Stats Modeling II, Section I

Jack Liu

February 16, 2018

Exercise 1.1. *Clearly, all iid sequences are exchangeable, but not all exchangeable sequences are iid. Consider an urn, containing r red balls and b blue balls. A sequence of colors is generated by repeatedly sampling a ball from the urn, noting its color, and then returning the ball, plus another ball of the same color, to the urn. Show that the resulting sequence is exchangeable, but not iid.*

Proof. Let r be the number of red balls in the urn initially, b be the number of blue balls in the urn initially, and $n = r + b$. Clearly, $P(X_1 = Red) = r/n$. Next, $P(X_2 = Red) = \frac{r}{n}\frac{r+1}{n+1} + \frac{b}{n}\frac{r}{n} = \frac{r}{n}$. Thus, by induction, $P(X_i = Red)$ is the same for all i, which implies probability for the sequence of colors is exchangeable. However, the sequence is by definition not independent. □

Exercise 1.2. *We will start off with a finite sequence $(X_1, \ldots, X_M)$. For any $N \leq M$, show that*

$$\mathbf{P}\left(\sum_{i=1}^{N} X_i = s \Big| \sum_{i=1}^{M} X_i = t\right) = \frac{\binom{t}{s}\binom{M-t}{N-s}}{\binom{m}{n}}$$

Proof.

$$\mathbf{P}\left(\sum_{i=1}^{N} X_i = s \Big| \sum_{i=1}^{M} X_i = t\right) = \frac{\mathbf{P}\left(\sum_{i=1}^{N} X_i = s, \sum_{i=1}^{M} X_i = t\right)}{\mathbf{P}\left(\sum_{i=1}^{M} X_i = t\right)}$$

Because the sequences are exchangeable, we don't care about the order of sequence. The probability reduces to the number of ways the top can occur divided by the number of ways the bottom can occur:

$$\frac{\binom{N}{s}\binom{M-s}{t-s}}{\binom{m}{t}} = \frac{\binom{t}{s}\binom{M-t}{N-s}}{\binom{m}{n}}$$

□

Exercise 1.3. *Show that, as $M \to \infty$, we can write*

$$\mathbf{P}(X_1 = x_1, \cdots, X_N = x_N) \to \int_0^1 \theta^s (1-\theta)^{N-s} dF_M(\theta)$$

Proof. We have $P(\sum_{i=1}^N X_i = s) = \binom{N}{s} P(X_1, \cdots, X_N)$, so $P(X_1, \cdots, X_N) = \int_0^1 \frac{(M\theta)_s (M(1-\theta))_{n-s}}{(M)_N} dF_M(\theta)$.

$$\begin{aligned}
\lim_{M\to\infty} \frac{(M\theta)_s (M(1-\theta))_{N-s}}{(M)_N} &= \lim_{M\to\infty} \frac{(M\theta)\cdots(M\theta - s + 1)M(1-\theta)\cdots(M(1-\theta)-(N-s)+1)}{M\cdots(M-N+1)} \\
&= \lim_{M\to\infty} \frac{M\theta}{M} \cdots \lim_{M\to\infty} \frac{M\theta - s + 1}{M - s + 1} \lim_{M\to\infty} \frac{M(1-\theta)}{M-s} \cdots \lim_{M\to\infty} \frac{M(1-\theta)-(N-s)+1}{M-N+1} \\
&= \theta^s (1-\theta)^{N-s}
\end{aligned}$$

□

Exercise 1.4. *The Poisson random variable has PDF*

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Re-write the density of the Poisson random variable in exponential family form. What are η, $T(x)$, $A(\eta)$ and $h(x)$? What about if we have n independent samples $x_1, \ldots, x_n$?

Proof.

$$P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!}e^{-\lambda}e^{ln\lambda^x} = \frac{1}{x!}e^{xln(\lambda)-\lambda}$$

We have $h(x) = 1/x!, \eta(\lambda) = ln\lambda, T(x) = x, A(\eta) = e^{\eta}$.
For n independent x_i, we have:

$$P(\vec{x}|\lambda) = \prod_i P(x_i|\lambda) = \frac{1}{\prod_i x_i!}e^{ln(\lambda)(\sum_i x_i)-n\lambda}$$

Thus, $h(\vec{x}) = \frac{1}{\prod_i x_i!}, T(\vec{x}) = \sum x_i, \eta(\lambda) = ln(\lambda), A(\eta) = ne^{\eta}$. □

Exercise 1.5. *The gamma random variable has PDF*

$$p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$$

What are the natural parameters and sufficient statistics for the gamma distribution, given n observations $x_1, \ldots, x_N$?

Proof.

$$\begin{aligned} L(\mathbf{x}|\alpha,\beta) &= (\frac{\beta^\alpha}{\Gamma(\alpha)})^n \prod_{i=1}^{n} x_i^{\alpha-1}e^{-\beta x_i} \\ &= exp\left((\alpha-1)\sum_i ln(x_i) \;\; - \beta\sum_i x_i \;\; + n\ ln\Big(\frac{\beta^\alpha}{\Gamma(\alpha)}\Big)\right) \end{aligned}$$

Thus, $T(x) = [\sum ln(x_i), \sum x_i], \eta(\theta) = [\alpha - 1, -\beta]$ □

Exercise 1.6. *For exponential family random variables, we know that the sufficient statistic $T(X)$ contains all the information about X, so (for univariate X) we can write the moment generating function of the sufficient statistic as $\mathbb{E}[e^{sT(x)}|\eta]$. Show that the moment generating function for the sufficient statistic of an arbitrary exponential family random variable with natural parameter η can be written as*

$$M_{T(X)}(s) = \exp A(\eta + s) - A(\eta)$$

Proof.

$$\begin{aligned} M_{T(x)}(s) = E(e^{T(x)s}) &= \int e^{T(x)s}h(x)e^{\eta T(x)-A(\eta)}dx \\ &= \int h(x)e^{T(x)(\eta+s)-A(\eta)}dx \\ &= \int h(x)e^{T(x)(\eta+s)-A(\eta+s)+A(\eta+s)-A(\eta)}dx \\ &= exp(A(\eta+s) - A(\eta)). \end{aligned}$$

□

Exercise 1.7. *It is usually easier to calculate mean and variance using the cumulant generating function rather than the moment generating function. Starting from the exponential family representation of the Poisson distribution from Exercise 1.4, calculate the mean and variance of the Poisson using a) the moment generating function, and b) the cumulant generating function.*

Proof. a) moment generating function: $M_x(s) = exp(e^{\eta+s} - e^{\eta})$
Take first derivative at s=0 gives: $m_1 = \frac{dM_x(s)}{ds} = exp(e^{\eta} + s - e^{\eta})exp(\eta + s)|_{s=0} = e^{\eta} = \lambda$
Take second derivative at s=0 gives: $m_2 = \frac{d^2M_x(x)}{ds^2} = e^{2\eta} + e^{\eta} = \lambda^2 + \lambda$
$\mu = m_1 = \lambda, \sigma^2 = m_2 - m_1^2 = \lambda$

b) cumulant generating function: $C_x(x) = e^{\eta+s} - e^{\eta}$
Take first derivative at s=0 gives $c_1 = e^{\eta} = \lambda$
Take second derivative at s=0 gives $c_2 = e^{\eta} = \lambda$
It is easier to compute mean and variance with cumulant generating function. □

Exercise 1.8. *Suppose we have N independent observations $x_1, \ldots, x_N \overset{iid}{\sim} Normal(\mu, \sigma^2)$. If σ^2 is known and $\mu \sim Normal(\mu_0, \sigma_0^2)$, derive the posterior for $\mu|x_1, \ldots, x_N$*

Proof.

$$\begin{aligned}
f(\mu|x_1, \cdots, x_N) &= \frac{L(\mathbf{x}|\mu)f(\mu)}{f(x_1, \cdots, x_N)} \\
&\propto e^{-[\frac{1}{2\sigma^2}\sum_i(\mu-x_i)^2+\frac{1}{2\sigma_0^2}(\mu-\mu_0^2)]} \\
&\propto e^{-[\frac{1}{2\sigma^2\sigma_0^2}(\sum_i\sigma_0^2(\mu-x_i)^2+\sigma^2(\mu-\mu_0)^2)]} \\
&\propto e^{-[\frac{1}{2\sigma^2\sigma_0^2}(n\sigma_0^2\mu^2-2\mu(\sum x_i)\sigma_0^2+\sigma^2\mu^2-2\sigma^2\mu\mu_0)]} \\
&\propto e^{-[\frac{1}{2\sigma^2\sigma_0^2}((n\sigma_0^2+\sigma^2)\mu^2-2\mu(\sigma_0^2\sum x_i+\sigma^2\mu_0))]} \\
&\propto e^{-[\frac{1}{2\sigma^2\sigma_0^2}(n\sigma_0^2+\sigma^2)(\mu-\frac{\sigma_0^2\sum_i x_i+\sigma^2\mu_0}{n\sigma_0^2+\sigma^2})^2]}
\end{aligned}$$

The above equation has distribution $N(\frac{\sigma_0^2\sum_i x_i+\sigma^2\mu_0}{n\sigma_0^2+\sigma^2}, \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2})$. □

Exercise 1.9. *Now, let's assume $x_1, \ldots, x_N \overset{iid}{\sim} Normal(\mu, \sigma^2)$ with known mean μ but unknown variance σ^2. Let's express the likelihood in terms of the precision, $\omega = \frac{1}{\sigma^2}$:*

$$f(x_i|\mu, \omega) = \sqrt{\frac{\omega}{2\pi}} \exp\left\{-\frac{\omega}{2}(x_i - \mu)^2\right\}$$

Let ω have a gamma prior (this is also known as putting an inverse-gamma prior on σ^2):

$$p(\omega) = \frac{\beta^\alpha}{\Gamma(\alpha)}\omega^{\alpha-1}e^{-\beta\omega}$$

Derive the posterior distribution for ω

Proof.

$$\begin{aligned}
f(w|x_1, \cdots, x_N) &= \frac{L(\mathbf{x}|w)f(w)}{f(x_1, \cdots, x_N)} \\
&\propto f(w)\prod_i f(x_i|w) \\
&\propto w^{\alpha-1}e^{-\beta w}\prod_i \left(w^{\frac{1}{2}}e^{-\frac{w}{2}(x_i-\mu)^2}\right) \\
&\propto w^{n/2+\alpha-1} exp\left(-w\Big(\frac{\sum_i(x_i-\mu)^2}{2}+\beta\Big)\right)
\end{aligned}$$

The above equation has distribution $Gamma\left(n/2+\alpha, \frac{\sum_i(x_i-\mu)^2}{2}+\beta\right)$ □

Exercise 1.10. *Let's assume $x \sim Normal(0,\sigma^2)$ and that $\sigma^2 \sim InvGamma(\alpha,\beta)$ (i.e. $1/\sigma^2 \sim Gamma(\alpha,\beta)$). Show that the marginal distribution of x is given by a Student's t distribution.*

Proof.

$$\begin{aligned} f(x,w) &= f(x|w)f(w) \\ &= (2\pi)^{-1/2}\frac{\beta^\alpha}{\Gamma(\alpha)}\omega^{\frac{1}{2}+\alpha-1}e^{-\omega(\beta+\frac{1}{2}x^2)} \\ &\sim NormalGamma(0,1,\alpha,\beta) \end{aligned}$$

Thus,

$$\begin{aligned} f(x) = \int_0^\infty f(x,w)dw &= c\int_0^\infty Gamma(1/2+\alpha,\beta+\frac{1}{2}x^2)dw \\ = c = \frac{\Gamma(\alpha+\frac{1}{2})}{\sqrt{2\pi\beta}\Gamma(\alpha)}(1+\frac{x^2}{2\beta})^{-\frac{2\alpha+1}{2}} &= \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi\sigma^2}\Gamma(\frac{v}{2})}\left(1+\frac{1}{v}\left(\frac{x}{\sigma}\right)^2\right)^{-\frac{v+1}{2}} \end{aligned}$$

where $v=2\alpha, \sigma^2=\frac{\beta}{\alpha}$. This is a non-standardized Student's t-distribution. □

Exercise 1.11. *The covariance matrix Σ of a vector-valued random variable x is the matrix whose entries $\Sigma(i,j)=cov(x_i,x_j)$ are given by the covariance between the ith and jth elements of x, giving*

$$\Sigma = \mathbb{E}\left[(x-\mu)(x-\mu)^T\right]$$

Show that a) $\Sigma = E[xx^T]-\mu\mu^T$; b) if the covariance of x is σ, then the covariance of $Ax+b$ is $A\Sigma A^T$

Proof.
a) $\mathbb{E}\left[(x-\mu)(x-\mu)^T\right] = \mathbb{E}\left[xx'-x\mu'-\mu x'+\mu\mu'\right] = \mathbb{E}[xx']-\mu\mu'$
b)

$$\begin{aligned} cov(Ax+b) = cov(Ax) &= \mathbb{E}\big[(Ax-A\mu)(Ax-A\mu)'\big] \\ &= \mathbb{E}[Ax(Ax)'] - A\mu(A\mu)' \\ &= A(\mathbb{E}[xx']-\mu\mu')A' = A\Sigma A' \end{aligned}$$

□

Exercise 1.12 (Standard multivariate normal)**.** *The simplest multivariate normal, known as the standard multivariate normal, occurs where the entries of x are independent and have mean 0 and variance 1. a) What is the moment generating function of a univariate normal, with mean m and variance v^2? b) Express the PDF and moment generating function of the standard multivariate normal, in vector notation.*

Proof.
a)

$$\begin{aligned} M_X(t) &= \int e^{tx}\frac{1}{\sqrt{2\pi v^2}}e^{-\frac{(x-m)^2}{2v^2}}dx \\ &= \frac{1}{\sqrt{2\pi v^2}}\int e^{-\frac{1}{2v^2}(x^2-2mx-2tv^2x+m^2)}dx \\ &= \frac{1}{\sqrt{2\pi v^2}}\int e^{-\frac{1}{2v^2}[(x-(m+v^2t))^2+m^2-(m+v^2t)^2]}dx \\ &= e^{-\frac{1}{2v^2}(m^2-(m+v^2t)^2)} \\ &= e^{mt+\frac{v^2t^2}{2}}. \end{aligned}$$

b)

$$f(x) = (2\pi)^{-n/2} e^{-\frac{1}{2}x'x}$$

$$M_X(t) = \int e^{t'x} f(x) dx = (2\pi)^{-n/2} \int e^{-\frac{1}{2}(x'x - 2t'x + t't - t't)} dx = e^{\frac{1}{2}t't}$$

□

Exercise 1.13 (Multivariate normal)**.** *A random vector x has multivariate normal distribution if and only if every linear combination of its elements is univariate normal, i.e. if the scalar value $z = a^T x$ is normally distributed for all possible x. Prove that this implies that x is multivariate normal if and only if its moment generating function takes the form $M_X(s) = \exp\{s^T \mu + s^T \Sigma s\}$, where μ and Σ are the mean and covariance of x. Hint: We know the moment generating function of z in terms of the mean and variance of z, from the previous question...*

Proof.
$\Rightarrow$) x is multivariate normal $\Rightarrow z = a'x$ is univariate normal $\forall x \in \mathbb{R}^n$
$\Rightarrow M_x(s) = E(e^{s'x}) = E(e^z) = e^{\mu_z + 0.5\sigma_z^2} = e^{s'\mu + 0.5s'\Sigma s}$

$\Leftarrow) M_x(s) = E(e^{s'x}) = e^{s'\mu + 0.5s'\Sigma s} = e^{E(s'x) + 0.5var(s'x)} = M_{s'x}(1)$
$\Rightarrow z = s'x$ is univariate normal $\Rightarrow$ x is multivariate normal. □

Exercise 1.14 (Relationship to standard multivariate normal)**.** *An equivalent statement is that a random vector x has multivariate normal distribution if and only if it can be written in the form*

$$x = Dz + \mu$$

for some matrix D, real-valued vector μ, and vector z distributed according to a standard multivariate normal. Express the moment generating function of x in terms of D, and uncover the relationship between D and Σ. Use this result to suggest a method for generating multivariate normal random variables, if you have a method for generating Normal(0,1) univariate random variables.

Proof.

$$M_x(s) = E(e^{s'(Dz+\mu)}) = e^{s'\mu} E(e^{s'DZ}) = e^{s'\mu + \frac{1}{2}s'DD's} = e^{s'\mu + \frac{1}{2}s'\Sigma s}$$

Thus, we have $\Sigma = DD'$. Generate $x = \mu + Dz$. □

Exercise 1.15. *Use the result from the previous question to show that the PDF of a multivariate normal random vector $x \sim$ Normal(μ, Σ) takes the form*

$$p(x) = \frac{1}{(2\pi)^{n/2}} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\},$$

by using a change-of-variables from the standard multivariate normal distribution.

Proof. Apply change of variable in density function: $f_X(x) = f_Z(g^{-1}(x))|J(x)|$, where $x = g(z) = \mu + Dz$, $z = g^{-1}(x) = D^{-1}(x-\mu)$, and $|J(x)| = |D^{-1}|$ is the determinant of the Jacobian matrix for $g^{-1}(x)$. Thus:

$$f_Z(z) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}z'z} \Rightarrow f_X(x) = (2\pi)^{-\frac{n}{2}} |D^{-1}| e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

□

Exercise 1.16 (marginal distribution)**.** *Let us assume that $x \sim$ Normal(μ, Σ), and let us partition x into 2 components x_1 and x_2. Let us similarly partition μ and σ so that*

$$\mu = (\mu_1, \mu_2)^T \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$$

Derive the marginal distribution of x_1.

Proof. Let $U \equiv [I, O]$, where I is the identity matrix and O is the zero matrix. Then $x_1 = Ux$ is normal because linear combination of all components of x is normal. Moreover, $E(x1) = E(Ux) = \mu_1$, and $Cov(X_1) = Cov(Ux) = UCov(x)U' = U\Sigma U' = \Sigma_{11}$. □

Exercise 1.17 (Precision matrix)**.** *Earlier, we chose to express a univariate normal random variable in terms of its precision, to make math easier. We can also express a multivariate normal in terms of a precision matrix $\Omega = \Sigma^{-1}$. Partition Ω as*

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix}$$

and express Ω_{11}, Ω_{12} and Ω_{22} in terms of Σ_{11}, Σ_{12} and Σ_{22}. Hint: You'll need the matrix inversion lemma

Proof. Directly apply matrix inversion lemma, we get:

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} K_1^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}K_2^{-1} \\ -K_2^{-1}\Sigma_{12}'\Sigma_{11}^{-1} & K_2^{-1} \end{pmatrix}$$

where $K_1 = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}'$ and $K_2 = \Sigma_{22} - \Sigma_{12}'\Sigma_{11}^{-1}\Sigma_{12}$. □

Exercise 1.18 (Conditional distribution)**.** *The conditional distribution of $x_1|x_2$ is also normal, with mean $\mu_1 + \Sigma_{12}\sigma_{22}^{-1}(x_2 - \mu_2)$ and covariance $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$. Prove this for the case where μ is zero (the general case isn't really harder, just more tedious). Hint: ignore any constants that don't involve x_1. You might want to work with the log conditional density.*

Proof.
Define $z \equiv x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2$, then $Cov(z, x_2) = \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0$. Two normal random variable are uncorrelated iff they are independent:

$$\Rightarrow f(z|x_2) = f(z) \Rightarrow E(z|x_2) = E(z) = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \Rightarrow Cov(z|x_2) = Cov(z) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}'$$

$$\Rightarrow f(x_1|x_2) = f(z + \Sigma_{12}\Sigma_{22}^{-1}x_2|x_2) \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}')$$

□

Exercise 1.19 (method of moments)**.** *To obtain the theoretical moments, we can assume that $E[y_i|x_i] = x_i^T\beta$, implying that the covariance between the predictors x_i and the residuals is zero. By setting the sample covariance between the x_i and the ϵ_i to zero, derive a method of moments estimator $\hat{\beta}_{MM}$*

Proof.

$$cov(x, \epsilon) = 0 \Rightarrow E(x'\epsilon) - E(x')E(\epsilon) = 0 \Rightarrow E[x'(y - xb)] = 0 \Rightarrow E(x'y) = E(x'xb) \Rightarrow \hat{b} = (X'X)^{-1}(X'Y)$$

□

Exercise 1.20 (maximum likelihood)**.** *Show that, if we assume $\epsilon_i \sim Normal(0, \sigma^2)$, then the ML estimator $\hat{\beta}_{ML}$ is equivalent to the method of moments estimator.*

Proof.

$$L = c_1 exp(-\frac{1}{2\sigma^2}(Y - XB)'(Y - XB) \Rightarrow ln(L) = c_2 - \frac{1}{2\sigma^2}(Y - XB)'(Y - XB)$$

Maximize $ln(L)$, we obtain $-2X'(Y - XB) = 0 \Rightarrow \hat{B} = (X'X)^{-1}X'Y$. □

Exercise 1.21 (Least squares loss function)**.** *Show that if we assume a quadratic loss function, i.e.* $\hat{\beta}_{LS} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$*, we recover the same estimator again.*

Proof. $\hat{B}_{LS}$ tries to minimize $(Y - XB)'(Y - XB)$, which is the same optimization problem as the previous exercise in maximum likelihood. □

Exercise 1.22 (Ridge regression)**.** *We may wish to add a regularization term to our loss term. For example, ridge regression involves adding an L2 penalty term, so that*

$$\hat{\beta}_{ridge} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \ s.t. \ \sum_{j=1}^{p} \beta_j^2 \leq t$$

for some $t \geq 1$*.*

Reformulate this constrained optimization using a Lagrange multiplier, and solve to give an expression for $\hat{\beta}_{ridge}$*. Comparing this with the least squares estimator, comment on why this estimator might be prefered in practice.*

Proof. The Lagrange set up with Kuhn-Tucker Conditions are as follows

$$\begin{aligned} \mathcal{L} = &-(Y - XB)'(Y - XB) + \lambda(t - B'B) \\ &\lambda \geq 0 \\ &\lambda(t - B'B) = 0 \\ &t - B'B \geq 0 \end{aligned}$$

Take derivative, $\frac{\partial L}{\partial \beta} = -2X'(Y - XB) + 2\lambda B = 0 \Rightarrow \hat{B_{ridge}} = (X'X - \lambda I)^{-1} X'Y$.
Decomposition of MSE shows that $MSE = (E[\hat{\theta}] - \theta])^2 + var(\hat{\theta})$. Sometimes ridge regression may be preferred because the estimator can have less variance. □

Exercise 1.23. *What is the sampling distribution for* $\hat{\beta}_{LS}$ *(*$= \hat{\beta}_{MM} = \hat{\beta}_{ML}$*)?*

Proof.

$$\begin{aligned} \hat{B} &= (X'X)^{-1} X'(XB + \epsilon) \\ &= (X'X)^{-1}(X'X)B + (X'X)^{-1} X'\epsilon \\ &\sim N(B, (X'X)^{-1}\sigma^2) \end{aligned}$$

□

Exercise 1.24. *How about the sampling distribution for* $\hat{\beta}_{ridge}$*?*

Proof.

$$\begin{aligned} \hat{B}_{ridge} &= (X'X - \lambda I)^{-1} X'(XB + \epsilon) \\ &= (X'X - \lambda I)^{-1}(X'X)B + (X'X - \lambda I)^{-1} X'\epsilon \\ &\sim N((X'X - \lambda I)^{-1}(X'X)B,\ \sigma^2 (X'X - \lambda I)^{-1}(X'X)(X'X - \lambda I)^{-1}) \end{aligned}$$

□

Exercise 1.25. *The two exercises above assumed the residual variance* σ^2 *is known. This is unlikely to be the case. Propose a strategy for estimating the standard error of* $\hat{\beta}_{LS}$ *from data, when* σ^2 *is unknown. Implement it in R, and test it on the dataset* `Prestige` *in the R package* `cars` *(there's a starter script,* `prestige.R` *on Github). Do you get the same standard errors as the built-in function* `lm`*?*

Proof.

$$SSE = (y-\hat{y})'(y-\hat{y}) = y'(I-M)'(I-M)y = y'(I-M)y$$

$$\begin{aligned} E(SSE) &= E(y'(I-M)y) = tr(\sigma^2(I-M)) + E(y)'(I-M)E(y) \\ &= tr(\sigma^2(I-M) + B'X(I-M)XB \\ &= \sigma^2 tr(I-M) = \sigma^2(n-p) \end{aligned}$$

So σ^2 can be estimated by MSE, which is $\frac{SSE}{n-p}$. The standard errors of $\hat{B}_{LS}$ match the standard errors from built-in lm function completely. □

Exercise 1.26. *Let's assume we care about $f(\theta) = \sum_i \theta_i$. What is the standard error of $f(\theta)$?*

Proof.

$$Var(\sum_i \theta_i) = \sum_i Var(\theta_i) + 2\sum_{i>j} Cov(\theta_i, \theta_j)$$

Thus, $Var(\sum_i \theta_i)$ equals sum of all elements of covariance matrix of θ. □

Exercise 1.27. *How about the standard error of some arbitrary non-linear function $f(\theta)$? Hint: Try a Taylor expansion*

Proof. Expand the Taylor series around a:

$$\begin{aligned} & f(\theta) = f(a) + f'(a)(\theta - a) + ... \\ & Var(f(\theta)) = Var(f(a) + f'(a)(\theta - a) + ...) \approx Var(f'(a)(\theta - a)) \end{aligned}$$

Let $a \to \theta, Var(f'(a)(\theta - a)) \approx f'(a)Cov(\theta)f'(a)^t$

□