline{t}'\underline{\mu} + \frac{1}{2}\underline{t}'\Sigma\underline{t})$"$\}$

(D) I'm skeptical about the "full col rank" statement here, in that as far as I can see, it is more appropriate for L to be "full row rank" than "full col rank".

Basically, the idea is to show that $\forall \underline{a} \in \mathbb{R}^p / \{0\}$, $\underline{b}' = \underline{a}' L$ is also non-zero, which requires the ROWS of L being linearly independent.

(E) Without loss of generality, let $\underline{X}$ follows a non-singular normal distribution, i.e., $\underline{X} \sim N(\underline{\mu}, \Sigma)$ and Σ is non-singular.

$\Rightarrow$ Σ is positive definite. Then we can apply spectral decomposition to Σ. Let $v_1, v_2, \cdots, v_p$ are eigen vectors of Σ, then:

(i) Σ is symmetric $\Rightarrow$ eigen vectors with different eigen values are orthogonal;

(ii) for eigenvectors with the same eigen value, apply Gram-Schmidt orthogonalization;

(iii) normalize all the eigen vectors.

At this stage we will have an orthogonal mat $P = [v_1, v_2, \cdots, v_p]$, where $\|v_i\| = 1$, $v_i' v_j = 0$, $\forall i \neq j \in \{1, 2, \cdots, p\}$, and

$$\Sigma = P \Lambda P' \quad \text{(Spectral Decomposition)},$$

$$\Lambda = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_p\}.$$

Consider the affine transformation

$$\underline{X}^* = L\underline{Z} + \underline{\mu}^*, \text{ where } \underline{Z} \text{ is standard multivariate normal, } \underline{\mu}^* = \underline{\mu}$$

and $L = P\Lambda^{1/2}$, then by what we proved in (D),

$\mathrm{rank}(P\Lambda^{1/2}) = p \Rightarrow \underline{X}^* \sim N(\underline{\mu}^*, (P\Lambda^{1/2})(P\Lambda^{1/2})') = N(\underline{\mu}^*, P\Lambda P')$

As mean and var characterize a normal distribution, so $\underline{X}^*$ is just $\underline{X}$

In this way, for any $\underline{X} \sim N(\underline{\mu}, \Sigma)$, we can construct the affine transformation $(L, \underline{\mu})$ s.t. $\underline{X} = L\underline{Z} + \underline{\mu}$.

(F) For $\underline{x} \sim N(\underline{\mu}, \Sigma)$, we know from (E) that we are able to transform them back to standard normal variables:

$$\underline{x} = L\underline{z} + \underline{\mu} \iff \underline{z} = L^{-1}(\underline{x} - \underline{\mu}),\ L^{-1} = \Lambda^{-1/2}P'$$

Using the R.V. transformation property,

$$f_{\underline{x}}(\underline{x}) = f_{\underline{z}}(\underline{z}(\underline{x})) \cdot |J|_{x \to z}$$

$$= (2\pi)^{-p/2} \exp\{-\tfrac{1}{2}(\underline{x}-\underline{\mu})' P\Lambda^{-1/2}\Lambda^{-1/2}P'(\underline{x}-\underline{\mu})\} \cdot |\Lambda^{-1/2}P'|$$

$$= (2\pi)^{-p/2} \exp\{-\tfrac{1}{2}(\underline{x}-\underline{\mu})'\Sigma^{-1}(\underline{x}-\underline{\mu})\}\ |\Lambda^{-1/2}| \cdot |P|$$

$$= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\tfrac{1}{2}(\underline{x}-\underline{\mu})'\Sigma^{-1}(\underline{x}-\underline{\mu})\}\quad (*)$$

$(*)$: $|P| = 1$, $|\Sigma| = |P\Lambda P'| = |P|\cdot|\Lambda|\cdot|P'| = |\Lambda|$, $|\Lambda^{1/2}| = |\Lambda|^{1/2}$

$\Rightarrow |\Lambda^{-1/2}| = |\Lambda|^{-1/2} = |\Sigma|^{-1/2}$

(G) Both matrices A and B are full ROW rank, not full col rank.

Let the row number of both A and B be p, then $\text{rank}(A) = \text{rank}(B) = p$.

A, B full row rank $\iff$ Any $\underline{l} \in \mathbb{R}^p/\{0\}$, $\underline{l}'A$ and $\underline{l}'B$ are not zero.

By the definition of "multivariate normal" in (C), it's safe to claim

$\underline{x}_1, \underline{x}_2$ are multivariate Gaussian; A, B are of full row rank $\}$ Any $\underline{l} \in \mathbb{R}^p/\{0\}$, $(\underline{l}'A)\underline{x}_1$ and $(\underline{l}'B)\underline{x}_2$ are univariate normal R.V.'s

$\iff$ both $A\underline{x}_1$ and $B\underline{x}_2$ are multivariate normal.

$\iff$ MGF of $A\underline{x}_1$ is $\exp\{\underline{t}'(A\underline{\mu}_1), \tfrac{1}{2}\underline{t}'(A\Sigma_1A')\underline{t}\}$

MGF of $B\underline{x}_2$ is $\exp\{\underline{t}'(B\underline{\mu}_2), \tfrac{1}{2}\underline{t}'(B\Sigma_2B')\underline{t}\}$

Beyond that, $\underline{x}_1$ is independent with $\underline{x}_2$, indicating that $A\underline{x}_1$ is independent with $B\underline{x}_2$, then MGF of $A\underline{x}_1 + B\underline{x}_2$ is:

$$M_{A\underline{x}_1+B\underline{x}_2}(\underline{t}) = M_{A\underline{x}_1}(\underline{t}) \cdot M_{B\underline{x}_2}(\underline{t}) = \exp\{\underline{t}'(A\underline{\mu}_1 + B\underline{\mu}_2) + \tfrac{1}{2}\underline{t}'(A\Sigma_1A' + B\Sigma_2B')\underline{t}\}$$

From the MGF of $A\underline{x}_1 + B\underline{x}_2$, we know it's a multivariate Gaussian R.V. with mean $A\underline{\mu}_1 + B\underline{\mu}_2$ and variance $A\Sigma_1A' + B\Sigma_2B'$.

6/10

The multivariate normal distribution - Conditionals and marginals.

(A) " Any marginal distribution of a partition of $\underline{x}$ is normal distribution if $\underline{x} \sim N(\underline{\mu}, \Sigma)$ ".

Let $A = [\, I_K \,\vdots\, 0_{k\times(p-k)} \,]$. Clearly A has size $k \times p$ and is of full row rank. By what we have proven in (G) of last section, we have

$$\underline{x_1} = A\underline{x} \text{ follows } N(A\underline{\mu}, A\Sigma A') = N(\underline{\mu_1}, \Sigma_{11})$$

(B) As long as x follows a non-singular multivariate normal distribution, it's safe to assert that both Σ_{11} and Σ_{22} are invertible.

Then by the inverse of block matrix:

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1} & (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \end{bmatrix}$$

we have

$$\Omega_{11} = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}, \quad \Omega_{12} = -\Omega_{11}\Sigma_{12}\Sigma_{22}^{-1}$$

$$\Omega_{22} = (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}, \quad \Omega_{21} = -\Omega_{22}\Sigma_{21}\Sigma_{11}^{-1}$$

(C) $\log f(\underline{x_1} \mid \underline{x_2}) = \log f(\underline{x_1}, \underline{x_2}) - \log f(\underline{x_2})$

$$= c + \left(-\tfrac{1}{2}(\underline{x} - \underline{\mu})'\Omega(\underline{x} - \underline{\mu})\right) - \left(-\tfrac{1}{2}(\underline{x_2} - \underline{\mu_2})'\Omega_{22}(\underline{x_2} - \underline{\mu_2})\right)$$

$$= c - \tfrac{1}{2}\left[(\underline{x_1} - \underline{\mu_1})'\Omega_{11}(\underline{x_1} - \underline{\mu_1}) + 2(\underline{x_1} - \underline{\mu_1})'\Omega_{12}(\underline{x_2} - \underline{\mu_2})\right]$$

$$= c - \tfrac{1}{2}(\underline{x_1} - \underline{\mu_1}^*)'\Omega_{11}(\underline{x_1} - \underline{\mu_1}^*), \text{ where } \underline{\mu}^* \text{ satisfies}$$

$$2\underline{x_1}'\Omega_{11}\underline{\mu_1}^* = 2\underline{x_1}'\Omega_{11}\underline{\mu_1} - 2\underline{x_1}'\Omega_{12}(\underline{x_2} - \underline{\mu_2})$$

$$= 2\underline{x_1}'\Omega_{11}\left(\underline{\mu_1} - \Omega_{11}^{-1}\Omega_{12}(\underline{x_2} - \underline{\mu_2})\right)$$

so $x_1 \mid x_2$ follows a multivariate Gaussian distribution with mean $\underline{\mu_1} - \Omega_{11}^{-1}\Omega_{12}(\underline{x_2} - \underline{\mu_2})$ and var Ω_{11}. It is also the distribution of a regression target of model

$$y = \underbrace{(\underline{\mu_1} + \Omega_{11}^{-1}\Omega_{12}\underline{\mu_2})}_{\text{Intercept}} - \underbrace{(\Omega_{11}^{-1}\Omega_{12})}_{\text{slope}} x + \underbrace{\epsilon}_{\text{where randomness comes from}}, \quad \epsilon \sim N(0, \Omega_{11})$$

Multiple regression: three classical principles for inference.

(A) Probably the coolest aspect of linear regression!

Least squares: $\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^n (y_i - \underline{x_i}'\beta)^2 \right\}$

Maximum of Gaussianity:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\max} \left\{ \prod_{i=1}^n p(y_i \mid \beta, \sigma^2) \right\} = \underset{\beta \in \mathbb{R}^p}{\arg\max} \ \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \underline{x_i}'\underline{\beta})^2 \right\}$$

$$= \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^n (y_i - \underline{x_i}'\underline{\beta})^2 \right\},$$ which reduces to Least Squares problem.

Method of moments

Denote matrix by $J_n = \{\frac{1}{n}\}_{n\times n}$, then $(I - J_n)$ is the "centralize operator" i.e., $(I - J_n)\underline{x} = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x})'$. It's not hard to see that matrix $(I - J_n)$ is symmetric and idempotent, i.e.,

$$\begin{cases} (I - J_n)' = (I - J_n) \\ (I - J_n)^m = (I - J_n) \end{cases}$$

Denote error by vector $\underline{e}$, $\underline{e} = \{e_i\}_{n\times 1}$, $e_i = y_i - \underline{x_i}'\underline{\beta}$, or $\underline{e} = \underline{y} - \mathbf{X}\underline{\beta}$, $\underline{x_i}'$ is the i-th column of $\mathbf{X}$. Then,

$\mathrm{cov}(\underline{x}, \underline{e}) = 0 \iff \mathbf{X}'(I - J_n)'(I - J_n)\underline{e} = 0$, by the symmetry and idempotency of $(I - J_n)$, we have

$\mathrm{cov}(\underline{x}, \underline{e}) = 0 \iff \mathbf{X}'(I - J_n)'\underline{e} = 0$

$$\iff \begin{bmatrix} x_{11} - \bar{x}_{\cdot 1} & x_{21} - \bar{x}_{\cdot 1} & \cdots & x_{n1} - \bar{x}_{\cdot 1} \\ x_{12} - \bar{x}_{\cdot 2} & x_{22} - \bar{x}_{\cdot 2} & \cdots & x_{n2} - \bar{x}_{\cdot 2} \\ \vdots & & & \\ x_{1p} - \bar{x}_{\cdot p} & x_{2p} - \bar{x}_{\cdot p} & \cdots & x_{np} - \bar{x}_{\cdot p} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (*)$$

If we use standardized data, i.e., $\bar{x}_{\cdot 1} = \bar{x}_{\cdot 2} = \cdots = \bar{x}_{\cdot p} = \bar{y} = 0$, then Eq $(*)$ is equivalent to Least Squares statement.

(B) $\underset{\beta \in \mathbb{R}^p}{\arg\min} \{ \sum_{i=1}^{n} w_i (y_i - \underline{x_i}'\underline{\beta})^2 \}$

$= \underset{\beta \in \mathbb{R}^p}{\arg\max} \exp\{-(\underline{y} - X\underline{\beta})' W (\underline{y} - X\underline{\beta})\}$, $W = \mathrm{diag}\{w_1, \cdots, w_p\}$

The RHS can be interpreted as maximizing likelihood of a heteroscedastic Gaussian with precision matrix W, and

$$W = \Sigma^{-1} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)^{-1} = \mathrm{diag}(1/\sigma_1^2, \cdots 1/\sigma_p^2)$$

Quantifying uncertainty: some basic frequentist ideas.

In linear regression

Notation: Both Y and β are vectors, X is data matrix.

(A) $\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\, \mathcal{L}$ where $\mathcal{L} = (Y - X\beta)'(Y - X\beta)$, then

$\nabla_\beta \mathcal{L} \,|_{\beta = \hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0 \Rightarrow \hat{\beta} = (X'X)^{-1}X'Y$

since $Y = X\beta + \epsilon \sim \mathcal{N}(X\beta, \sigma^2 I)$, then

$\hat{\beta} = (X'X)^{-1}X'Y \sim \mathcal{N}((X'X)^{-1}(X'X)\beta,\ (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1}) = \mathcal{N}(\beta, (X'X)^{-1}\sigma^2$

(B) Apparently, the standard errors for each β_j is just the square root of the (j, j) element in $(X'X)^{-1}\sigma^2$. By the relationship between adjoin and inverse of a matrix, if mat A is invertible then

$$A^{-1} = \frac{\mathrm{Adj.}\,A}{|A|}$$

so the (j, j) element in A^{-1} is $\mathrm{Adj.}\,A\,[j, j] / |A|$.

Going back to our problem,

$\mathrm{var}(\beta_j) = \sigma^2 \frac{\mathrm{Adj.}\,(X'X)\,[j, j]}{|X'X|} = \sigma^2 \frac{|M_{jj}| \cdot (-1)^{(j+j)}}{|X'X|} = \frac{\sigma^2 |M_{jj}|}{|X'X|}$, when M_{jj} is the minor of $X'X$ that removes the j-th row and j-th column in $(X'X)$. Denote the i-th column in X by X_i, then $(X'X)[i, j] = X_i'X_j$, therefore by removing the j-th row and column from $(X'X)$, we are virtually doing $(X_{-j}'X_{-j})$, where X_{-j} denotes the remaining cols in X after removing column j

As a result, $\mathrm{var}(\beta_j) = \sigma^2 \cdot \frac{|X_{-j}'X_{-j}|}{|X'X|}$

Propagating uncertainty.

(A) $\underline{\theta} = (\theta_1, \cdots, \theta_p)'$, $f(\underline{\theta}) = \theta_1 + \theta_2 \iff f(\underline{\theta}) = [1, 1, 0, \cdots, 0]\, \underline{\theta}$,

then $var(f(\hat{\underline{\theta}})) = [1, 1, 0, \cdots, 0]\, \hat{\Sigma} \begin{bmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \hat{\Sigma}_{11} + 2\hat{\Sigma}_{12} + \hat{\Sigma}_{22}$

Similarly, if $f(\theta) = [1, 1, \cdots, 1]\, \underline{\theta}$, then

$var(f(\hat{\underline{\theta}})) = [1, 1, \cdots, 1]\, \hat{\Sigma} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = sum_{i=1, j=1}^{p \quad p} \hat{\Sigma}_{ij}$ (i, j are allowed to be identical)

(B) What (A) tells is how to calculate var of $f(\underline{\theta})$ when f itself is a linear function w.r.t. $\underline{\theta}$. So the idea of calculating var of $f(\underline{\theta})$ when f is not linear is to make a linear approximation of f and calculate the var of this linear approximation.

$f(\hat{\underline{\theta}}) \approx f(\underline{\theta}) + (\nabla_\theta f)'(\hat{\underline{\theta}} - \underline{\theta})$. (1-st order Taylor approximation)

Then $var(f(\hat{\underline{\theta}})) \approx var\left[f(\underline{\theta}) + (\nabla_\theta f)'(\hat{\underline{\theta}} - \underline{\theta}) \right]$

$= (\nabla_\theta f)'\, var(\hat{\theta} - \theta)\, (\nabla_\theta f)$

Caveats:

(i) To make linear approximation make sense, f should not change violenty (have a large second order derivative) w.r.t. $\underline{\theta}$; and $\hat{\underline{\theta}}$ should not be far from $\underline{\theta}$, so $\hat{\underline{\theta}}$ almost must be an unbiased estimator for $\underline{\theta}$.