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

Section 1: Preliminaries

1.1 Exchangeability and de Finetti's theorem

A standard situation in statistics is to be presented with a sequence of observations, and use them to make predictions about future observations. In order to do so, we need to make certain assumptions about the nature of the statistical relationships between the sequence of observations.

A common assumption is that our data are **exchangeable**, meaning that their joint probability is invariant under permutations. More concretely, we say a sequence of N observations is finitely exchangeable if $\mathbf{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_N \in A_n) = \mathbf{P}(X_{\sigma(1)} \in A_1, X_{\sigma(2)} \in A_2, \ldots, X_{\sigma(N)} \in A_n)$ for any permutation of the integers 1 through N, and that an infinite sequence is infinitely exchangeable if this invariance holds for all values of N.

***

Difference between frequentest and Bayesian is in Bayesian probability is a measure of uncertainty but observation is certain. In frequentest approach, beliefs cant be cast into probability, probability corresponds to repetition: Fixed unknown params and random observation. The difference is what is the uncertain quantity or param.
The urn example is cast into β distribution with α and β being the initial weight of red and blue. X_i is iid Bernoulli θ which follows the β distribution. $P(X_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$ where θ follows the β distribution.
***

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.*

Solution:

For any sequence of N observations $\mathbf{S}$ $(s_1, s_2, \ldots, s_\mathbf{N})$, there are B blue ball observations and R red ball observations where B+R=N.

The probability of seeing a sequence $\mathbf{S}$ is given by:

$$\mathbf{P}(\mathbf{S}) = \mathbf{P}(s_N|s_{N-1}, s_{N-2}, \ldots, s_1)\mathbf{P}(s_{N-1}|s_{N-2}, \ldots, s_1)\ldots\mathbf{P}(s_1).$$

For each element in the right-hand side of the equation, the value of the probability is equal to the number of balls available from the particular color divided by the total number of balls in the urn. Therefore, the denominator of this ratio is incremented by 1 after each draw regardless of the type of drawn ball. Hence, the product of all denominators is equal to

$$(N+1)! = (N+1)(N)(N-1)\ldots(3)(2).$$

On the other hand, the nominator of these terms represent the number of balls available from a particular call. Considering the red balls, the product of all nominator values for events of observing red balls is equal

to R!. This is because, regardless of the order, the first red ball is drawn when only 1 is present in the urn, then one red ball is drawn every time the urn contains t red balls where t $\in [2, R]$. Similar reasoning applies for blue balls, and the probability of **S** can be given as:

$$\mathbf{P}(\mathbf{S}) = \frac{R!B!}{(R+B+1)!}.$$

Therefore, the order does not play a role in the probability which means that the sequence is exchangeable. However, the probability of observing a red (or blue) ball is not always the same, so the observations are not *iid*.

Solution End

Loosely speaking, de Finetti's Theorem states if a sequence of random variables is infinitely exchangeable, those random variables must be conditionally i.i.d. given some set of parameters. More formally,

Theorem 1.1 (de Finetti) *Let $(X_1, X_2, \dots)$ be an infinite sequence of random variables in some space $\mathcal{X}$. This sequence is infinitely exchangeable if and only if there exists a probability distribution $\mathbf{Q}_\theta$, parametrized by some random parameter $\theta \sim \nu$, such that the X_i are conditionally iid given $\mathbf{Q}_\theta$ and such that*

$$\mathbf{P}(X_1 \in A_1, X_2 \in A_2, \dots) = \int_\Theta \prod_{i=1}^{\infty} \mathbf{Q}_\theta(A_i)\nu(d\theta).$$

This means we can imagine that any exchangeable sequence has been generated as a sequence of i.i.d. random variables with some unknown law. This provides a motivation for Bayesian inference: We have a hierarchical model, where data are generated according to some distribution parametrized by a random (in the Bayesian context – i.e. unknown/uncertain) variable θ, and our uncertainty about θ is characterized by some distribution ν.

Let's consider the 0/1 form of de Finetti's theorem, for exchangeable sequences of binary variables:

Theorem 1.2 (de Finetti 0/1) *An infinite sequence $(X_1, X_2, \dots)$ of binary random variables is exchangeable if and only if its distribution can be written as*

$$\begin{aligned}\mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_N = x_N) &= \int_0^1 \prod_{i=1}^{N} \left\{\theta^{x_i}(1-\theta)^{1-x_i}\right\} d\nu(\theta) \\ &= \int_0^1 \prod_{i=1}^{N} \theta^k (1-\theta)^{N-k} d\nu(\theta)\end{aligned}$$

where $k = \sum_i x_i$.

We will now work through (most of) a proof in the next two exercises.

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

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

Solution:

Given that there are t ones out of M elements in the sequence and exchangeability holds (all sequences are equally probable) , the above probability can be obtained by the counting the number of possible subsequences with N elements that contain s ones, then dividing the number by the total possible subsequences. In the nominator of the right hand side, the first term forces the subsequence to contain s out of the t available ones. Then, the second term requires the remaining elements in the subsequence to be equal to zero by choosing N-s zeros of the M-t zeros in the original sequence. The denominator counts the number of all possible subsequences.

Solution End

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

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

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

Solution:

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

In the equation above the the first fraction in the right hand side is equal to $\prod_{i=0}^{s-1} M\theta - i$. When $M \to \infty$, this product goes to $M\theta^s$. Similarly, the second fraction goes to $M(1-\theta)^{N-s}$, and the last fraction goes to M^{-N}. Therefore, the integral can be written as

$$\binom{N}{s} \int_0^1 \frac{(M\theta)^s (M(1-\theta))^{N-s}}{M^N} dF_M(\theta) \to \binom{N}{s} \int_0^1 \theta^s (1-\theta)^{N-s} dF_M(\theta).$$

Solution End

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

De Finetti's theorem can be seen as a motivation for Bayesian inference. If our data are exchangeable, we know that they are iid according to some unknown probability distribution $F_\theta(X)$, which we can think of as a **likelihood function**, and that they can be represented using an mixture of such iid sequences. As we saw from the 0/1 case, the distribution over probabilities is given by the limit of the empirical distribution function. When not working in this limit, we may choose to model this distribution over the parameters of our likelihood function using a **prior** distribution $\pi(\theta)$ – ideally one that both assigns probability mass to where we expect the empirical distribution might concentrate, and for which $\int_\Theta F_\theta(X)\pi(d\theta)$ is tractable.

The exponential family of probability distributions is the class of distributions parametrized by θ whose density can be written as

$$p(x|\theta) = h(x)\exp\{\eta(\theta)^T T(x) - A(\eta(\theta))\}$$

where

- $\eta(\theta)$ (sometimes just written as η), is a transformation of θ that is often referred to as the **natural or canonical parameter**.
- $T(X)$ is known as a **sufficient statistic** of X. We see that $p(x|\theta)$ depends only on X through $T(X)$, implying that $T(X)$ contains all the relevant information about X.
- $A(\eta(\theta)$ (or $A(\eta)$) is known as the **cumulant function** or the **log partition function** (remember, a partition function provides a normalizing constant).

Example 1.1 (The Bernoulli distribution) *A Bernoulli random variable X takes the value $X = 1$ with probability π and $X = 0$ with probability $1 - \pi$; it's density can be written:*

$$\begin{aligned} p(x|\pi) =& \pi^x(1-\pi)^{1-x} \\ =& \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\} \end{aligned}$$

By rewriting in this exponential family form, we see that

- $\eta = \log\frac{\pi}{1-\pi}$
- $T(x) = x$
- $A(\eta) = -\log(1-\pi) = \log(1+e^\eta)$
- $h(x) = 1$

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 $x_1, \ldots, x_n$?

Solution:

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!}\exp\{\log\lambda x - \lambda\}$$

- $\eta = \log\lambda$
- $T(x) = x$
- $A(\eta) = \lambda = e^\eta$
- $h(x) = \frac{1}{x!}$

In case of n independent samples:

- $\eta = \log\lambda$
- $T(x) = \sum_{i=1}^n x_i$
- $A(\eta) = n\lambda = ne^\eta$
- $h(x) = \prod_{i=1}^n \frac{1}{x_i!}$

Solution End

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 $x_1,\ldots,x_N$*?*

Solution:

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

For n observations:

- $\eta = \alpha - 1, -\beta$
- $T(x) = \sum_{i=1}^n \log x_i, \sum_{i=1}^n x_i$

Solution End

1.2.1 Cumulants and moments of exponential families

We are probably most familiar with using the PDF or the CDF of a random variable to describe its distribution, but there are other representations that can be useful. The **moment generating function** $M_X(s) = \mathbb{E}[\exp(s^T x)] = \int_{\mathcal{X}} e^{s^T x} p_X(x) dx$ is the Laplace transform of the PDF $p_X(x)$. As the name suggests, we can use the moment-generating function to generate the (uncentered) moments of a random variable; the nth moment is given by

$$m_n = \frac{d^n M_X}{ds^n}\Big|_{s=0}$$

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*

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

Solution:

$$\mathbb{E}[e^{sT(x)}|\eta] = h(x) \int e^{sT(x)} e^{\eta T(x) - A(\eta)} dT(x) = \int h(x) e^{T(x)(s+\eta) - A(\eta)} dT(x)$$
$$= \int h(x) e^{T(x)(s+\eta) - A(\eta) + A(s+\eta) - A(s+\eta)} dT(x) = e^{A(s+\eta) - A(\eta)} \int e^{T(x)(s+\eta) - A(s+\eta)} dT(x)$$

The integral in the equation above is equal to one since it represents a new exponential family probability; hence:

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

Solution End

A related representation is the **cumulant generating function** $C_X(s) = \log \mathbb{E}[e^{s^T x}] = \log(M_X(s))$. Clearly, for exponential families this takes the form $C_{T(X)}(s) = A(\eta+s) - A(\eta)$. This explains why $A(\eta)$ is sometimes called the cumulant function! The cumulant function can be used to generate the cumulants of a distribution as

$$\kappa_n = \frac{d^n C_X}{ds^n}\Big|_{s=0}$$

The first three cumulants are the same as the first three central moments of the distribution – meaning, the cumulant generative function is a useful tool for calculating mean, variance and the third central moment.

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.*

Solution:

$A\eta = \lambda = e^{\eta}$ so $M_{T(X)}(s) = e^{\lambda e^s - \lambda}$ and $C_X(s) = \lambda e^s - \lambda$.

Using the **moment generating function**, the first moment (mean) is:

$$\frac{\partial M_{T(X)}(s)}{\partial s}|_{s=0} = \lambda e^{\lambda(e^s-1)+s}|_{s=0} = \lambda$$

and the second moment is:

$$\frac{\partial^2 M_{T(X)}(s)}{\partial^2 s}|_{s=0} = \lambda(\lambda e^s + 1)e^{\lambda(e^s-1)+s}|_{s=0} = \lambda(1+\lambda)$$

the variance is the difference between the second moment and the square of the mean and can be given by $var = \lambda(1+\lambda) - \lambda^2 = \lambda$.

Using the **cumulant generating function**, mean is:

$$\frac{\partial C_{T(X)}(s)}{\partial s}|_{s=0} = \lambda$$

Similarly taking the second derivative gives the variance which is equal to λ as well.

Solution End

1.2.2 Conjugate priors

Exponential families are very important in Bayesian statistics because, for any exponential family likelihood, we can find an conjugate exponential family prior. If our likelihood takes the form

$$f(x|\eta) = h(x)\exp\left\{\eta^T T(x) - A(\eta)\right\}$$

then a conjugate prior is given by

$$p(\eta|\xi,\nu) = g(\xi,\nu)\exp\left\{\eta^T\xi - \nu A(\eta)\right\}$$

Below are some exercises based on common conjugate priors.

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 $\mu|x_1,\ldots,x_N$*

Solution:

$$p(\mu|x_1,\ldots,x_N) \propto p(x_1,\ldots,x_N|\mu)p(\mu) = \prod_{i=1}^{N} p(x_i|\mu)p(\mu)$$
$$= \sqrt{(\frac{1}{2\pi\sigma^2})}^N \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}\sqrt{\frac{1}{2\pi\sigma_0^2}}\exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}$$
$$= K\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\} = K\exp\left\{\mu^2(\frac{N}{2\sigma^2}+\frac{1}{2\sigma_0^2}) - 2\mu(\frac{\sum_{i=1}^{N}x_i}{2\sigma^2}+\frac{\mu_0}{2\sigma_0^2}) + cts\right\}$$
$$= K\exp\left\{\frac{1}{2}(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2})\left\{\mu^2 - 2\mu(\frac{\sum_{i=1}^{N}x_i}{2\sigma^2}+\frac{\mu_0}{2\sigma_0^2})(\frac{N}{2\sigma^2}+\frac{1}{2\sigma_0^2})^{-1} + cts\right\}\right\}$$

In the equation above, a perfect square can be obtained in the exponential expression by adding some constants and adjusting for them by changing the constant K. The resultant form is:

$$K\exp\left\{\frac{1}{2}(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2})\left\{\mu - (\frac{\sum_{i=1}^{N}x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2})^{-1}\right\}^2\right\}$$

which is the form of a normal distribution with $mean = (\frac{\sum_{i=1}^{N}x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2})^{-1}$ and $var = (\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2})^{-1}$.

Solution End

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,* $\omega = 1/\sigma^2$*:*

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

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

Solution:

$$p(\omega|x_1,\ldots,x_N) \propto p(x_1,\ldots,x_N|\omega)p(\omega) = \sqrt{(\frac{\omega}{2\pi})}^N\exp\left\{-\frac{\omega}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}\frac{\beta^\alpha}{\Gamma(\alpha)}\omega^{\alpha-1}e^{-\beta\omega}$$
$$= \sqrt{(\frac{\omega}{2\pi})}^N\frac{\beta^\alpha}{\Gamma(\alpha)}\omega^{\alpha+N/2-1}e^{-\omega(\beta+\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2)}$$

The posterior is a gamma distribution with $\alpha' = \alpha + N/2$ and $\beta' = \beta + \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2$).

Solution End

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.*

Solution:

$$\int_\omega \sqrt{\frac{\omega}{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \omega^{\alpha+1/2-1} e^{-\omega(\beta+0.5(x-\mu)^2)} d\omega$$
$$= \sqrt{\frac{1}{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \left\{ \frac{\Gamma(\alpha+1/2)}{(\beta+0.5(x-\mu)^2))^{\alpha+1/2}} \right\} \int_\omega \frac{(\beta+(x-\mu)^2))^{\alpha+1/2}}{\Gamma(\alpha+1/2)} \omega^{\alpha+1/2-1} e^{-\omega(\beta+(x-\mu)^2)} d\omega$$
$$= \sqrt{\frac{1}{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \left\{ \frac{\Gamma(\alpha+1/2)}{(1+0.5(x-\mu)^2))^{\alpha+1/2}} \right\}$$

The integral in the equal above goes to one since it is the integral of a pdf. The remaining terms match a student's t distribution.

Solution End

1.3 Multivariate normal distribution

So far, we have looked at univariate random variables - particularly, the univariate normal random variable, which is characterized by its mean and variance. We will often work with the multivariate normal distribution, a natural generalization characterized by a mean vector and a covariance matrix.

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

Solution:
a)

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

but $\mathbb{E}\left[\mu x^T\right] = \mu\mathbb{E}\left[x^T\right] = \mu\mu^T$ and $\mathbb{E}\left[x\mu^T\right] = \mathbb{E}\left[x\right]\mu^T = \mu\mu^T$. Hence,

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

b)

$$\Sigma' = \mathbb{E}\left[(Ax+b-\mathbb{E}(Ax+b))(Ax+b-\mathbb{E}(Ax+b))^T\right] = E\left[(Ax-A\mu)(Ax-A\mu)^T\right] = \mathbb{E}\left[A(x-\mu)(x-\mu)^T A^T\right]$$
$$= A\mathbb{E}\left[(x-\mu)(x-\mu)^T\right] A^T = A\Sigma A^T$$

Solution End

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.*

Solution:
a)

$$M(s) = \int e^{sx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^2-2x\mu+\mu^2-2sx\sigma^2)} dx$$
$$= e^{\frac{2\mu s\sigma^2}{2\sigma^2}} e^{\frac{s^2\sigma^4}{2\sigma^2}} \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^2-2x(\mu+s\sigma^2)+\mu^2+2\mu s\sigma^2+\sigma^4 s^s)} dx$$
$$= e^{\mu s} e^{\frac{s^2\sigma^2}{2}} \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-(\mu+s\sigma^2))^2} dx = e^{\mu s} e^{\frac{s^2\sigma^2}{2}}$$

In the equation above, the integral evaluates to zero because it is the *pdf* of a normal distribution.
b) Since the variables are independent in a multivariate normal distribution, then for a case with k-variables

$$\mathbb{E}[e^{s^T x}] = \prod_{i=1}^{k} \mathbb{E}[e^{s_i x_i}]$$

using the results of part a) and with $\mu = 0$

$$\mathbb{E}[e^{s^T x}] = \prod_{i=1}^{k} e^{\frac{s_i^2}{2}} = e^{\sum_{i=1}^{k} \frac{s_i^2}{2}} = e^{0.5 s^T s}$$

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

Solution End

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...*

Solution:
Using the same reasoning as in 1.12

$$\mathbb{E}[e^{(as)^T x}] = \prod_{i=1}^{k} \mathbb{E}[e^{s_i a_i x_i}] = e^{\sum_{i=1}^{k} \mu s a_i + \frac{s^2 a_i^2 \sigma^2}{2}} = e^{\mu' s + \frac{s^2 \sigma'^2}{2}}$$

The equation shows that the moment generation function of z is that of a univariate normal distribution with mean $\mu' = a^T\mu$ and variance $\sigma'^2 = (a)^T\Sigma a$. The other direction of the if statement can be achieved in a similar reasoning.

Solution End

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.

Solution:

$$M_x(s) = e^{s^T\mu + \frac{s^T\Sigma s}{2}} = e^{s^T\mu + \frac{s^T(D^TD)s}{2}}$$

where $\Sigma^{=} D^TD$. To generate samples for multivariate normal, get z from standard multivariate normal, get D from the desired Σ, then generate the new samples using $x = Dz + \mu$.

Solution End

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

Solution:
if $x = Dz+\mu$, then $z = D^{-1}(x-\mu)$. The *pdf* of x can be related to that of z according to: $pdf(x) = pdf(z)|\frac{\partial z}{\partial x}|$. The determinant of the partial derivative is equal to $|D^{-1} = |\frac{1}{\sqrt{|\Sigma|}}$ since $D = \Sigma^{\frac{1}{2}}$. Then:

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

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

Solution End

1.3.1 Manipulation of multivariate normals

Like its univariate counterpart, the multivariate normal distribution is closed under a number of operations, which we will explore here.

Exercise 1.16 (marginal distribution) *Let us assume that $x \sim Normal(\mu, \Sigma)$, 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.

Solution:
Given $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$, we define the inverse of Ω as $\Sigma^{-1} = \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix}$ where (details in exercise 1.17):

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

Starting from the joint distribution

$$\frac{1}{(2\pi)^{n/2}}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}\left\{(x_1-\mu_1)^T\Omega_{11}(x_1-\mu_1) + 2(x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2) + (x_2-\mu_2)^T\Omega_{22}(x_2-\mu_2)\right\}\right\}$$
$$= \frac{1}{(2\pi)^{n/2}}|\Sigma|^{-1/2}\exp-\frac{1}{2}\left\{(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\right\}$$
$$\exp\left\{-\frac{1}{2}(x_1-\mu_1)^T(\Omega_{11}-\Sigma_{11}^{-1})(x_1-\mu_1) + 2(x_1-\mu_1)^T\Omega_{12}(x_2-\mu_2) + (x_2-\mu_2)^T\Omega_{22}(x_2-\mu_2)\right\}$$

By substituting the expressions of Ω in the equation above, we can get:

$$\frac{1}{(2\pi)^{n/2}}|\Sigma|^{-1/2}\exp-\frac{1}{2}\left\{(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\right\}$$
$$\exp\left\{-\frac{1}{2}((x_2-\mu_2) - \Sigma_{12}^T\Sigma_{11}^{-}1(x_1-\mu_1))^T(\Sigma_{22}-\Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1}(((x_2-\mu_2) - \Sigma_{12}^T\Sigma_{11}^{-1}(x_1-\mu_1))\right\}$$

If the dimensions of x_1 and x_2 are given by n_1 and n_2 respectively and knowing that $|\Sigma| = |\Sigma_{11}||\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}|$, the distribution can be expressed as:

$$\frac{1}{(2\pi)^{n_1/2}}|\Sigma_{11}|^{-1/2}\exp-\frac{1}{2}\left\{(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\right\}$$
$$\frac{1}{(2\pi)^{n_2/2}}|\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}|^{-1/2}\exp\left\{-\frac{1}{2}((x_2-\mu_2) - \Sigma_{12}^T\Sigma_{11}^{-}1(x_1-\mu_1))^T(\Sigma_{22}-\Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1}(((x_2-\mu_2) - \Sigma_{12}^T\Sigma_{11}^{-1}(x_1-\mu_1))\right.$$

The marginal distribution of x_1 is give by:

$$\int pdf(x_1,x_2)dx_2 = \frac{1}{(2\pi)^{n_1/2}}|\Sigma_{11}|^{-1/2}\exp-\frac{1}{2}\left\{(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\right\}$$

The other term in the joint distribution is a normal distribution of x_2, therefore it goes to one after integration. The marginal distribution is therefore Normal with mean μ_1 and covariance Σ_{11}. This problem can also be solved by using the fact that the linear combination of multivariate variables is a normal distribution.
Solution End

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

Solution:
$\Sigma\Omega = I$, so:

$$I = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix} = \begin{pmatrix} \Sigma_{11}\Omega_{11} + \Sigma_{12}\Omega_{12}^T & \Sigma_{11}\Omega_{12} + \Sigma_{12}\Omega_{22} \\ \Sigma_{12}^T\Omega_{11} + \Sigma_{22}\Omega_{12}^T & \Sigma_{12}^T\Omega_{12} + \Sigma_{22}\Omega_{22} \end{pmatrix}$$

$$THEN,$$

$$\Sigma_{11}\Omega_{11} + \Sigma_{12}\Omega_{21}^T = I$$

$$\Sigma_{11}\Omega_{12} + \Sigma_{12}\Omega_{22} = 0 \Rightarrow \Omega_{12} = -\Sigma_{11}^{-1}\Sigma_{12}\Omega_{22}$$

$$\Sigma_{21}^T\Omega_{11} + \Sigma_{22}\Omega_{21} = I \Rightarrow \Omega_{21} = -\Sigma_{22}^{-1}\Sigma_{21}\Omega_{11}$$

$$\Sigma_{12}^T\Omega_{12} + \Sigma_{22}\Omega_{22} = 0$$

Substituting $\Omega_{12} and \Omega_{21}$ in the first and last equation above gives:

$$\Omega_{11} = [\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]^{-1}$$
$$\Omega_{22} = [\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}]^{-1}$$

The equation for Ω_{11} can be rewritten as:

$$\Omega_{11} = [\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]^{-1} = \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_1 2)^{-1}\Sigma_{12}^T\Sigma_{11}^{-1}$$

This can be proven by showing that:

$$[\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}][\Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{12}^T\Sigma_{11}^{-1}] = I$$

$$[\Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{12}^T\Sigma_{11}^{-1}] = [\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]\Sigma_{11}^{-1} +$$
$$[\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}][\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{12}^T\Sigma_{11}^{-1}]$$
$$= I - [\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]\Sigma_{11}^{-1} + [\Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}][(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{12}^T\Sigma_{11}^{-1}]$$
$$= I - [\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]\Sigma_{11}^{-1} + \Sigma_{12}\Sigma_{22}^{-1}[\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}][(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{12}^T\Sigma_{11}^{-1}]$$
$$= I - [\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]\Sigma_{11}^{-1} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T\Sigma_{11}^{-1} = I$$

And

$$\Omega_{12} = \Omega_{21}^T = -\Sigma_{11}^{-1}\Sigma_{12}\Omega_{22} = -\Sigma_{11}^{-1}\Sigma_{12}[\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}]^{-1}$$

The final equations are:

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

Solution End

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.*

Solution:
Getting the conditional probability either of x_1 or x_2 is the same. To make use of the results of the previous

questions, the conditional probability of x_2 is considered here:

$$pdf(x_2|x_1) = \frac{pdf(x_1, x_2)}{pdf(x_1)}$$
$$= \frac{1}{(2\pi)^{n_2/2}}|\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}|^{-1/2}$$
$$\exp\left\{-\frac{1}{2}((x_2-\mu_2) - \Sigma_{12}^T\Sigma_{11}^{-1}(x_1-\mu_1))^T(\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12})^{-1}(((x_2-\mu_2) - \Sigma_{12}^T\Sigma_{11}^{-1}(x_1-\mu_1))\right\}$$

Which is a normal distribution with mean $\mu_2 + \Sigma_{12}^T\Sigma_{11}^{-1}(x_1-\mu_1))$ and covariance $\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}$.

Solution End

1.4 Frequentist estimation and uncertainty quantification

In this section, we're going to go over basic frequentist approaches to inference, with a focus on multiple linear regression (since we're next going to look at Bayesian regression). Some of this should be familiar to you, although we will go into quite some depth. Throughout the remainder of this section, we are going to assume our data follow a linear model, of the form

$$y_i = x_i^T\beta + \epsilon_i, \qquad i = 1, \ldots, N$$

There are a number of options for estimating β. Three commonly used techniques are:

1. **Method of Moments** Select $\hat{\beta}$ so that the empirical moments of the observations match the theoretical moments.
2. **Maximum likelihood** Assume a model for generating the ϵ_i, and find the value of $\hat{\beta}$ that maximizes the likelihood.
3. **Loss function**: Construct a loss function between the y_i and $x_i^T\hat{\beta}$, and minimize that loss function.

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 $\hat{\beta}_{MM}$*

Solution:

$$Cov(x, \epsilon) = E[x\epsilon] - E[x]E[\epsilon] = \frac{1}{N}[\sum_{i=1}^N x_i(y_i - x_i^T\beta)] - 0 = \frac{1}{N}[\sum_{i=1}^N x_iy_i - \sum_{i=1}^N x_ix_i^T\beta)]$$

Since $Cov(x, \epsilon) = 0$, then:

$$\frac{1}{N}[\sum_{i=1}^N x_iy_i - \sum_{i=1}^N x_ix_i^T\beta)] = 0$$
$$\hat{\beta}_{MM} = \left\{\frac{1}{N}\sum_{i=1}^N x_ix_i^T\right\}^{-1}\frac{1}{N}\sum_{i=1}^N x_iy_i$$

Solution End

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.*

Solution:

$$likelihood = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \right\}^N exp \left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \right\}$$

The negative log likelihood can be expressed as (K being a constant):

$$n\_log\_likelihood = K + \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

$\hat{\beta}_{ML}$ is the value of β that minimizes the negative log likelihood.

$$\frac{\partial [K + \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2]}{\partial \beta} = 0$$

$$\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i x_i^T \beta = 0$$

$$\hat{\beta}_{ML} = \left\{ \sum_{i=1}^{N} x_i x_i^T \right\}^{-1} \sum_{i=1}^{N} x_i y_i = \hat{\beta}_{MM}$$

Solution End

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.*

Solution:
The quadratic loss optimization problem is equivalent to that of minimizing the negative log likelihood and will eventually give the same estimator for $\hat{\beta}_{QL}$.
Solution End

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 $t \geq 0$.

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.

Solution:
The problem can be expressed as:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \frac{\alpha}{2} \sum_{j=1}^{p} \beta^T \beta - t$$

To solve the problem, the derivative with respect to β is set to zero:

$$-\sum_{i=1}^{N} x_i y_i + \sum_{i=1}^{N} x_i x_i^T \beta + \alpha\beta = 0$$

$$\left\{\sum_{i=1}^{N} x_i x_i^T + \alpha I\right\}\beta = \sum_{i=1}^{N} x_i y_i$$

$$\hat{\beta}_{ridge} = \left\{\sum_{i=1}^{N} x_i x_i^T + \alpha I\right\}^{-1} \sum_{i=1}^{N} x_i y_i = W\hat{\beta}_{MLE}$$

$$where; W = \left\{\sum_{i=1}^{N} x_i x_i^T + \alpha I\right\}^{-1} \sum_{i=1}^{N} x_i x_i^T$$

Adding the regularization term will help avoid overfitting by forcing the estimators to take small values if possible.
Solution End

1.4.1 Uncertainty quantification

In a frequentist context, we typically quantify our uncertainty by looking at the sampling distribution of our estimator. Let's assume that our errors are normally distributed, i.e. (in vector notation)

$$y = X\beta + \epsilon, \qquad \epsilon \sim \text{Normal}(0, \sigma^2 I)$$

.

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

Solution:
Assuming x has zero mean (only to simplify notation), we can instead define $z = x - \mu_x$, work with z and express the final answer in terms of x

$$\hat{\beta}_{MM} = \left\{\sum_{i=1}^{N} x_i x_i^T\right\}^{-1} \sum_{i=1}^{N} x_i y_i$$

$$\mu(\hat{\beta}_{MM}) = E[\left\{\sum_{i=1}^{N} x_i x_i^T\right\}^{-1} \sum_{i=1}^{N} x_i y_i] = E[\left\{\sum_{i=1}^{N} x_i x_i^T\right\}^{-1} \sum_{i=1}^{N} x_i (x_i^T\beta - \epsilon)]$$

$$\mu(\hat{\beta}_{MM}) = \beta + E[\left\{\sum_{i=1}^{N} x_i x_i^T\right\}^{-1} \sum_{i=1}^{N} x_i (-\epsilon)] = \beta$$

This holds since the the predictors and residuals have zero covariance and the mean of the residual is zero.

$$var(\hat{\beta}_{MM}) = var\left\{\left\{\sum_{i=1}^{N} x_i x_i^T\right\}^{-1} \sum_{i=1}^{N} x_i y_i\right\} = \left\{\sum_{i=1}^{N} x_i x_i^T\right\}^{-2} var\left\{\sum_{i=1}^{N} x_i y_i\right\}$$

Since $\left\{\sum_{i=1}^{N} x_i x_i^T\right\}$ is assumed to be constant. Then:

$$var(\hat{\beta}_{MM}) = \left\{\sum_{i=1}^{N} x_i x_i^T\right\}^{-2} \sigma^2 \left\{\sum_{i=1}^{N} x_i x_i^T\right\} = \left\{\sum_{i=1}^{N} x_i x_i^T\right\}^{-1} \sigma^2$$

Solution End

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

Solution:

$$\hat{\beta}_{ridge} = \left\{\sum_{i=1}^{N} x_i x_i^T + \alpha I\right\}^{-1} \sum_{i=1}^{N} x_i y_i$$

The variance can be expressed as:

$$var(\hat{\beta}_{ridge}) = W var(\hat{\beta}_{MM}) W^T = \sigma^2 W \left\{\sum_{i=1}^{N} x_i x_i^T\right\}^{-1} W^T$$

$$\mu(\hat{\beta}_{ridge}) = W\beta$$

since W is constant in this expectation.

Solution End

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`*?*

Solution:
Since the residuals are normally distributed, then:

$$\frac{\sum_{i=1}^{N}(y-\hat{y})^2}{\sigma^2} \sim \chi^2(N-p)$$

So,

$$E(\frac{\sum_{i=1}^{N}(y-\hat{y})^2}{\sigma^2}) = N - p \Rightarrow \sigma^2 = \frac{\sum_{i=1}^{N}(y-\hat{y})^2}{N-p}$$

which gives the unbiased estimator for σ^2.

Code available on Github
Solution End

1.4.2 Propagation 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 $f(\theta)$?*

Solution:

$$Var(\sum_{i=1}^{P} \theta) = \sum_{i=1}^{P}\sum_{j=1}^{P} Cov(\theta_i, \theta_j) = \sum_{i=1}^{P} Var(\theta_i) + 2 \sum_{1 \le i < j \le P} Cov(\theta_i, \theta_j)$$

$$s.e(\sum_{i=1}^{P} \theta) = \sqrt{\sum_{i=1}^{P} Var(\theta_i) + 2 \sum_{1 \le i < j \le P} Cov(\theta_i, \theta_j)}$$

Solution End

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

Solution:
The Taylor expansion of $f(\theta)$ at $\theta = \mu_\theta$ is given by:

$$f(\theta) = f(\mu_\theta) + \frac{\partial f}{\partial \theta}^T |_{\theta=\mu_\theta}(\theta - \mu_\theta) + \cdots \approx f(\mu_\theta) + \frac{\partial f}{\partial \theta}|_{\theta=\mu_\theta}(\theta - \mu_\theta)$$

$$Var(f(\theta)) = \frac{\partial f}{\partial \theta}^T |_{\theta=\mu_\theta} Cov(\theta) \frac{\partial f}{\partial \theta}|_{\theta=\mu_\theta}$$

The standard error is the square root of the variance above when $\mu_\theta = \hat{\theta}$.
Solution End