STAT 543/641 – WINTER 2019 – HOMEWORK #1

DUE FEBRUARY 11, 2019

(1) Let P be a disribution on $\mathbb{R}$ with variance σ^2 Let $X_1, \ldots, X_n$ be a random sample from P and let S^2 be the associated unbiased estimator of σ^2:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Show that

$$\operatorname{Var} S^2 = \frac{2\sigma^4}{n-1}.$$

Feel free to "cheat" and use the fact that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

(Can you do it without "cheating"?)

(2) (a) Let $\widetilde{x}$ be the median of $x_1, \ldots, x_n$, n odd. Prove that the identity

$$\sum_{i=1}^{n} |x_i - z| = \min_{y \in \mathbb{R}} \sum_{i=1}^{n} |x_i - y|$$

holds if and only if $z = \widetilde{x}$.

(b) Let $X_1, \ldots, X_n$ be a random sample from $\mathscr{L}(\mu, b)$, where $\mathscr{L}(\mu, b)$ is the *Laplace distribution* with density

$$f(x|\mu, b) = \frac{1}{2b^2} e^{-|x-\mu|/b}.$$

Assuming that b is known and that n is odd, Show that the MLE of μ is the sample median, $\widetilde{X}$. (Hint: Use (a).)

(3) [2, Exercise 7.1.3] Let $Y_1 < Y_2 < Y_3$ be the order statistics of a random sample of size three drawn from the uniform distribution having density function

$$f(x|\theta) = \begin{cases} 1/\theta & \text{if } 0 < x < \theta \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$. Show that $4Y_1$, $2Y_2$, and $\frac{4}{3}Y_3$ are all unbiased estimators of θ. Find the variance of each of these estimators.

(4) Suppose that

$$(X, Y) \sim N((\mu_X, \mu_Y), \Sigma), \quad \text{where} \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

(a) Write down the conditional density of Y given X.

(b) Show that $\mathrm{E}[Y|X]$ is has the form $a+bX$. Express a and b in terms of μ_X, μ_Y, σ_X, σ_Y, and ρ. (Hint: Use (a).)

(c) Confirm your answer to (b) experimentally by finding the least-squares line for data sampled from a bivariate normal distribution with randomly generated mean and covariance matrix.

(5) Let $x_0, x_1, \ldots, x_n \in \mathbb{R}$, let $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_n$ be independent normally distributed random variables with common mean 0 and common variance σ^2, and suppose

$$Y_i = a + bx_i + \varepsilon_i, \quad i = 0, 1, \ldots, n.$$

Recall our notation:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}), \quad S_{xY} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})$$

Let $\widehat{b}$, $\widehat{a}$, and $\widehat{\sigma}^2$ be the maximum likelihood estimators of b, a, and σ^2, respectively:

$$\begin{aligned}
\widehat{b} = \widehat{b}(Y_1, \ldots, Y_n) &= \frac{S_{xY}}{S_{xx}}, \\
\widehat{a} = \widehat{a}(Y_1, \ldots, Y_n) &= \overline{Y} - \widehat{b}\,\overline{x}, \\
\widehat{\sigma}^2 = \widehat{\sigma}^2(Y_1, \ldots, Y_n) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{a} - \widehat{b}x_i)^2.
\end{aligned}$$

Note that these expressions involve only the *training data* $(x_1, Y_1), \ldots, (x_n, Y_n)$. They omit the *test data* (x_0, Y_0).

The training error of our regression model is

$$\mathrm{MSE}_{\text{train}} = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - (\widehat{a} + \widehat{b}x_i)\right)^2\right],$$

while its test (prediction) error is

$$\mathrm{MSE}_{\text{test}} = \mathrm{E}\left[\left(Y_0 - (\widehat{a} + \widehat{b}x_0)\right)^2\right].$$

We know that

$$\mathrm{MSE}_{\text{train}} = \mathrm{E}\left[\widehat{\sigma}^2\right] = \frac{n-2}{n}\sigma^2.$$

In this exercise, we prove

$$\mathrm{MSE}_{\text{test}} = \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\sigma^2.$$

Note that

$$\mathrm{MSE}_{\text{train}} \leq \mathrm{MSE}_{\text{test}},$$

as one would expect (why?).

(a) Show that

$$\widehat{b} = \sum_{i=1}^{n} d_i Y_i \quad \text{and} \quad \widehat{a} = \sum_{i=1}^{n} c_i Y_i,$$

where

$$d_i = \frac{(x_i - \overline{x})}{S_{xx}} \quad \text{and} \quad c_i = \frac{1}{n} - \frac{\overline{x}(x_i - \overline{x})}{S_{xx}}.$$

(b) Prove that $\widehat{b}$ and $\widehat{a}$ are unbiased estimators of b and a, respectively. (Hint: Use (5a).)

(c) Establish the following identities:

$$\operatorname{Var}\widehat{b} = \frac{1}{S_{xx}}\sigma^2, \quad \operatorname{Var}\widehat{a} = \left(\frac{1}{nS_{xx}}\sum_{i=1}^{n} x_i^2\right)\sigma^2, \quad \operatorname{Cov}(\widehat{a}, \widehat{b}) = -\frac{\overline{x}}{S_{xx}}\sigma^2$$

(Hint: Use (5a) and the independence of $Y_1, \ldots, Y_n$.)

(d) What are the distributions of $\widehat{b}$ and $\widehat{a}$? (Hint: Use (5b) and (5c).)

(e) Establish the following identities:

$$\mathrm{E}[\widehat{a} + \widehat{b}x_0], \quad \operatorname{Var}(\widehat{a} + \widehat{b}x_0) = \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\sigma^2.$$

What is the distribution of $\widehat{a} + \widehat{b}x_0$? (Hint: For the variance, use (5c). The calculation is a bit tricky; if you get stuck, see [1, §11.3.5].)

(f) Prove that

$$\mathrm{E}\left[\left(Y_0 - \widehat{a} - \widehat{b}x_0\right)^2\right] = \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\sigma^2.$$

(Hint: Use the fact that Y_0 and $\widehat{a} + \widehat{b}x_0$ are independent (why?) and (5f).)

References

[1] Casella, Bergger, *Statistical Inference (2nd ed.)*, Duxbury, 2002.

[2] Hogg, McKean, Craig, *Introduction to Mathematical Statistics (7th ed.)*, Pearson, 2013.