Problem Set 3, (Lectures 5 and 6)

Problem 1 (Identification, 50 points): Say that the parameters of a statistical model $\{P_\theta\}_{\theta \in \Theta}$ are *identified* if for any $\theta_1, \theta_2 \in \Theta$, $\theta_1 \neq \theta_2 \implies P_{\theta_1} \neq P_{\theta_2}$. Identification means that there are no two different members in the statistical model that yield the same distribution over the data.

1. (10 points) Show that the parameters (μ, σ^2) are identified in the model $X \sim N(\mu, \sigma^2)$.

2. (10 points) Show that the parameter p is identified in the model $X \sim \text{Bernoulli}(p)$.

3. (5 points) Show that θ_1 and θ_2 are not identified in the model $X \sim \mathcal{N}(\theta_1 + \theta_2, 1)$.

4. (5 easy points) The parameter θ is identified in the model $X \sim \mathcal{N}(\theta, 1)$ but the parameters (θ_1, θ_2) are not identified in the model $X \sim \mathcal{N}(\theta_1 + \theta_2, 1)$. Fix $\theta = \theta_0$ and define the "identified set" at θ_0 to be the values of (θ_1, θ_2) such that $P_{\theta_0} = P_{(\theta_1, \theta_2)}$. What is the identified set at $\theta_0 = 0$?

Problem 2 (Homoskedastic Linear Regression with Normal errors and Normal distribution of covariates, 50 points): Suppose we have a data set containing an outcome variable y_i and a vector of k controls $x_i = (x_{i1}, \ldots, x_{ik})'$ for n individuals. Assume that the controls are $N_k(\mathbf{0}, \Sigma)$ (where Σ is a known positive definite matrix) and let the outcome variable be modeled as

$$y_i = x_i'\beta + \epsilon_i,$$

where $\beta \in \mathbb{R}^k$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is assumed to be i.i.d. across individuals, and we treat σ^2 as known. If we collect the outcome variables in the $n \times 1$ vector

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

and the covariates in the $n \times k$ matrix

$$X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix},$$

A statistical model for (Y, X) can be described using

$$Y|X \sim \mathcal{N}(X\beta, \sigma^2 \mathbb{I}_n). \tag{0.1}$$

and X is an $n \times k$ matrix where each row is the transpose of an independent draw from a $N_k(0, \Sigma)$. Since Σ and σ^2 are both known, the parameter of this statistical model is $\beta \in \mathbb{R}^k$. The model in (0.1) is known as the Homoskedastic Linear Regression model with normal/Gaussian errors, known variance, and Gaussian distribution of covariates.

1. (Identification, 20 points) Is the parameter β identified? Does your answer depend on whether $n \geq k$?

2. (Statistical Sufficiency, 20 points) Let us define a *statistic* S as a mapping from the data D to some euclidean space $\mathbb{R}^p$. A statistic S is said to be sufficient for a parameter θ in a statistical model $\{P_\theta\}_{\theta \in \Theta}$ if the conditional distribution of the data, given the sufficient statistic, does not depend on θ. That is:

   $$\mathbb{P}_\theta(D|S(D)),$$

   does not depend on θ. Since, after conditioning on S, the distribution of the data does not depend any longer on θ, the statistic S is usually interpreted as carrying all the relevant information that the data has to give about θ. This typically means that having S is as good as having the whole data D.

   Suppose $n > k$ and $(X'X)$ is invertible for almost every data realization. Consider the $\mathbb{R}^p$ valued statistic

   $$S = (X'X)^{-1}X'Y,$$

   which is called the Ordinary Least Squares estimator of β in a linear regression model. Is it true that S is a sufficient statistic for β?