University of California, Los Angeles
Department of Statistics

Statistics C173/C273 **Instructor: Nicolas Christou**

Homework 1

EXERCISE 1

Suppose $Y_i = \mu + \epsilon_i$, with $E(\epsilon_i) = 0, \text{var}((Y_i) = \sigma^2$, and $\text{cov}(Y_i, Y_j) = \sigma^2 \rho^{|i-j|}$. The inverse of the variance covariance matrix is given as follows:

$$\mathbf{\Sigma^{-1}} = \frac{1}{\sigma^2(1-\rho^2)} \begin{pmatrix} 1 & -\rho & 0 & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & \dots & 0 \\ 0 & 0 & -\rho & 1+\rho^2 & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots. & \ddots & -\rho \\ 0 & 0 & \dots & 0 & -\rho & 1 \end{pmatrix}.$$

a. It is given that $\hat{\mu} = \frac{\mathbf{1'\Sigma^{-1}Y}}{\mathbf{1'\Sigma^{-1}1}}$. Show that $\text{var}(\hat{\mu}) = \sigma^2 \frac{1+\rho}{(n-2)(1-\rho)+2}$. Explain what happens to $\text{var}(\hat{\mu})$ when $\rho = 0$ and when $\rho = 1$.

b. For the same sample size n, the variance of $\hat{\mu}$ for this model (correlated data) is less precise than $\text{var}(\bar{Y}) = \frac{\sigma^2}{n}$ in the i.i.d. case. Show that $\text{var}(\hat{\mu})$ is more variable than $\text{var}(\bar{Y})$ by a factor of $\left[\frac{1-\rho}{1+\rho} + \frac{2}{n}\frac{\rho}{1+\rho}\right]^{-1}$.

EXERCISE 2

Use `R` to access the Maas river data. These data contain the concentration of lead and zinc in `ppm` at 155 locations at the banks of the Maas river in the Netherlands. You can read the data in `R` as follows:

```
a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics_c173_c273/soil.txt",
     header=TRUE)
```

a. Compute the summary statistics for `lead` and `zinc`.

b. Plot the histogram of `lead` and `log(lead)`.

c. Plot `log(lead)` against `log(zinc)`. What do you observe?

d. The level of risk for surface soil based on lead concentration in `ppm` is given on the table below:

Mean concentration (`ppm`)	Level of risk
Below 150	Lead-free
Between 150-400	Lead-safe
Above 400	Significant environmental lead hazard

Use similar `R` commands as in pages 8-11 of handout #4 to create a bubble plot with different colors and sizes of the lead concentration at these 155 locations.

EXERCISE 3

Consider the equal correlation model, $cov(Y_i, Y_j) = \rho\sigma^2$. The variance covariance matrix is of the form $(a-b)\mathbf{I} + b\mathbf{J}$, where $a = 1, b = \rho, \mathbf{J} = \mathbf{1}\mathbf{1}'$. Therefore, in our model $\mathbf{\Sigma} = \sigma^2\left[(1-\rho)\mathbf{I} + \rho\mathbf{J}\right]$. The inverse of this special matrix can be obtained as follows: $\mathbf{\Sigma^{-1}} = \frac{1}{\sigma^2(1-\rho)}\left[\mathbf{I} - \frac{\rho}{1+(n-1)\rho}\mathbf{J}\right]$. The GLS estimator of μ is $\hat{\mu} = \frac{\mathbf{1'\Sigma^{-1}Y}}{\mathbf{1'\Sigma^{-1}1}}$ with variance $var(\hat{\mu}) = \frac{1}{\mathbf{1'\Sigma^{-1}1}}$ and using the inverse of the variance covariance matrix above we get $var(\hat{\mu}) = \frac{\sigma^2}{n}\left[1 + (n-1)\rho\right]$. Show that the variance of the error of prediction is $\sigma^2_{pred} = \sigma^2\left[1 + \frac{1}{n} \times c\right]$. Find the constant c (it will be a function of n and ρ) and find the condition on ρ that will make the prediction variance σ^2_{pred} in the correlated data more precise than the $var(Y_0 - \bar{Y}) = \sigma^2(1 + \frac{1}{n})$ in the i.i.d. case.

EXERCISE 4

Suppose $Y_1, , Y_n$ are normally distributed with $E(Y_i) = \mu, var(Y_i) = \sigma^2, cov(Y_i, Y_j) = \rho\sigma^2$. Define $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$. Are $\bar{Y}, S^2$ independent? Is S^2 an unbiased estimator of σ^2?

EXERCISE 5

Assume that $Y_1, \ldots, Y_n$ follow the normal distribution with mean μ_y, variance σ^2, and $cov(Y_i, Y_j) = \rho\sigma^2$. Consider a second sample such that $X_1, \ldots, X_n$ follow the normal distribution with mean μ_x, variance σ^2, and $cov(X_i, X_j) = \rho\sigma^2$. The two samples are independent.

a. Show that $var(\bar{Y}) = \frac{\sigma^2}{n}[1 + (n-1)\rho]$.

b. Assume that $\rho > 0$. Which sample mean is more dispersed? The one that takes into account the correlation or the one that ignores correlation (i.i.d. case).

c. Consider the case with the correlated $Y_i's$. Find $E\bar{Y}$. Is $\bar{Y}$ unbiased estimator of μ_y? Find $\lim_{n\to\infty} var(\bar{Y})$. Is $\bar{Y}$ consistent estimator of μ?

d. Assume that σ^2 is known and that the data are uncorrelated. Write the test statistic for testing the hypothesis $H_0 : \mu_x = \mu_y$.

e. Refer to question (d). Assume now that the data are correlated. Write the test statistic for testing the hypothesis $H_0 : \mu_x = \mu_y$.

f. Compare (d) with (e). Which tests has smaller p-value? Which test rejects more often than it should?

g. If n denotes the sample size of the correlated observations find the corresponding sample size of uncorrelated observations.

h. Consider a confidence interval for μ_x using the correlated data of the second sample: $\bar{x} \pm 1.96\sigma\sqrt{\frac{1+(n-1)\rho}{n}}$. Now, suppose we fail to see the dependence in the random variables, and we decided to use $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$. What is the actual coverage of our confidence interval if $n = 25, \rho = 0.2$?