University of Ljubljana, Faculty of Economics

Quantitative finance and actuarial science

Probability and statistics

Written examination

February 1st, 2018

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) Let the observed values $x_1, x_2, \ldots, x_n$ be generated as independent, identically distributed random variables $X_1, X_2, \ldots, X_n$ with distribution

$$P(X_1 = x) = \frac{(\theta-1)^{x-1}}{\theta^x}$$

for $x = 1, 2, 3, \ldots$ and $\theta > 1$.

a. (10) Find the MLE estimate of θ based on the observations.

Solution: we find

$$\ell(\theta, \mathbf{x}) = \left(\sum_{k=1}^{n} x_k - n\right)\log(\theta-1) - \left(\sum_{k=1}^{n} x_k\right)\log\theta\,.$$

Taking the derivative we have

$$\ell'(\theta, \mathbf{x}) = \frac{\sum_{k=1}^{n} x_k - n}{\theta - 1} - \frac{\sum_{k=1}^{n} x_k}{\theta} = 0\,.$$

It follows that

$$\hat{\theta} = \frac{1}{n}\sum_{k=1}^{n} x_k = \bar{x}\,.$$

b. (15) Write an approximate 99%-confidence interval for θ based on the observations. Assume as known that

$$\sum_{x=1}^{\infty} x a^{x-1} = \frac{1}{(1-a)^2}$$

for $|a| < 1$.

Solution: we have

$$\ell''(\theta, x) = -\frac{x-1}{(\theta-1)^2} + \frac{x}{\theta^2}\,.$$

To find the Fisher information we need

$$E(X_1) = \sum_{x=1}^{\infty} x\frac{(\theta-1)^{x-1}}{\theta^x}\,.$$

Using the hint we get

$$E(X_1) = \frac{1}{\theta}\cdot\left(1 - \frac{\theta-1}{\theta}\right)^{-2} = \theta\,.$$

We have

$$I(\theta) = \frac{1}{\theta(\theta-1)}\,.$$

An approximate 99%-confidence interval is

$$\hat{\theta} \pm 2.56\cdot\sqrt{\frac{\hat{\theta}(\hat{\theta}-1)}{n}}\,.$$

3. (25) Assume that your observations are pairs $(x_1, y_1), \ldots, (x_n, y_n)$. Assume the pairs are an i.i.d. sample from the density

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

for $\sigma > 0$, $x > 0$, $-\infty < y < \infty$. We would like to test the hypothesis

$$H_0\colon \theta = 0 \quad \text{versus} \quad H_1\colon \theta \neq 0\,.$$

a. (10) Find the maximum likelihood estimates for θ and σ.

Solution: the log-likelihood function is

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

Take partial derivatives to get

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

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

Set the partial derivatives to 0. *From the first equation we have*

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

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. (15) Find the likelihood ratio statistic for testing the above hypothesis. What is the approximate distribution of the test statistic under H_0?

Solution: if $\theta = 0$ *the log-likelihood functions attains its maximum for*

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

It follows that

$$\lambda = -n\log\left(1 - \frac{\left(\sum_{k=1}^n y_k\right)^2}{\sum_{k=1}^n x_k \sum_{k=1}^n \frac{y_k^2}{x_k}}\right).$$

The approximate distribution od λ *is* $\chi^2(1)$.

4. (25) Assume the regression model

$$\begin{aligned} Y_1 &= \alpha + \beta x_1 + \epsilon_1 \\ Y_2 &= \alpha + \beta x_2 + \epsilon_1 + \epsilon_2 \\ \cdots &= \cdots\cdots\cdots \\ Y_n &= \alpha + \beta x_2 + \epsilon_1 + \epsilon_2 + \cdots + \epsilon_n \end{aligned}$$

where we assume $E(\epsilon_k) = 0$, $\mathrm{var}(\epsilon_k) = \sigma^2$ for all $k = 1, 2, \ldots, n$, and $\mathrm{cov}(\epsilon_k, \epsilon_l) = 0$ for $k \neq l$. Assume that all $x_1, x_2, \ldots, x_n$ are different.

a. (10) Find explicitly the best unbiased linear estimators of α and β.

Solution: define

$$\boldsymbol{U} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_n \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 - Y_1 \\ Y_3 - Y_2 \\ \vdots \\ Y_n - Y_{n-1} \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} 1 & x_1 \\ 0 & x_2 - x_1 \\ 0 & x_3 - x_2 \\ \vdots & \vdots \\ 0 & x_n - x_{n-1} \end{bmatrix}, \quad \boldsymbol{\gamma} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

We have $\mathbf{U} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ *and the usual assumptions of the Gauss-Markov theorem are met. Denote*

$$\begin{aligned} Q_1 &:= \sum_{k=2}^{n} (x_k - x_{k-1})^2 = x_1^2 + 2\sum_{k=2}^{n-1} x_k^2 + x_n^2 + 2\sum_{k=2}^{n} x_{k-1}x_k\,, \\ S_1 &:= \sum_{k=2}^{n} (x_k - x_{k-1})(Y_k - Y_{k-1}) \\ &= x_1 Y_1 + 2\sum_{k=2}^{n-1} x_k Y_k + x_n Y_n + \sum_{k=2}^{n} \bigl(x_{k-1}Y_k + x_k Y_{k-1}\bigr) \end{aligned}$$

and compute

$$\mathbf{Z}^T\mathbf{Z} = \begin{bmatrix} 1 & x_1 \\ x_1 & x_1^2 + Q_1^2 \end{bmatrix}, \qquad \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} = \frac{1}{Q_1} \begin{bmatrix} x_1^2 + Q_1 & -x_1 \\ -x_1 & 1 \end{bmatrix},$$

$$\mathbf{Z}^T\boldsymbol{U} = \begin{bmatrix} Y_1 \\ x_1 Y_1 + S_1 \end{bmatrix}.$$

By Gauss-Markov theorem the BLUE for the parameter $\boldsymbol{\gamma}$ *given by*

$$\hat{\boldsymbol{\gamma}} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\boldsymbol{U} = \frac{1}{Q_1} \begin{bmatrix} Q_1 Y_1 - x_1 S_1 \\ S_1 \end{bmatrix},$$

The best unbiased linear estimators for α *and* β *are:*

$$\hat{\alpha} = Y_1 - \frac{x_1 S_1}{Q_1}\,, \qquad \hat{\beta} = \frac{S_1}{Q_1}\,.$$

b. (5) Find explicitly the standard errors of the best unbiased linear estimates of α and β.

Solution: from

$$\mathrm{var}(Y_k - Y_{k-1}) = \mathrm{var}(\epsilon_k) = \sigma^2 \,; \quad k = 2, 3, \ldots, n$$

we get

$$\mathrm{var}(\hat{\beta}) = \frac{\sigma^2}{Q_1^2} \sum_{k=2}^{n} (x_k - x_{k-1})^2 = \frac{\sigma^2}{Q_1} \,.$$

Note that $\hat{\alpha} = Y_1 - x_1 \hat{\beta}$. *The random variables* Y_1 *and* $\hat{\beta}$ *are independent because* Y_1 *depends on* ϵ_1 *only, and* $\hat{\beta}$ *on* $\epsilon_2, \ldots, \epsilon_n$ *only. We have*

$$\mathrm{var}(\hat{\alpha}) = \mathrm{var}(Y_1) + x_1^2 \, \mathrm{var}(\hat{\beta}) = \sigma^2 \left(1 + \frac{x_1^2}{Q_1}\right) \,.$$

c. (5) Suggest an unbiased estimator of σ^2.

Solution: we use the transformed model and known unbiased estimates for the standard regression model. We have

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2} \|\boldsymbol{U} - \mathbf{Z}\hat{\boldsymbol{\gamma}}\|^2 \\
&= \frac{1}{n-2} \big(\boldsymbol{U} - \mathbf{Z}\hat{\boldsymbol{\gamma}}\big)^T \big(\boldsymbol{U} - \mathbf{Z}\hat{\boldsymbol{\gamma}}\big) \\
&= \frac{1}{n-2} \big(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\gamma}}\big)^T \boldsymbol{\Sigma}^{-1} \big(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\gamma}}\big) \\
&= \frac{1}{n-2} \left[(Y_1 - \hat{\alpha} - \hat{\beta} x_1)^2 + \sum_{k=2}^{n} \big(Y_k - Y_{k-1} - \hat{\beta}(x_k - x_{k-1})\big)^2 \right] \\
&= \frac{1}{n-2} \sum_{k=2}^{n} \left(Y_k - Y_{k-1} - \frac{S_1}{Q_1}(x_k - x_{k-1}) \right)^2 \,.
\end{aligned}$$

d. (5) Let $\tilde{\alpha}$ and $\tilde{\beta}$ be ordinary least squares estimators of the parameters α and β. Show that the estimators are unbiased and find their standard errors explicitly.

Solution: the two estimators form the vector $\tilde{\boldsymbol{\gamma}} = \begin{bmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y} = \boldsymbol{\gamma} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\eta}$. *As* $E(\boldsymbol{\eta}) = 0$ *we have* $E(\tilde{\boldsymbol{\gamma}}) = \boldsymbol{\gamma}$, *so the estimators are unbiased.*

The standard errors are best expressed with matrices. We ned the diagonal elements of the matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$. *But a direct approach is quicker. Define*

$$S_x := \sum_{k=1}^{n} x_k \,, \qquad S_{xx} := \sum_{k=1}^{n} x_k^2 \,, \qquad S_{xY} := \sum_{k=1}^{n} x_k Y_k \,, \qquad \Delta := n\, S_{xx} - S_x^2 \,,$$

We have

$$\tilde{\alpha} = \frac{S_{xx}S_Y - S_xS_{xY}}{\Delta} = \frac{1}{\Delta}\sum_{k=1}^{n}(S_{xx} - S_x x_k)Y_k\,,$$
$$\tilde{\beta} = \frac{n\,S_{xY} - S_xS_Y}{\Delta} = \frac{1}{\Delta}\sum_{k=1}^{n}(n\,x_k - S_x)Y_k\,.$$

The random variables $U_1, U_2, \ldots, U_n$ *are independent so*

$$\tilde{\alpha} = \frac{1}{\Delta}\sum_{k=1}^{n}\sum_{l=1}^{k}(S_{xx} - S_x x_k)U_l = \frac{1}{\Delta}\sum_{l=1}^{n}\sum_{k=l}^{n}(S_{xx} - S_x x_k)U_l\,,$$
$$\tilde{\beta} = \frac{1}{\Delta}\sum_{k=1}^{n}\sum_{l=1}^{k}(nx_k - S_x)U_l = \frac{1}{\Delta}\sum_{l=1}^{n}\sum_{k=l}^{n}(nx_k - S_x)U_l$$

The variances are

$$\begin{aligned}
\mathrm{var}(\tilde{\alpha}) &= \frac{\sigma^2}{\Delta^2}\sum_{l=1}^{n}\left(\sum_{k=l}^{n}(S_{xx} - S_x x_k)\right)^2 = \frac{\sigma^2}{\Delta^2}\sum_{l=1}^{n}\sum_{j=l}^{n}\sum_{k=l}^{n}(S_{xx} - S_x x_j)(S_{xx} - S_x x_k) \\
&= \frac{\sigma^2}{\Delta^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\min\{j,k\}(S_{xx} - S_x x_j)(S_{xx} - S_x x_k)\,, \\
\mathrm{var}(\tilde{\beta}) &= \frac{\sigma^2}{\Delta^2}\sum_{l=1}^{n}\left(\sum_{k=l}^{n}(nx_k - S_x)\right)^2 = \frac{\sigma^2}{\Delta^2}\sum_{l=1}^{n}\sum_{j=l}^{n}\sum_{k=l}^{n}(nx_j - S_x)(nx_k - S_x) \\
&= \frac{\sigma^2}{\Delta^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\min\{j,k\}(nx_j - S_x)(nx_k - S_x)\,.
\end{aligned}$$

Standard errors are obtained by taking square roots.