18.650. Statistics for Applications
Fall 2016. Problem Set 8

Due Friday, Nov. 4 at 12 noon

Problem 1 Heteroscedastic regression

Let the characteristics $(\mathbf{X}_i, y_i)$ of n individuals ($i = 1, \ldots, n$) be observed, where $y_i \in \mathbb{R}$ is the dependent variable and $X_i \in \mathbb{R}^p$ is the vector of deterministic explanatory variables. Our goal is to estimate the coefficients of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ in the linear regression:

$$y_i = X_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n.$$

We assume that the model is heteroscedastic, i.e., the error terms ε_i are not i.i.d.. In this exercise, we are interested in the case where the vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is Gaussian, centered, with known covariance matrix Σ and we assume that Σ is invertible. We denote by $\mathbb{X}$ the matrix in $\mathbb{R}^{n \times p}$ whose rows are $X_1', \ldots, X_n'$ and by $\mathbb{Y}$ the vector with coordinates $y_1, \ldots, y_n$.

Consider the estimator $\hat{\beta}$ that minimises

$$(\mathbb{Y} - \mathbb{X}\boldsymbol{\beta})'\Sigma^{-1}(\mathbb{Y} - \mathbb{X}\boldsymbol{\beta}),$$

over $\boldsymbol{\beta} \in \mathbb{R}^p$.

1. Show that in the homoscedastic case, i.e., when $\Sigma = \sigma^2 I_n$ for some $\sigma^2 > 0$, $\hat{\boldsymbol{\beta}}$ reduces the least square error estimator.
2. Prove that $\hat{\boldsymbol{\beta}}$ is equal to the maximum likelihood estimator.
3. Propose a sufficient condition on the matrix $\mathbb{X}$ for $\hat{\boldsymbol{\beta}}$ to be uniquely defined.
4. From now on, we assume that the previous condition is satisfied. Compute $\hat{\boldsymbol{\beta}}$. What is the distribution of $\hat{\boldsymbol{\beta}}$?
5. Compute the bias and the quadratic risk of $\hat{\boldsymbol{\beta}}$.

Problem 2 Linear regression with random design

Consider n i.i.d. pairs of random variables $(\mathbb{X}_i, Y_i), i = 1, \ldots, n$, where $\mathbb{X}_i \in \mathbb{R}^p$ ($p \geq 1$) and $Y_i \in \mathbb{R}$. For each i, write

$$Y_i = \mathbb{X}_i'\boldsymbol{\beta} + \varepsilon_i,$$

where $\mathbb{E}[\varepsilon_i] = 0, cov(\mathbb{X}_i, \varepsilon_i) = 0$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector, that we want to estimate. In Questions 1,2 and 3, we assume that for all $x \in \mathbb{R}^p$, ε_1 has a conditional density given $X_1 = x$, denoted by f_x and that $\mathbb{X}_1$ has a density, which we denote by g.

1. Write the likelihood in terms of the unknown parameter $\boldsymbol{\beta}$, f_x and g.
2. Show that the maximum likelihood estimator of $\boldsymbol{\beta}$ does not depend on g, which may be unknown.
3. Assume that ε_1 is independent of $\mathbb{X}_1$ and that $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$.
   a) Compute f_x, for $x \in \mathbb{R}^p$.
   b) Since the $\mathbb{X}_i$'s are independent continuous random vectors of size p, it is possible to prove that the rank of the family $\{\mathbb{X}_1, \ldots, \mathbb{X}_n\}$ is equal to p almost surely. Here, we use this result without proving it.
   Show that the maximum likelihood estimator of $\boldsymbol{\beta}$ is equal to the least square error estimator and compute it.
   c) Conditionally on the $\mathbb{X}_i$'s, what is the distribution of the MLE ?
   d) Is the MLE biased ?
   Hint: First compute its expectation conditionally on the $\mathbb{X}_i$'s.
   e) What is the maximum likelihood estimator of σ^2 ?
   f) Propose an unbiased estimator $\hat{\sigma}^2$ of σ^2. What is the conditional distribution of $\dfrac{(n-p)\hat{\sigma}^2}{\sigma^2}$ given the X_i's ?
4. Assume that $p = 2$ and $\mathbb{X}_i = (1, X_i), i = 1, \ldots, n$, where X_i is a random variable with finite, non zero variance. Of course, we no longer assume that $\mathbb{X}_1$ has a density. Denote $\boldsymbol{\beta} = (a, b)$, so:

$$Y_i = a + bX_i + \varepsilon_i, \;\; i = 1, \ldots, n.$$

   a) Recall the least square estimator $(\hat{a}, \hat{b})$ of (a, b).
   b) Prove that it is consistent.
   c) Assume that X_1 and ε are independent, and denote by σ^2 the variance of ε_1. Show that $(\hat{a}, \hat{b})$ is asymptotically normal, and compute its asymptotic covariance matrix in terms of σ^2 and the moments of X_1.
   d) Propose a test with asymptotic level at most $\alpha \in (0, 1)$ for the null hypothesis H_0 : "$b > 0$" (the moments of X_1 and σ^2 are **not** known).

Problem 3 Logistic regression

Consider independent random pairs $(\mathbb{X}_1, Y_1), \ldots, (\mathbb{X}_n, Y_n)$, such that:

- $Y_i \in \{0, 1\}$ is a binary variable,
- $\mathbb{X}_i \in \mathbb{R}^p$,
- $\ln\left(\dfrac{\mathbb{P}[Y_i = 1|\mathbb{X}_i]}{\mathbb{P}[Y_i = 0|\mathbb{X}_i]}\right) = \mathbb{X}_i'\boldsymbol{\beta}$, for some $\boldsymbol{\beta} \in \mathbb{R}^p$.

For the sake of simplicity, we assume that $\mathbb{X}_1$ has a density, that is unknown. We denote it by f.

1. Compute $\mathbb{P}[Y_i = 1|\mathbb{X}_i]$ (for $i = 1, \ldots, n$).
2. Write the likelihood of the model in terms of $\boldsymbol{\beta}$ and f.
3. Show that the maximum likelihood estimator of $\boldsymbol{\beta}$ does not depend on the unknown density f.
   Remark: *In practice, there is no closed form for the maximum likelihood estimator, but there are some algorithms that allow to approach it.*