Practice Final

Question 1

A researcher studies a latent variable generated by the population model

$$Y_i^\star = X_i\beta + U_i, \qquad U_i \mid X_i \sim \mathcal{N}(0, \sigma^2),$$

but the sampling scheme only allows observation of $Y_i^\star$ over a restricted region. Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal pdf and cdf.

(a) Suppose the data are *doubly truncated*: the pair (Y_i, X_i) is observed only when $a < Y_i^\star < b$, for known constants $a < b$. Derive the conditional mean $\mathbb{E}[Y_i \mid X_i,\, a < Y_i^\star < b]$. Express your answer in terms of ϕ and Φ evaluated at the standardized cut-points $\alpha_i = (a - X_i\beta)/\sigma$ and $\beta_i = (b - X_i\beta)/\sigma$, and comment on the two limiting cases $a \to -\infty$ and $b \to +\infty$.

(b) Now drop the upper truncation and let the lower threshold be *individual-specific and observed*: (Y_i, X_i, Z_i) enters the sample if and only if $Y_i^\star > c_i$, where $c_i = Z_i\pi$ with Z_i observed, $\pi \neq 0$, and Z_i independent of U_i. Derive $\mathbb{E}[Y_i \mid X_i, Z_i,\, Y_i^\star > c_i]$ and write the implied regression in the additive form $Y_i = X_i\beta + \sigma\,\lambda(a_i) + e_i$, identifying a_i and the inverse Mills ratio $\lambda(\cdot)$, and stating $\mathbb{E}[e_i \mid X_i, Z_i, Y_i^\star > c_i]$.

(c) Derive the maximum likelihood estimator of the truncated regression model under the single-threshold scheme of (b). That is, derive the conditional density of Y_i given (X_i, Z_i) in the truncated sample, write down the log-likelihood, and state the maximization problem that defines the MLE $(\hat\beta, \hat\sigma)$. Then suppose the researcher estimates the *standard* truncated model, incorrectly imposing a fixed, known constant threshold c in place of the true $c_i = Z_i\pi$ (with $\pi \neq 0$). By examining the population score of the misspecified likelihood evaluated at the true (β, σ), show that the resulting MLE is inconsistent for β in general. Identify the source of the inconsistency, and state the condition on $\mathrm{Cov}(Z_i, X_i)$ under which consistency is restored.

(d) Using the representation in (b), show that as $c_i \to -\infty$ for all i the selection-correction term vanishes and the OLS estimator of β (regressing Y_i on X_i in the observed sample) becomes consistent. Make precise what happens to $\lambda(a_i)$, to the score of the model, and hence to the OLS normal equations in this limit.

(e) Suppose instead the data are *censored* at zero (Tobit): $Y_i = \max\{0, Y_i^\star\}$ and X_i is always observed. Show that

$$\mathbb{E}[Y_i \mid X_i] = \Phi\left(\frac{X_i\beta}{\sigma}\right) X_i\beta + \sigma\,\phi\left(\frac{X_i\beta}{\sigma}\right),$$

indicating where you use the symmetry of the standard normal, and derive the partial effect $\partial\,\mathbb{E}[Y_i \mid X_i]/\partial X_{ij}$, showing it equals $\Phi\left(\frac{X_i\beta}{\sigma}\right)\beta_j$.

(f) For the Tobit model in (e), define the indicator $\ell(Y_i) = \mathbf{1}\{Y_i > 0\}$. Write down the log-likelihood, explain why Y_i has a mixed (continuous plus discrete) distribution and identify the size of the probability mass at zero. Then explain why OLS of Y_i on X_i using the full sample is inconsistent for β, and contrast this with the linear-model case where the conditional mean is correctly specified.