International School of Economics at TSU
Econometrics II
Lasha Chochua

Problem Set 6

Instructions: You are encouraged to solve the problems before the recitation. Additionally, you are encouraged to work in groups. It is **not mandatory** to submit solutions unless stated otherwise. However, if you would like to share your solution, I would be happy to review it.

Problem 1 (Truncated Normal Moment). Let $Y^* \sim N(\alpha, \sigma_Y^2)$. Prove that

$$\mathbb{E}[(Y^* - \alpha)\, Y^* \mid Y^* > 0] = \sigma_Y^2.$$

You may use without proof the standard truncated normal formulas: for $W \sim N(\mu, \sigma^2)$ with $c = -\mu/\sigma$,

$$\mathbb{E}[W \mid W > 0] = \mu + \sigma\lambda(c), \qquad \mathrm{Var}(W \mid W > 0) = \sigma^2[1 - \delta(c)],$$

where $\lambda(c) = \phi(c)/\Phi(-c)$ and $\delta(c) = \lambda(c)[\lambda(c) - c]$. Explain in one sentence why this identity matters for deriving OLS bias under censoring.

Problem 2 (Tobit Model). Consider the Tobit model $Y^* = \beta_0 + \beta_1 X + e$, $e \mid X \sim N(0, \sigma^2)$, $Y = \max(Y^*, 0)$. You estimate the model on $n = 1{,}000$ households and obtain $\hat{\beta}_0 = -5$, $\hat{\beta}_1 = 2$, $\hat{\sigma} = 4$, where X denotes years of education.

(a) Compute the predicted censoring probability at $X = 3$.

(b) Compute the predicted censored mean $\hat{m}(3)$ and verify that $\hat{m}(3) \geq \hat{\beta}_0 + \hat{\beta}_1 \cdot 3$.

(c) An OLS regression on the full sample gives $\hat{\beta}_1^{\text{OLS}} = 0.7$ with estimated censoring proportion $\hat{\pi} = 0.65$. Is this consistent with Greene's formula? What does it imply about the reliability of OLS here?

Problem 3 (Selection Bias and Heckman). A researcher wants to estimate the returns to education using observed wages. Wages are only observed for employed workers; employment itself is a choice.

(a) Write down Heckman's two-equation model (outcome equation and selection equation) and state the joint normality assumption. Define σ_{21} and explain its role.

(b) Derive the selected-sample CEF $\mathbb{E}[Y \mid X, Z, S = 1]$ and explain why OLS on the selected sample is biased when $\sigma_{21} \neq 0$.

(c) Describe the Heckit two-step estimator. Why must standard errors be corrected after step two?

(d) Propose an exclusion restriction for this wage application and explain why it must satisfy two conditions.

Problem 4 (CQR – Identification and Estimation). Consider the censored quantile regression (CQR) model of Powell (1986):

$$Y^* = X'\beta + e, \qquad Q_\tau[e \mid X] = 0, \qquad Y = \max(Y^*, 0).$$

(a) Using the equivariance property of quantiles, show that $Q_\tau[Y \mid X] = \max(X'\beta, 0)$.

(b) For what values of X is $Q_\tau[Y \mid X]$ identified from the censored data? What goes wrong for observations where $X'\beta \leq 0$?

(c) Write down the CQR criterion function $M_n(\beta; \tau)$ using the check function $\rho_\tau(u) = u(\tau - \mathbf{1}\{u < 0\})$. What special case does $\tau = 0.5$ reduce to?

(d) State one key computational difficulty with CQR that does not arise with Tobit, and explain how practitioners typically deal with it.

Problem 5 (CLAD). Consider the censored regression model

$$Y^* = \beta_0 + \beta_1 X + e, \qquad \text{med}[e \mid X] = 0, \qquad Y = \max(Y^*, 0).$$

(a) State what $\text{med}[Y \mid X]$ equals under this model and justify your answer using the equivariance property of the median.

(b) Write down the CLAD criterion function $M_n(\beta)$ and explain intuitively why minimizing it recovers β consistently even when the Tobit normality assumption fails.

(c) You have five observations: $(X_i, Y_i) = (1, 0), (2, 3), (3, 0), (4, 7), (5, 4)$. For $\hat{\beta}_0 = 0$, compare M_n at $\hat{\beta}_1 = 1$ vs. $\hat{\beta}_1 = 2$ and identify which gives the lower criterion value.

Problem 6 (Quantile Regression by Hand). You observe $n = 5$ observations:

i	X_i	Y_i
1	1	2
2	2	1
3	3	5
4	4	4
5	5	7

Consider the linear quantile regression model with no intercept: $Q_\tau[Y \mid X] = \beta X$.

(a) Write down the check function $\rho_\tau(u)$ and explain asymmetry: why does $\tau = 0.9$ push the fitted line upward relative to $\tau = 0.1$?

(b) For $\tau = 0.5$ (median regression), compute the LAD criterion $M_n(\beta) = \frac{1}{5}\sum_{i=1}^{5} |Y_i - \beta X_i|$ at $\beta = 1$ and $\beta = 1.5$. Which gives a lower value?

(c) For $\tau = 0.9$, compute the check-function criterion $M_n(\beta; 0.9)$ at $\beta = 1$ and $\beta = 1.5$. Which gives a lower value, and does the ranking flip relative to part (b)? Explain why or why not.

(d) At the optimal $\hat{\beta}$, what fraction of observations should lie below the fitted line $\hat{\beta}X_i$? Verify this approximately for the better candidate in part (b).

Problem 7 (CLAD by Hand). You observe $n = 5$ observations:

i	X_i	Y_i
1	1	0
2	2	0
3	3	2
4	4	5
5	5	6

Assume the model $Y^* = \beta X + e$, $\text{med}[e \mid X] = 0$, $Y = \max(Y^*, 0)$, with no intercept.

(a) Two observations are censored ($Y_i = 0$). For each, compute the predicted latent value $\hat{Y}_i^* = \beta X_i$ at $\beta = 0.8$ and $\beta = 1.2$. Are these consistent with $Y^* \leq 0$ having occurred?

(b) Write down the CLAD criterion $M_n(\beta) = \frac{1}{5}\sum_{i=1}^{5} |Y_i - \max(\beta X_i, 0)|$ and evaluate it at $\beta = 0.8$ and $\beta = 1.2$.

(c) Which β gives the lower criterion value? Compare to what a naive LAD (ignoring censoring, fitting $Y_i = \beta X_i$ directly) would give at the same two candidate values. Does ignoring censoring bias the estimate upward or downward here?

(d) Explain why CLAD is consistent even if $e \mid X$ is heteroskedastic, whereas Tobit MLE would not be.