Final exam

Due on June 23, 24:00

The problems are from sections 3.6, 4.7, 5.6, 10.7 of Tsiatis, A. (2006) Semiparametric Theory and Missing Data.

Problem 1

Let $Z_1, \ldots, Z_n$ be iid $p(z, \beta, \eta)$, where $\beta \in \mathbb{R}^q$ and $\eta \in \mathbb{R}^r$. Assume all the usual regularity conditions that allow the maximum likelihood estimator to be a solution to the score equation,

$$\sum_{i=1}^{n} \begin{pmatrix} S_\beta\left(Z_i, \beta, \eta\right) \\ S_\eta\left(Z_i, \beta, \eta\right) \end{pmatrix} = 0^{(q+r)\times 1},$$

and be consistent and asymptotically normal.

a) Show that the influence function for $\hat{\beta}_n$ is the efficient influence function.

b) Sketch out an argument that shows that the solution to the estimating equation

$$\sum_{i=1}^{n} S_{\text{eff}}^{q\times 1}\{Z_i, \beta, \hat{\eta}_n^*(\beta)\} = 0^{q\times 1},$$

for any root-n consistent estimator $\hat{\eta}_n^*(\beta)$, yields an estimator that is asymptotically linear with the efficient influence function.

Problem 2

Let Y be a one-dimensional response random variable. Consider the model

$$Y = \mu(X, \beta) + \varepsilon,$$

where $\beta \in \mathbb{R}^q$, and $E\{h(\varepsilon)|X\} = 0$ for some arbitrary function $h(\cdot)$. Up to now, we considered the identity function $h(\varepsilon) = \varepsilon$, but this can be generalized to arbitrary $h(\varepsilon)$. For example, if we define $h(\varepsilon) = \{I(\varepsilon \leq 0) - 1/2\}$, then this is the median regression model. That is, if we define $F(y|x) = P(Y \leq y|X = x)$, then med $(Y|x) = F^{-1}(1/2, x)$, the value $m(x)$ such that $F(m(x)|x) = 1/2$. Therefore, the model with this choice of $h(\cdot)$ is equivalent to

$$\text{med}\,(Y|X) = \mu(X, \beta).$$

Assume no other restrictions are placed on the model but $E\{h(\varepsilon)|X\} = 0$ for some function $h(\cdot)$. For simplicity, assume h is differentiable, but this can be generalized to nondifferentiable h such as in median regression.

a) Find the space $\Lambda^{\perp}$ (i.e., the space perpendicular to the nuisance tangent space).
b) Find the efficient score vector for this problem.
c) Describe how you would construct a locally efficient estimator for β from a sample of data $(Y_i, X_i), i = 1, \ldots, n$.
d) Find an estimator for the asymptotic variance of the estimator defined in part (c).

Problem 3

Heteroscedastic models
Consider the semiparametric model for which, for a one-dimensional response variable Y, we assume

$$Y = \mu(X, \beta) + V^{1/2}(X, \beta)\varepsilon, \;\; \beta \in \mathbb{R}^q,$$

where ε is an arbitrary continuous random variable such that ε is independent of X. To avoid identifiability problems, assume that for any scalars α, α'

$$\alpha + \mu(x, \beta) = \alpha' + \mu(x, \beta') \quad \text{for all} \quad x$$

implies

$$\alpha = \alpha' \quad \text{and} \quad \beta = \beta',$$

and for any scalars $\sigma, \sigma' > 0$ that

$$\sigma\{V(x, \beta)\} = \sigma'\{V(x, \beta')\} \quad \text{for all} \quad x$$

implies

$$\sigma = \sigma' \quad \text{and} \quad \beta = \beta'.$$

For this model, describe how you would derive a locally efficient estimator for β from a sample of data

$$(Y_i, X_i), i = 1, \ldots, n.$$

Problem 4

Consider the simple linear regression restricted moment model where with full data $(Y_i, X_{1i}, X_{2i}), i = 1, \ldots, n$, we assume

$$E(Y_i|X_{1i}, X_{2i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}.$$

In such a model, we can estimate the parameters $(\beta_0, \beta_1, \beta_2)^T$ using ordinary least squares; that is, the solution to the estimating equation

$$\sum_{i=1}^{n} (1, X_{1i}, X_{2i})^T (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}) = 0. \tag{10.104}$$

In fact, this estimator is locally efficient when $\text{var}(Y_i|X_{1i}, X_{2i})$ is constant. The data, however, are missing at random with a monotone missing pattern. That is, Y_i is observed on all individuals in the sample; however, for some individuals, only X_{2i} is missing, and for others both X_{1i} and X_{2i} are missing. Therefore, we define the missingness indicator

$$(\mathcal{C}_i = 1) \text{ if we only observe } Y_i,$$
$$(\mathcal{C}_i = 2) \text{ if we only observe } (Y_i, X_{1i}),$$

and

$$(\mathcal{C}_i = \infty) \text{ if we observe } (Y_i, X_{1i}, X_{2i}).$$

We will define the missingness probability model using discrete-time hazards, namely

$$\lambda_1(Y) = P(\mathcal{C} = 1|Y),$$
$$\lambda_2(Y, X_1) = P(\mathcal{C} = 2|\mathcal{C} \geq 2, Y, X_1).$$

a) In terms of λ_1 and λ_2, what is

$$P(\mathcal{C} = \infty|Y, X_1, X_2)?$$

In order to model the missingness process, we assume logistic regression models; namely,

$$\text{logit}\,\{\lambda_1(Y)\} = \psi_{10} + \psi_{11} Y, \text{where logit}\,(p) = \log\left(\frac{p}{1-p}\right),$$

and

$$\text{logit}\,\{\lambda_2(Y, X_1)\} = \psi_{20} + \psi_{21} X_1 + \psi_{22} Y.$$

b) Using some consistent notation to describe the observed data, write out the estimating equations that need to be solved to derive the maximum likelihood estimator for

$$\psi = (\psi_{10}, \psi_{11}, \psi_{20}, \psi_{21}, \psi_{22})^T.$$

c) Describe the linear subspace Λ_ψ.
d) Describe the linear subspace Λ_2. Verify that $\Lambda_\psi \subset \Lambda_2$.
e) Describe the subspace $\Lambda^\perp$, the linear space orthogonal to the observed-data nuisance tangent space. An initial estimator for β can be obtained by using an inverse probability weighted complete-case estimator that solves the equation

$$\sum_{i=1}^{n} \frac{I(\mathcal{C}_i = \infty)}{\varpi(\infty, Y_i, X_{1i}, X_{2i}, \hat{\psi}_n)} (1, X_{1i}, X_{2i})^T (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}) = 0,$$

where $\hat{\psi}_n$ is the maximum likelihood estimator derived in (b). Denote this estimator by $\hat{\beta}_n^I$.
f) What is the i-th influence function for $\hat{\beta}_n^I$?
g) Derive a consistent estimator for the asymptotic variance of $\hat{\beta}_n^I$.
In an attempt to gain efficiency, we consider

$$\frac{I(\mathcal{C}_i = \infty)}{\varpi(\infty, Y_i, X_{1i}, X_{2i}, \psi_o)} \varphi^{*F}(Y_i, X_{1i}, X_{2i}) - \Pi\left[\frac{I(\mathcal{C}_i = \infty)\varphi^{*F}(Y_{1i}, X_{1i}, X_{2i})}{\varpi(\infty, Y_i, X_{1i}, X_{2i}, \psi_o)} \middle| \Lambda_2 \right],$$

where $\varphi^{*F}(\cdot) = (1, X_{1i}, X_{2i})^T (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})$.
h) Compute

$$\Pi\left[\frac{I(\mathcal{C}_i = \infty)\varphi^{*F}(Y_{1i}, X_{1i}, X_{2i})}{\varpi(\infty, Y_i, X_{1i}, X_{2i}, \psi_o)} \middle| \Lambda_2 \right].$$

In practice, we need to estimate (h) using a simplifying model. For simplicity, let us use as a working model that $(Y_i, X_{1i}, X_{2i})^T$ is multivariate normal with mean $(\mu_Y, \mu_{X_1}, \mu_{X_2})^T$ and covariance matrix

$$\begin{bmatrix} \sigma_{YY} & \sigma_{YX_1} & \sigma_{YX_2} \\ \sigma_{YX_1} & \sigma_{X_1X_2} & \sigma_{X_1X_2} \\ \sigma_{YX_2} & \sigma_{X_1X_2} & \sigma_{X_2X_2} \end{bmatrix}.$$

i) With the observed data, how would you estimate the parameters in the multivariate normal?
j) Assuming the simplifying multivariate normal model and the estimates derived in (i), estimate the projection in (h).
k) Write out the estimating equation that needs to be solved to get an improved estimator.
l) Find a consistent estimator for the asymptotic variance of the estimator in (k). (Keep in mind that the simplifying model of multivariate normality may not be correct.)