School of Economics

Doctoral Programme in Economics

Probability and statistics

Written examination

February 11th, 2022

Name and surname: ____________________ ID: □□□□□□□□

Instructions

Read the problems carefull before staring your work. There are four problems. You can have a sheet with formulae and a mathematical handbook. Write your solutions on the paper provided. You have two hours.

Problem	a.	b.	c.	d.	
1.				•	
2.				•	
3.			•	•	
4.			•	•	
Total					

1. (25) Assume that every unit in a population of size N has two values of statistical variables X and Y. Denote the values by $(x_1, y_1), \ldots, (x_N, y_N)$. Assume that the population mean μ_X and the population variance σ_X^2 of the variable X are known.

Suppose a simple random sample of size n is selected from the population. Denote by $(X_1, Y_1), \ldots, (X_n, Y_n)$ the sample values. The above assumptions are that

$$E(X_k) = \mu_X \qquad \text{and} \qquad \mathrm{var}(X_k) = \sigma_X^2$$

for $k = 1, 2, \ldots, n$.

a. (10) Denote $c = \mathrm{cov}(X_1, Y_1)$. Compute $\mathrm{cov}(X_k, Y_l)$ for $k \neq l$.

Hint: what would be $\mathrm{cov}(X_k, Y_1 + Y_2 + \cdots + Y_N)$*? Use symmetry.*

Solution: by symmetry the covariances $\mathrm{cov}(X_k, Y_l)$ *are the same for all* $k \neq l$*. The covariance in the hint is* 0 *because the second sum is a constant. By properties of covariance we have*

$$\mathrm{cov}(X_k, Y_k) + (N-1)\,\mathrm{cov}(X_k, Y_l) = 0\,,$$

and hence

$$\mathrm{cov}(X_k, Y_l) = -\frac{c}{N-1}\,.$$

b. (10) Assume the quantity $c = \mathrm{cov}(X_1, Y_1)$ is known. We would like to estimate the population mean μ_Y of the variable Y. The following estimator is proposed:

$$\hat{\mu}_Y = \bar{Y} - \frac{c}{\sigma_X^2}\left(\bar{X} - \mu_X\right)\,.$$

Argue that the estimator is unbiased and compute its variance.

Solution: The estimators $\bar{X}$ *and* $\bar{Y}$ *are unbiased and the claim follows by linearity. We compute*

$$\begin{aligned}
\mathrm{var}(\tilde{Y}) &= \mathrm{var}(\bar{Y}) + \frac{c^2}{\sigma_X^4}\,\mathrm{var}(\bar{X}) - \frac{2c}{\sigma_X^2}\,\mathrm{cov}(\bar{Y}, \bar{X}) \\
&= \frac{\sigma_Y^2}{n} \cdot \frac{N-n}{N-1} + \frac{c^2}{\sigma_X^4} \cdot \frac{\sigma_X^2}{n} \cdot \frac{N-n}{N-1} - \frac{2c}{n^2\sigma_X^2}\left(nc - (n^2 - n)\,\frac{c}{N-1}\right) \\
&= \frac{N-n}{N-1}\,\frac{1}{n}\left(\sigma_Y^2 - \frac{c^2}{\sigma_X^2}\right)\,.
\end{aligned}$$

c. (5) Assume the quantity $c = \mathrm{cov}(X_1, Y_1)$ is known. Another possible estimator of μ_Y is $\tilde{\mu}_Y = \bar{Y}$ which is unbiased. Under which circumstances is the estimator

$$\hat{\mu}_Y = \bar{Y} - \frac{c}{\sigma_X^2}\left(\bar{X} - \mu_X\right)\,.$$

more accurate than the estimator $\tilde{\mu}_Y$? Explain your answer.

Solution: Both estimators are unbiased and the variance of $\tilde{Y}$ *is always smaller than the variance of* $\bar{X}$ *unless* $c = 0$*.*

2. (25) Assume that our observations are pairs $(x_1, y_1), \ldots, (x_n, y_n)$. We assume that the pairs are independent samples from the distribution with density

$$f(x,y) = e^{-x} \cdot \frac{1}{\sigma\sqrt{2\pi x}}\, e^{-\frac{(y-\theta x)^2}{2\sigma^2 x}}$$

for $x > 0$, $-\infty < y < \infty$ and $\sigma^2 > 0$. Assume as known that the random variable

$$Z = \frac{Y_1 - \theta X_1}{\sqrt{X_1}}$$

is distributed normally as $\mathrm{N}(0, \sigma^2)$ and is independent of X_1.

a. (5) Find the maximum likelihood estimates for the parameters θ and σ.

Solution: the log-likelihood function is

$$\ell\,(\theta, \sigma | \mathbf{x}, \mathbf{y}) = \sum_{k=1}^{n} \left(-\frac{n}{2} \log 2\pi - n \log \sigma - \frac{(y_k - \theta x_k)^2}{2\sigma^2 x_k} \right) .$$

Computing the partial derivatives we get

$$\frac{\partial \ell}{\partial \theta} = \sum_{k=1}^{n} \frac{(y_k - \theta x_k)}{\sigma^2}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{k=1}^{n} \frac{(y_k - \theta x_k)^2}{\sigma^3 x_k}$$

Setting the partial derivatives to zero, from the first equation we get

$$\hat{\theta} = \frac{\sum_{k=1}^{n} y_k}{\sum_{k=1}^{n} x_k}$$

and from the second

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} \frac{(y_k - \hat{\theta} x_k)^2}{x_k} .$$

b. (10) Compute the Fisher information matrix and give approximate standard errors for the above estimators.

Solution: compute for $n = 1$:

$$\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{x}{\sigma^2} ,$$

$$\frac{\partial^2 \ell}{\partial \theta\, \partial \sigma} = -2 \frac{y - \theta x}{\sigma^3} ,$$

$$\frac{\partial^2 \ell}{\partial \sigma^2} = \frac{1}{\sigma^2} - 3 \frac{(y - \theta x)^2}{\sigma^4 x} .$$

Replace x by X and y by Y. From the first part we infer that $X \sim \exp(1)$, so

$$E\left[\frac{\partial^2 \ell}{\partial \theta^2}(\theta, \sigma | X, Y)\right] = -\frac{E(X)}{\sigma^2} = -\frac{1}{\sigma^2}\,.$$

For the other two expectation we use the known fact from the text to get

$$\begin{aligned} E\left[\frac{\partial^2 \ell}{\partial \theta\, \partial \sigma}(\theta, \sigma | X, Y)\right] &= -\frac{2E\big(Z\sqrt{X}\big)}{\sigma^3} = 0\,, \\ E\left[\frac{\partial^2 \ell}{\partial \sigma^2}(\theta, \sigma | X, Y)\right] &= \frac{1}{\sigma^2} - \frac{3E(Z^2)}{\sigma^4} = -\frac{2}{\sigma^2}\,. \end{aligned}$$

The Fisher information matrix is

$$I(\theta, \sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}\,,$$

and the approximate standard errors are

$$\mathrm{se}(\hat{\theta}) = \frac{\sigma}{\sqrt{n}} \quad \textit{and} \quad \mathrm{se}(\hat{\sigma}) = \frac{\sigma}{\sqrt{2n}}\,.$$

c. (10) Compute the exact standard error of the maximum likelihood estimator $\hat{\theta}$. Assume as known that the density of the pair $(\sum_{k=1}^n X_k, \sum_{k=1}^n Y_k)$ is

$$f(x, y) = \frac{1}{(n-1)!} x^{n-1} e^{-x} \cdot \frac{1}{\sqrt{2\pi x}\sigma} e^{-\frac{(y-\theta x)^2}{2\sigma^2 x}}\,.$$

Solution: using the given density we get

$$\begin{aligned} E\left(\hat{\theta}\right) &= \int_0^\infty \int_{-\infty}^\infty \frac{y}{x} \frac{1}{(n-1)!} x^{n-1} e^{-x} \cdot \frac{1}{\sqrt{2\pi x}\sigma} e^{-\frac{(y-\theta x)^2}{2\sigma^2 x}} dy dx \\ &= \frac{\theta}{(n-1)!} \int_0^\infty x^{n-1} e^{-x} dx \\ &= \theta\,, \end{aligned}$$

and

$$\begin{aligned} E\left(\hat{\theta}^2\right) &= \int_0^\infty \int_{-\infty}^\infty \frac{y^2}{x^2} \frac{1}{(n-1)!} x^{n-1} e^{-x} \cdot \frac{1}{\sqrt{2\pi x}\sigma} e^{-\frac{(y-\theta x)^2}{2\sigma^2 x}} dy dx \\ &= \frac{1}{(n-1)!} \int_0^\infty \frac{\sigma^2 x + \theta^2 x^2}{x^2} x^{n-1} e^{-x} dx \\ &= \frac{\sigma^2}{n-1} + \theta^2\,. \end{aligned}$$

Finally,

$$\mathrm{var}(\hat{\theta}) = \frac{\sigma^2}{n-1}\,.$$

3. (25) Assume the observed values are pairs $(x_1, y_1), \ldots, (x_n, y_n)$. We assume that the pairs are an i.i.d. sample from the bivariate normal density given by

$$f(x,y) = \frac{1}{2\pi\sqrt{ab-c^2}} e^{-\frac{bx^2-2cxy+ay^2}{2(ab-c^2)}}$$

where $a, b > 0$ and $ab - c^2 > 0$. We would like to test the hypothesis

$$H_0\colon c = 0 \qquad \text{versus} \qquad H_1\colon c \neq 0\,.$$

a. (15) Assume as known that the unrestricted maximum likelihood estimates of the parameters are given by

$$\begin{pmatrix} \hat{a} & \hat{c} \\ \hat{c} & \hat{b} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{k=1}^n x_k^2 & \frac{1}{n}\sum_{k=1}^n x_k y_k \\ \frac{1}{n}\sum_{k=1}^n x_k y_k & \frac{1}{n}\sum_{k=1}^n y_k^2 \end{pmatrix}$$

Find the likelihood ratio statistic λ for the testing problem.

Solution: the log-likelihood function is given by

$$\ell\,(a,b,c|\mathbf{x},\mathbf{y}) = -n\log 2\pi - \frac{n}{2}\log(ab-c^2) - \frac{1}{2(ab-c^2)}\sum_{k=1}^n (bx_k^2 - 2cx_ky_k + ay_k^2)\,.$$

Using the known unrestricted maximum likelihood estimates we get

$$\ell\left(\hat{a},\hat{b},\hat{c}|\mathbf{x},\mathbf{y}\right) = -n\log 2\pi - \frac{n}{2}\log(\hat{a}\hat{b}-\hat{c}^2) - \frac{1}{2(\hat{a}\hat{b}-\hat{c}^2)}\sum_{k=1}^n (\hat{b}x_k^2 - 2\hat{c}x_ky_k + \hat{a}y_k^2)\,.$$

We need to simplify the last expression. Summing up we get

$$\sum_{k=1}^n (\hat{b}x_k^2 - 2\hat{c}x_ky_k + \hat{a}y_k^2) = \hat{b}n\hat{a} - 2\hat{c}n\hat{c} + \hat{a}n\hat{b}\,.$$

It follows that

$$\ell\left(\hat{a},\hat{b},\hat{c}|\mathbf{x},\mathbf{y}\right) = -n\log 2\pi - \frac{n}{2}\log(\hat{a}\hat{b}-\hat{c}^2) - n\,.$$

In the restricted case we need to maximize

$$\ell\,(a,b|\mathbf{x},\mathbf{y}) = -n\log 2\pi - \frac{n}{2}\log a - \frac{n}{2}\log b - \frac{1}{2a}\sum_{k=1}^n x_k^2 - \frac{1}{2b}\sum_{k=1}^n y_k^2\,.$$

The above expression is maximized when the terms containing a and b are maximized. We get

$$\tilde{a} = \frac{1}{n}\sum_{k=1}^n x_k^2 \quad \textit{and} \quad \tilde{b} = \frac{1}{n}\sum_{k=1}^n y_k^2\,.$$

It follows

$$\ell\left(\tilde{a},\tilde{b},0|\mathbf{x},\mathbf{y}\right) = -n\log 2\pi - \frac{n}{2}\log\tilde{a} - \frac{n}{2}\log\tilde{b} - n\,.$$

We have

$$\lambda = n\left(-\log(\hat{a}\hat{b}-\hat{c}^2) + \log\tilde{a} + \log\tilde{b}\right)\,.$$

b. (10) What is the approximate distribution of λ under H_0?

Solution: by Wilks's theorem $\lambda \sim \chi^2(r)$ *where* $r = 3 - 2 = 1$.

4. (25) Assume the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $E(\boldsymbol{\epsilon}) = 0$ and $\text{var}(\boldsymbol{\epsilon}) = \sigma(\mathbf{I} + a\mathbf{1}\mathbf{1}^T)$ for $a > 0$. The constant a is assumed to be known.

a. (15) Prove that

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}^T\left(\mathbf{I} + c\mathbf{1}\mathbf{1}^T\right)\mathbf{X}\right]^{-1}\mathbf{X}^T\left(\mathbf{I} + c\mathbf{1}\mathbf{1}^T\right)\mathbf{Y}$$

for

$$c = -\frac{a}{1 + an}$$

is the best unbiased linear estimator of $\boldsymbol{\beta}$.

Hint: Check that

$$\left(\mathbf{I} + a\mathbf{1}\mathbf{1}^T\right)\left(\mathbf{I} + c\mathbf{1}\mathbf{1}^T\right) = \mathbf{I}\,.$$

Solution: let $\tilde{\boldsymbol{\beta}}$ be an arbitrary unbiased estimator of $\boldsymbol{\beta}$. This means that

$$\tilde{\boldsymbol{\beta}} = \mathbf{L}\mathbf{Y}$$

for a matrix $\mathbf{L}$ such that

$$\mathbf{L}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}\,.$$

Compute

$$\begin{aligned} \text{var}\left(\tilde{\boldsymbol{\beta}}\right) &= \text{var}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}\right) \\ &= \text{var}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right) + \text{var}\left(\hat{\boldsymbol{\beta}}\right) + 2\,\text{cov}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{\beta}}\right). \end{aligned}$$

Denote

$$\mathbf{A} = \left(\mathbf{I} + a\mathbf{1}\mathbf{1}^T\right) \quad \textit{and} \quad \mathbf{C} = \mathbf{I} + c\mathbf{1}\mathbf{1}^T\,.$$

We have

$$\mathbf{AC} = \mathbf{I} + a\mathbf{1}\mathbf{1}^T + c\mathbf{1}\mathbf{1}^T + ac\mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T = \mathbf{I} + (a + c + nac)\mathbf{1}\mathbf{1}^T = \mathbf{I}\,.$$

Using $\text{cov}\left(\mathbf{Y}, \mathbf{Y}\right) = \sigma^2\mathbf{A}$ *compute*

$$\begin{aligned} \text{cov}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}\right) &= \text{cov}\left(\left(\mathbf{L} - \left(\mathbf{X}^T\mathbf{C}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{C}\right)\mathbf{Y},\ \left(\mathbf{X}^T\mathbf{C}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{C}\mathbf{Y}\right) \\ &= \sigma^2\left(\mathbf{L} - \left(\mathbf{X}^T\mathbf{C}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{C}\right)\mathbf{A}\mathbf{C}\mathbf{X}\left(\mathbf{X}^T\mathbf{C}\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{L}\mathbf{X} - \mathbf{I}\right)\left(\mathbf{X}^T\mathbf{C}\mathbf{X}\right)^{-1} \\ &= 0\,. \end{aligned}$$

The assertion follows as in the proof of the Gauss-Markov theorem.

b. (10) Suggest an unbiased estimator for the parameter σ^2. Show that it is unbiased.

Solution: compute

$$\hat{\boldsymbol{\epsilon}} = \begin{bmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \vdots \\ \hat{\epsilon}_n \end{bmatrix} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\,.$$

We have

$$\hat{\boldsymbol{\epsilon}} = \Big(\mathbf{I} - \mathbf{X}\big(\mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{C}\Big)\boldsymbol{\epsilon}$$

and

$$\begin{aligned}\sum_{k=1}^n \hat{\epsilon}_k^2 &= \big(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\big)^T\big(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\big) \\ &= \boldsymbol{\epsilon}^T\Big(\mathbf{I} - \mathbf{C}\mathbf{X}\big(\mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1}\mathbf{X}^T\Big)\Big(\mathbf{I} - \mathbf{X}\big(\mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{C}\Big)\boldsymbol{\epsilon} \\ &= \mathrm{Tr}\Big[\Big(\mathbf{I} - \mathbf{C}\mathbf{X}\big(\mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1}\mathbf{X}^T\Big)\Big(\mathbf{I} - \mathbf{X}\big(\mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{C}\Big)\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\Big]\,.\end{aligned}$$

Since $\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T = \sigma^2\mathbf{A}$, *we have*

$$\begin{aligned}E\left(\sum_{k=1}^n \hat{\epsilon}_k^2\right) &= \sigma^2\,\mathrm{Tr}\Big[\Big(\mathbf{I} - \mathbf{C}\mathbf{X}\big(\mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1}\mathbf{X}^T\Big)\Big(\mathbf{I} - \mathbf{X}\big(\mathbf{X}^T\mathbf{C}\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{C}\Big)\mathbf{A}\Big] \\ &= \sigma^2\,\mathrm{Tr}\big(\mathbf{A} - \mathbf{X}(\mathbf{X}^T\mathbf{C}\mathbf{X})^{-1}\mathbf{X}^T\big)\,.\end{aligned}$$

It follows that

$$\hat{\sigma}^2 = \frac{1}{\mathrm{Tr}\big(\mathbf{A} - \mathbf{X}(\mathbf{X}^T\mathbf{C}\mathbf{X})^{-1}\mathbf{X}^T\big)}\sum_{k=1}^n \hat{\epsilon}_k^2$$

is an unbiased estimator of σ^2.