SDS 383D: Modeling II **Spring 2017**

Section 1 (Solutions by Xinjie Fan)

1.1 Exchangeability and de Finetti's theorem

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: The sequence is obviously not iid. We only show the sequence is exchangeable in the following. Consider a sequence of N colors, $(X_1, ..., X_N)$,

$$\mathbf{P}(X_1, ..., X_N) = \mathbf{P}(X_1)\mathbf{P}(X_2|X_1)\cdots\mathbf{P}(X_N|X_1, ..., X_{N-1}),$$

where

$$\mathbf{P}(X_i|X_1, ..., X_{i-1}) = \begin{cases} \frac{r+j-1}{r+b+i-1} & X_i \text{ is the } j\text{th red ball in the sequence} \\ \frac{b+j-1}{r+b+i-1} & X_i \text{ is the } j\text{th blue ball in the sequence.} \end{cases}$$

Notice that the dominator does not depend on $X_1, ..., X_N$. Therefore, neither does the dominator of the joint distribution. We only need to show that the numerator of the joint distribution does not depend on the order of $X_1, ..., X_N$. If we consider the product of numerators corresponding to red balls first, it would be $r(r+1)\cdots(r+m-1)$, where m is the number of red balls in $X_1, ..., X_N$. Similarly, the product of numerators corresponding to blue balls would be $b(b+1)\cdots(b+n-1)$, where n is the number of blue balls in $X_1, ..., X_N$. Then the numerator of the joint distribution is $r(r+1)\cdots(r+m-1)b(b+1)\cdots(b+n-1)$ which does not depend on the order of $X_1, ..., X_N$. Therefore, they are exchangeable.

■

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: Our assumption is that $(X_1, ..., X_M)$ is exchangeable.
Claim:

$$\mathbf{P}\left(X_1, .., X_M | \sum_{i=1}^{M} X_i = t\right) = \begin{cases} 1/\binom{M}{t} & \sum_{i=1}^{M} X_i = t \\ 0 & \sum_{i=1}^{M} X_i \neq t \end{cases}$$

The reason is condition on $\sum_{i=1}^{M} X_i = t$, there are $\binom{M}{t}$ different sequences, and one sequence can be transformed to another by rearranging the order. Therefore, they share the same probability (exchangeable) and sum up to one.

Therefore, we have the following(the second equation is a simple combinatorial step, counting how many different sequences has s ones in the first N numbers; the third equation is easy to check by expanding both sides):

$$\mathbf{P}\left(\sum_{i=1}^{N} X_i = s \Big| \sum_{i=1}^{M} X_i = t\right) = \sum_{\sum_{i=1}^{N} X_i = s, \sum_{i=1}^{M} X_i = t} \mathbf{P}\left(X_1, .., X_M | \sum_{i=1}^{M} X_i = t\right)$$
$$= \frac{\binom{N}{s}\binom{M-N}{t-s}}{\binom{M}{t}} = \frac{\binom{t}{s}\binom{M-t}{N-s}}{\binom{M}{N}}$$

■

We can therefore write

$$\mathbf{P}\left(\sum_{i=1}^{N} X_i = s\right) = \binom{N}{s} \sum_{t=s}^{M-N+s} \frac{(t)_s (M-t)_{n-s}}{(M)_N} \mathbf{P}\left(\sum_{i=1}^{M} X_i = t\right), \tag{1.1}$$

where $(x)_y = x(x-1)\dots(x-y+1)$.

Let $F_M(\theta)$ be the distribution function of $\frac{1}{M}(X_1, + \dots, +X_M)$ – i.e. a step function between 0 and 1, with steps of size $\mathbf{P}(\sum_i X_i = t)$ at $t = 0, 1, \dots, M$. Then we can rewrite Equation 1.1 as

$$\mathbf{P}\left(\sum_{i=1}^{N} X_i = s\right) = \binom{N}{s} \int_0^1 \frac{(M\theta)_s (M(1-\theta))_{N-s}}{(M)_N} dF_M(\theta)$$

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

$$\mathbf{P}\left(\sum_{i=1}^{N} X_i = s\right) \to \binom{N}{s} \int_0^1 \theta^s (1-\theta)^{N-s} dF_M(\theta)$$

Proof: Loosely speaking, as $M \to \infty$, $(M\theta)_s \to (M\theta)^s$, and similarly $(M(1-\theta))_{N-s} \to (M(1-\theta))^{N-s}$, $(M)_N \to M^N$. Therefore, $\frac{(M\theta)_s (M(1-\theta))_{N-s}}{(M)_N} \to \theta^s (1-\theta)^{N-s}$, ■

The proof is completed using a result (the Helly Theorem), that shows that any sequence $\{F_M(\theta); M = 1, 2, \dot{\}}$ of probability distributions on [0,1] contains a subsequence that converges to $F(\theta)$.

1.2 The exponential family of distributions

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 $x_1, \dots, x_n$?

Proof:

$$\mathbf{P}(x|\lambda) = \frac{1}{x!} \exp(\log(\lambda)x - \lambda);$$
$$\mathbf{P}(x_1, ..., x_n|\lambda) = \frac{1}{\prod_{i=1}^{n} x_i!} \exp(\log(\lambda)(\sum_{i=1}^{n} x_i) - n\lambda).$$

■

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:

$$p(x|\alpha,\beta) = \exp\left((-\beta, \alpha-1) \begin{pmatrix} x \\ \log x \end{pmatrix} - (\log(\Gamma(\alpha)) - \alpha \log \beta) \right).$$

$$p(x_1, ..., x_n|\alpha,\beta) = \exp\left((-\beta, \alpha-1) \begin{pmatrix} \sum x_i \\ \sum \log x_i \end{pmatrix} - n(\log(\Gamma(\alpha)) - \alpha \log \beta) \right).$$

The natural parameters are $(-\beta, \alpha-1)$ and sufficient statistics are $(\sum x_i, \sum \log x_i)$. ■

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:

$$\begin{aligned} M_{T(X)}(s) = E[\exp(s^T T(X))] &= \int \exp\{s^T T(X)\} h(X) \exp\{\eta^T T(X) - A(\eta)\} dX \\ &= \int h(X) \exp\{(s+\eta)^T T(X) - A(\eta+s)\} dX \exp\{A(\eta+s) - A(\eta)\} \\ &= \exp\{A(\eta+s) - A(\eta)\} \end{aligned} \tag{1.2}$$

■

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:

$$E[e^{sx}] = e^{-\lambda} \int_0^\infty \frac{e^{sx}\lambda^x}{x!} dx = e^{-\lambda} \int_0^\infty \frac{(e^s\lambda)^x}{x!} dx = \frac{e^{e^s\lambda}}{e^\lambda}$$

(a) Moment generating function:

$$\frac{dE[e^{sX}]}{ds} = \lambda e^{\lambda e^s - \lambda + s}; \quad \frac{d^2E[e^{sX}]}{ds^2} = \lambda e^{\lambda e^s - \lambda + s}(\lambda e^s + 1).$$

Therefore, $E[X] = \frac{dE[e^{sX}]}{ds}\Big|_{s=0} = \lambda$, $E[X^2] = \frac{d^2E[e^{sX}]}{ds^2}\Big|_{s=0} = \lambda(\lambda+1)$, and $var(X) = \lambda(\lambda+1) - \lambda^2 = \lambda$.
(b) Cumulant generating function:

$$C_X(s) = (e^s - 1)\lambda; \quad \frac{dC_X(s)}{ds} = \lambda e^s; \quad \frac{d^2C_X(s)}{ds^2} = \lambda e^s.$$

Therefore, $E[X] = \frac{dC_X(s)}{ds}\Big|_{s=0} = \lambda$, $var(X) = \frac{d^2C_X(s)}{ds^2}\Big|_{s=0} = \lambda$. ■

Exercise 1.8 *Suppose we have N independent observations $x_1, \ldots, x_N \overset{iid}{\sim}$ Normal(μ, σ^2). If σ^2 is known and $\mu \sim$ Normal(μ_0, σ_0^2), derive the posterior for $\mu | x_1, \ldots, x_N$*

Proof:

$$\begin{aligned} p(\mu|x_1,...,x_N) &\propto \exp\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\}\prod_{i=1}^N \exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}\} \\ &\propto \exp\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\}\exp\{-\frac{n\mu^2-2\sum_{i=1}^N x_i\mu}{2\sigma^2}\} \\ &\propto \exp\{-\frac{(n/\sigma^2+1/\sigma_0^2)\mu^2-2(\mu_0/\sigma_0^2+\sum x_i/\sigma^2))\mu}{2}\}. \end{aligned} \tag{1.3}$$

Therefore, $\mu|x_1,...,x_N \sim N(\frac{\mu_0/\sigma_0^2+\sum x_i/\sigma^2}{n/\sigma^2+1/\sigma_0^2}, \frac{1}{n/\sigma^2+1/\sigma_0^2})$ ■

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

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

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:

$$\begin{aligned} p(\omega|x_1,...,x_N) &\propto \omega^{\alpha-1}e^{-\beta\omega}\prod_{i=1}^N \omega^{1/2}\exp\{-\frac{\omega(x_i-\mu)^2}{2}\} \\ &\propto \omega^{n/2+\alpha-1}\exp\{-(\beta+\sum(x_i-\mu)^2/2)\omega\}. \end{aligned} \tag{1.4}$$

Therefore, $\omega|x_1,...,x_N \sim Gamma(n/2+\alpha, \beta+\sum(x_i-\mu)^2/2)$. ■

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

Proof:

$$\begin{aligned} f(x) &= \int f(x|\sigma^2)f(\sigma^2)d\sigma^2 \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)\sqrt{2\pi}}\int_0^\infty (\frac{1}{\sigma^2})^{\alpha+3/2}e^{-(\frac{x^2/2+\beta}{\sigma^2})}d\sigma^2 \\ &= \frac{\beta^\alpha\Gamma(\alpha+1/2)}{\Gamma(\alpha)\sqrt{2\pi}(x^2/2+\beta)^{\alpha+1/2}} \end{aligned} \tag{1.5}$$

Therefore, $x \sim t(\nu = 2\alpha, \sigma = \sqrt{\beta/\alpha})$. ■

1.3 Multivariate normal distribution

Exercise 1.11 (covariance matrix) *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) $\Sigma = \mathbb{E}[xx^T] - \mu\mu^T$; b) if the covariance of x is σ, then the covariance of $Ax + b$ is $A\Sigma A^T$

Proof: (a)

$$\Sigma = \mathbb{E}\left[(x-\mu)(x-\mu)^T\right] = \mathbb{E}\left[xx^T - 2x\mu^T + \mu\mu^T\right] = \mathbb{E}\left[xx^T\right] - 2\mathbb{E}\left[x\right]\mu^T + \mu\mu^T = \mathbb{E}[xx^T] - \mu\mu^T$$

(b)

$$\begin{aligned} Cov(Ax+b) &= \mathbb{E}\left[(Ax+b-\mathbb{E}(Ax+b))(Ax+b-\mathbb{E}(Ax+b))^T\right] \\ &= \mathbb{E}\left[(Ax-\mathbb{E}(Ax))(Ax-\mathbb{E}(Ax))^T\right] = Cov(Ax) \\ &= \mathbb{E}[(Ax)(Ax)^T] - \mathbb{E}[Ax]\mathbb{E}[Ax]^T \\ &= A\{\mathbb{E}[xx^T] - \mathbb{E}[x]\mathbb{E}[x]^T\}A^T \\ &= ACov(x)A^T \end{aligned} \tag{1.6}$$

■

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)Suppose $X \sim \mathcal{N}(m, v^2)$, then

$$\begin{aligned} \mathbb{E}[e^{tX}] &= \int e^{tx} \frac{1}{\sqrt{2\pi v^2}} e^{-\frac{(x-m)^2}{2v^2}} dx \\ &= e^{\frac{(m+v^2t)^2-m^2}{2v^2}} \frac{1}{\sqrt{2\pi v^2}} \int e^{-\frac{(x-(m+v^2t))^2}{2v^2}} dx \\ &= e^{\frac{(m+v^2t)^2-m^2}{2v^2}} = e^{\frac{2mt+v^2t^2}{2}} \end{aligned} \tag{1.7}$$

(b) Suppose $X \sim \mathcal{N}(0, I)$, then

$$p(x) = \frac{1}{|det(2\pi I)|^{\frac{1}{2}}} \exp\{-\frac{X^TX}{2}\}.$$

The moment generating function can be calculated as follows:

$$\begin{aligned} \mathbb{E}[e^{tX}] = \exp\{\frac{t^Tt}{2}\} \frac{1}{|det(2\pi I)|^{\frac{1}{2}}} \int \exp\{-\frac{(X-t)^T(X-t)}{2}\} dX \\ = \exp\{\frac{t^Tt}{2}\} \end{aligned} \tag{1.8}$$

■

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 $z = a^Tx$ 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 $M_X(s) = \exp\{s^T\mu + \frac{1}{2}s^T\Sigma s\}$. Then, for any aany linear, 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: (a) First, we assume that X is a multivariate normal. Then s^TX is distributed as $\mathcal{N}(s^T\mu, s^T\Sigma s)$, so using equation 1.7 with $t = 1$, we have $M_X(s) = \mathbb{E}(e^{s^TX}) = e^{\frac{2s^T\mu+s^T\Sigma s}{2}}$.
(b) Now, we assume that the moment generating function takes the form $M_X(s) = \exp\{s^T\mu + \frac{1}{2}s^T\Sigma s\}$. Then, for any a, consider the moment generating function of a^TX:

$$E[e^{sa^TX}] = \exp\{sa^T\mu + sa^T\Sigma as\} = \exp\{sa^T\mu + a^T\Sigma as^2\}.$$

Therefore, a^TX is normally distributed for any a.

■

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:

$$\mathbb{E}[e^{s^Tx}] = \mathbb{E}[e^{s^TDx+s^T\mu}] = e^{s^T\mu}E[e^{s^TDz}] = e^{s^T\mu}M_z(s^TD).$$

Since $\Sigma = Cov(X) = Cov(Dz + \mu) = DD^T$, $\Sigma = DD^T$. To generate an arbitrary normal distribution, we can first generate samples standard normal distribution and then transform the samples with the linear transformation. ■

Exercise 1.15 *Use the result from the previous question to show that the PDF of a multivarite 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:

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

Since $DD' = \Sigma$, $det(D)$ is equal to $\sqrt{det(\Sigma)}$, which completes the proof. ■

1.3.1 Manipulation of multivariate normals

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*

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

Derive the marginal distribution of x_1.

Proof: Note that $x_1 = (I, 0)x$, so x_1 is a linear transformation of x and therefore normally distributed.

$\mathbb{E}[x_1] = (I, 0)\mu = \mu_1$, $Cov(x_1) = (I, 0)\Sigma(I, 0)^T = \Sigma_{11}$. Therefore, $x_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$. ■

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: With block matrix inversion formula, we have

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

■

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: Consider the random vector $(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2, x_2)$. Since it is a linear transformation of a normal random vector, so itself is normally distributed as well. Observe that $Cov(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2, x_2) = Cov(x_1, x_2) - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0$. The zero correlation implies independence within normally distributed random vectors.

Therefore, $x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2|x_2$ has the same distribution as $x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2$, namely $x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2|x_2 \sim \mathcal{N}(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$. Therefore, $x_1|x_2 \sim \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)$

■

1.4 Frequentist estimation and uncertainty quantification

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:

$$\begin{aligned} Cov(y-\beta^T x, x) &= \mathbb{E}[(y-\beta^T x)x^T] - \mathbb{E}[y-\beta^T x]\mathbb{E}[x^T] \\ &= \frac{\sum_i y_i x_i^T - \beta^T x_i x_i^T}{N} - \frac{\sum_i y_i - \beta^T x_i}{N}\frac{\sum_i x_i^T}{N} \\ &= \frac{\sum_i y_i x_i}{N} - \bar{y}\bar{x} - \beta^T(\frac{\sum x_i x_i^T}{N} - \bar{x}\bar{x}^T), \end{aligned} \tag{1.9}$$

where $\bar{y} = \sum_i y_i/N$, and $\bar{x} = \sum_i x_i/N$. Set $Cov(y-\beta^T x, x)$ to zero, we have $\hat{\beta}^T_{MM} = (\frac{\sum_i y_i x_i}{N} - \bar{y}\bar{x})(\frac{\sum x_i x_i^T}{N} - \bar{x}\bar{x}^T)^{-1}$. And furthermore, if we assume that $\mathbb{E}(\epsilon) = 0$, we have $\hat{\beta}^T_{MM} = (\sum_i y_i x_i)(\sum x_i x_i^T)^{-1}$. ■

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:

$$loglikelihood = -\sum_i \frac{(y_i - \beta^T x_i)^2}{2\sigma^2} + constant \tag{1.10}$$

Take the derivative with respect to β and set it to 0, we have $\sum_i (y_i - \beta^T x_i)x_i^T = 0$, which is solved by $\hat{\beta}^T_{ML} = (\sum_i y_i x_i^T)(\sum_i x_i x_i^T)^{-1}$ ■

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: As is shown in equation 1.10, maximizing loglikelihood is equivalent to minimizing $\sum_{i=1}^N (y_i - x_i^T\beta)^2$ ■

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: Unconstrained version of ridge regression:

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

Take the derivative with respect to β and set it to 0, we have $-2\sum_i x_i(y_i - x_i^T\beta) + 2\lambda\beta = 0$, which is solved by $\hat{\beta}_{ridge} = (\sum_i x_i x_i^T + \lambda I)^{-1}(\sum_i x_i y_i)$ ■

1.4.1 Uncertainty quantification

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

Proof: Since $\hat{\beta}_{LS} = (X^TX)^{-1}X^Ty$, which is a linear transformation of a normal random vector y, so itself is a normal random vector as well. $\mathbb{E}(\hat{\beta}_{LS}) = \beta$, and $Cov(\hat{\beta}_{LS}) = \sigma^2(X^TX)^{-1}$. Therefore, $\hat{\beta}_{LS} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$. ■

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

Proof: Since $\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^Ty$, which is a linear transformation of a normal random vector y, so itself is a normal random vector as well. $\mathbb{E}(\hat{\beta}_{ridge}) = (X^TX + \lambda I)^{-1}X^TX\beta$, and $Cov(\hat{\beta}_{ridge}) = \sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}$. Therefore, $\hat{\beta}_{ridge} \sim \mathcal{N}((X^TX + \lambda I)^{-1}X^TX\beta, \sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1})$. ■

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: We can estimate σ^2 with $\hat{\sigma^2} = \frac{\sum_i (y_i - x_i^T\beta)^2}{n-df}$. We can show that this is an unbiased estimator. Then $se(\hat{\beta}_{LS}) = diag(\sqrt{\frac{\sum_i (y_i - x_i^T\beta)^2}{n-df}(X^TX)^{-1}})$. The unbiasedness is shown in the following way:

$$\begin{aligned}
\mathbb{E}[\hat{\sigma^2}] &= \mathbb{E}[y^T(I-M)y]/(n-df) \\
&= [tr((I-M)Cov(y)) + (X\beta)^T(I-M)X\beta]/(n-df) \\
&= \sigma^2 tr((I-M))/(n-df) = \sigma^2,
\end{aligned}$$

where $M = X(X^TX)^{-1}X^T$.

For the experiments, I got the same results by using the above formula as the one returned by R built in function.

■

1.4.2 Propogation of uncertainty

Let's now consider the general case where we have a point estimate $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_P)^T$ to some set of parameters $\theta = (\theta_1, \ldots, \theta_P)^T$, and we have an estimate $\hat{\Sigma}$ to the covariance matrix of the sampling distribution of $\hat{\theta}$. If we want to describe our uncertainty about the individual θ_i (as was the case for calculating standard errors in the regression problems above), we can look at the diagonal terms in the covariance matrix, $\hat{\Sigma}_{ii} = \hat{\sigma}_i^2$. If we care, more generally, about a *function* of the θ_i, however, the cross terms will become important.

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

Proof:

$$sd(f(\hat{\theta})) = \sqrt{var(\sum \theta_i)} = \sqrt{\sum var(\theta_i) + 2\sum_{i<j} cov(\theta_i, \theta_j)}.$$

■

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

Proof: Suppose $\hat{\theta}$ is close to θ,

$$sd(f(\hat{\theta})) = \sqrt{var(f(\hat{\theta}))} \approx \sqrt{var(f'(\theta)^T(\hat{\theta} - \theta))} = \sqrt{f'(\theta)^T \hat{\Sigma} f'(\theta)}.$$

■