Problem Set 3, (Lectures 5 and 6)

Problem 1 (Identification, 30 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, 40 points): Suppose we have a data set containing an outcome variable y_i and a vector of k controls $x_i = (x_{i1}, \dots, x_{ik})'$ for n individuals. Assume that the controls are "fixed" (that is, they are treated as if they were non-random) 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 the response variables is

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

and the parameter of the model is $\beta \in \mathbb{R}^k$. The model in (0.1) is known as the Homoskedastic Linear Regression model with normal/Gaussian errors (and known variance).

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

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:

   $$\mathbb{P}_\theta(D|S(D)) = \mathbb{P}_{\theta'}(D|S(D)), \quad \text{for any } \theta, \theta'.$$

   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 assume $(X'X)$ is invertible. 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?

Problem 3 (A statistical problem with binary actions, binary data, two parameters, 30 points): Here is a problem to make you go through the concepts of decision problem, admissibility, and Bayes rules. This is going to look like a very stylized problem, but you will find this again in the context of hypothesis testing problems.

The data is binary $\{X_0, X_1\}$ (think of a coin flip) and the statistical model is

$$P(X_0|\theta_0) \equiv p_0 > p_1 \equiv P(X_0|\theta_1).$$

This means there are only two possible parameter values and that it is more likely to see X_0 realized whenever θ_0 generated the data. There are two actions (a_0, a_1). One way of thinking about them is that a_i is that action that θ_i generated the data. The loss function for this problem is

$$\mathcal{L}(a_0, \theta_0) = 0 = \mathcal{L}(a_1, \theta_1)$$

and

$$\mathcal{L}(a_1, \theta_0) = 1 = \mathcal{L}(a_0, \theta_1).$$

Which means that the statistician looses 1 unit if he supports action i when the data was not generated by θ_i.

1. (10 points) Decision rules are maps from $\{X_0, X_1\}$ to $\{a_0, a_1\}$. There are essentially 4 decision rules

$$\begin{aligned}
d_1(X_0) &= a_0, \quad & d_1(X_1) &= a_0 \text{ (always } a_0\text{)} \\
d_2(X_0) &= a_0, \quad & d_2(X_1) &= a_1 \text{ (}a_0 \text{ only if } X_0\text{)} \\
d_3(X_0) &= a_1, \quad & d_3(X_1) &= a_0 \text{ (}a_0 \text{ only if } X_1\text{)} \\
d_4(X_0) &= a_1, \quad & d_4(X_1) &= a_1 \text{ (}a_0 \text{ only if } X_1\text{)}
\end{aligned}$$

   Compute the risk function of each of these 4 decision rules and graph them as points in $\mathbb{R}^2$, $(R(d, \theta_0), R(d, \theta_1)$. Is it true that if $p_0 > .5 > p_1$ then d_3 is dominated?

2. (10 points) What priors would a Bayesian decision maker need to have in order to choose d_i, $i = 1, 2, 3, 4$?. Assume that $p_0 > .5 > p_1$.

3. (10 points) Imagine now that the action space becomes $[0, 1]$. Action a is interpreted as a randomized action: choose a_0 with probability a and a_1 with probability $1 - a$. Define

$$\mathcal{L}(a, \theta_i) = a\mathcal{L}(a_0, \theta_i) + (1 - a)\mathcal{L}(a_1, \theta_i).$$

   Consider an action of the form $d(X_0) = a'$ and $d(X_1) = a''$. What is the risk of the decision d? Is it true that if $p_0 > p_1$ the decision rule $d(X_0) = 0$ and $d(X_1) = 1$ is dominated?

   Optional: How would you plot the risk of all decision rules in $\mathbb{R}^2$? How does the set of all admissible decision rules look like?