Final Exam: Suggestive Solutions

Part I

Repeat of Project #2

Part 1: Marginal Effects and 'Munkit'

(1) Derive the conditional probabilities of Y^* being censored and uncensored (from below, at zero), respectively, conditional on $X = x$, and relate these probabilities to the probabilities $\mathrm{P}(Y = 0|\, X = x)$ and $\mathrm{P}(Y > 0|\, X = x)$ pertaining to the outcome Y.
Solution: The probability of being censored is

$$\begin{aligned}
\mathrm{P}\left(Y^* \leqslant 0|\, X = x\right) &= \mathrm{P}\left(\beta_0 X + \sigma_0 \varepsilon \leqslant 0|\, X = x\right) \\
&= \mathrm{P}\left(\varepsilon \leqslant -\beta_0 X/\sigma_0|\, X = x\right) \\
&= \mathrm{P}\left(\varepsilon \leqslant -\beta_0 x/\sigma_0|\, X = x\right) \\
&= \mathrm{P}\left(\varepsilon \leqslant -\beta_0 x/\sigma_0\right) \qquad (\varepsilon \text{ and } X \text{ independent}) \\
&= G\left(-\beta_0 x/\sigma_0\right).
\end{aligned}$$

By the law of total probability,

$$\begin{aligned}
\mathrm{P}\left(Y^* > 0|\, X = x\right) &= 1 - \mathrm{P}\left(Y^* \leqslant 0|\, X = x\right) \\
&= 1 - G\left(-\beta_0 x/\sigma_0\right).
\end{aligned}$$

Since $Y = \max\{0, Y^*\}$, we must have

$$\mathrm{P}\left(Y = 0|\, X = x\right) = \mathrm{P}\left(Y^* \leqslant 0|\, X = x\right) = G\left(-\beta_0 x/\sigma_0\right)$$

and

$$\mathrm{P}\left(Y > 0|\, X = x\right) = \mathrm{P}\left(Y^* > 0|\, X = x\right) = 1 - G\left(-\beta_0 x/\sigma_0\right).$$

(2) Derive the CDF $F_{Y|X}(\cdot|\, x)$ of Y conditional on $X = x$ and comment on the nature of $F_{Y|X}(\cdot|\, x)$.
Solution: The CDF is defined by

$$F_{Y|X}\left(y|\, x\right) := \mathrm{P}\left(Y \leqslant y|\, X = x\right).$$

Y is nonnegative, so $F_{Y|X}(y|\,x) = 0$ for $y < 0$. Since $\{Y = 0\}$ and $\{Y > 0\}$ are complements, for $y \geqslant 0$ we have

$$\begin{aligned}
F_{Y|X}(y|\,x) &= \mathrm{P}(Y \leqslant y \cap Y = 0|\,X = x) + \mathrm{P}(Y \leqslant y \cap Y > 0|\,X = x) \\
&= \mathrm{P}(Y = 0|\,X = x) + \mathrm{P}(0 < Y \leqslant y|\,X = x) \\
&= G(-\beta_0 x/\sigma_0) + [G((y - \beta_0 x)/\sigma_0) - G(-\beta_0 x/\sigma_0)] \\
&= G((y - \beta_0 x)/\sigma_0).
\end{aligned}$$

Hence,

$$F_{Y|X}(y|\,x) = \begin{cases} 0, & y < 0, \\ G((y - \beta_0 x)/\sigma_0), & y \geqslant 0. \end{cases}$$

The CDF $F_{Y|X}(y|\,x)$ is flat up to zero, jump discontinuous at $y = 0$ with jump equal to the censoring probability, and continuous (in fact continuously differentiable) for $y > 0$.

(3) Derive the likelihood contribution function of the ith observation and define the maximum likelihood estimator of $\boldsymbol{\theta}_0 := (\beta_0, \sigma_0)$ based on $\{(Y_i, X_i)\}_1^n$.

Solution: The CDF $F_{Y|X}(y|\,x)$ is flat up to zero, jump discontinuous at $y = 0$ with jump equal to the censoring probability, and is differentiable for $y > 0$, so the conditional outcome density is

$$f_{Y|X}(y|\,x) = \begin{cases} G(-\beta_0 x/\sigma_0), & y = 0, \\ g((y - \beta_0 x)/\sigma_0)/\sigma_0, & y > 0, \end{cases}$$

which we may write as

$$f_{Y|X}(y|\,x) = G\left(-\frac{\beta_0 x}{\sigma_0}\right)^{\mathbf{1}(y=0)} \left[\frac{1}{\sigma_0} g\left(\frac{y - \beta_0 x}{\sigma_0}\right)\right]^{\mathbf{1}(y>0)}.$$

The likelihood contribution of observation i as a function of $\boldsymbol{\theta} := (\beta, \sigma) \in \mathbf{R} \times \mathbf{R}_{++}$ is therefore

$$\ell_i(\boldsymbol{\theta}) := G\left(-\frac{\beta X_i}{\sigma}\right)^{\mathbf{1}(Y_i=0)} \left[\frac{1}{\sigma} g\left(\frac{Y_i - \beta X_i}{\sigma}\right)\right]^{\mathbf{1}(Y_i>0)},$$

The MLE is then any maximizer $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta} \mapsto \sum_{i=1}^n \ln \ell_i(\boldsymbol{\theta})$.

(4) Show that $\mathrm{E}[Y|\,X = x] = \beta_0 x [1 - G(-\beta_0 x/\sigma_0)] + \sigma_0 \int_{-\beta_0 x/\sigma_0}^{\infty} t g(t)\,\mathrm{d}t$.

Solution: The claim follows from the calculation

$$\begin{aligned}
\mathrm{E}\left[Y \mid X=x\right] &= \mathrm{E}\left[\max\left\{0, Y^*\right\} \mid X=x\right] \\
&= \mathrm{E}\left[\max\left\{0, \beta_0 X + \sigma_0\varepsilon\right\} \mid X=x\right] \\
&= \mathrm{E}\left[\left(\beta_0 X + \sigma_0\varepsilon\right) \mathbf{1}\left(\beta_0 X + \sigma_0\varepsilon \geqslant 0\right) \mid X=x\right] \\
&= \mathrm{E}\left[\left(\beta_0 x + \sigma_0\varepsilon\right) \mathbf{1}\left(\varepsilon \geqslant -\beta_0 x/\sigma_0\right) \mid X=x\right] \\
&= \beta_0 x \mathrm{E}\left[\mathbf{1}\left(\varepsilon \geqslant -\beta_0 x/\sigma_0\right) \mid X=x\right] + \sigma_0 \mathrm{E}\left[\varepsilon \mathbf{1}\left(\varepsilon \geqslant -\beta_0 x/\sigma_0\right) \mid X=x\right] \\
&= \beta_0 x \mathrm{E}\left[\mathbf{1}\left(\varepsilon \geqslant -\beta_0 x/\sigma_0\right)\right] + \sigma_0 \mathrm{E}\left[\varepsilon \mathbf{1}\left(\varepsilon \geqslant -\beta_0 x/\sigma_0\right)\right] \\
&\qquad\qquad (\varepsilon \text{ and } X \text{ independent}) \\
&= \beta_0 x \mathrm{P}\left(\varepsilon \geqslant -\beta_0 x/\sigma_0\right) + \sigma_0 \mathrm{E}\left[\varepsilon \mathbf{1}\left(\varepsilon \geqslant -\beta_0 x/\sigma_0\right)\right] \\
&= \beta_0 x \left[1 - G\left(-\beta_0 x/\sigma_0\right)\right] + \sigma_0 \int_{-\beta_0 x/\sigma_0}^{\infty} t g\left(t\right) \mathrm{d}t. \qquad (G \text{ continuous})
\end{aligned}$$

(5) Derive an expression for the marginal effect $\mathrm{ME}\left(x\right) := \left(\mathrm{d}/\mathrm{d}x\right) \mathrm{E}\left[Y \mid X=x\right]$ of X on the conditional mean of Y at x and comment on its dependence on x.

Solution: Since

$$\mathrm{E}\left[Y \mid X=x\right] = \beta_0 x \left[1 - G\left(-\beta_0 x/\sigma_0\right)\right] + \sigma_0 \int_{-\beta_0 x/\sigma_0}^{\infty} t g\left(t\right) \mathrm{d}t,$$

the marginal effect is

$$\frac{\mathrm{d}}{\mathrm{d}x} \mathrm{E}\left[Y \mid X=x\right] = \beta_0 \frac{\mathrm{d}}{\mathrm{d}x} x \left[1 - G\left(-\beta_0 x/\sigma_0\right)\right] + \sigma_0 \frac{\mathrm{d}}{\mathrm{d}x} \int_{-\beta_0 x/\sigma_0}^{\infty} t g\left(t\right) \mathrm{d}t.$$

By the product and chain rules

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x} x \left[1 - G\left(-\beta_0 x/\sigma_0\right)\right] &= 1 \cdot \left[1 - G\left(-\beta_0 x/\sigma_0\right)\right] + x \left[-G'\left(-\beta_0 x/\sigma_0\right)\left(-\beta_0/\sigma_0\right)\right] \\
&= 1 - G\left(-\beta_0 x/\sigma_0\right) + \left(\beta_0 x/\sigma_0\right) g\left(-\beta_0 x/\sigma_0\right).
\end{aligned}$$

Applying Leibniz rule [with $a(x) = -\beta_0 x/\sigma_0$ and $b(x)$ constant in x],

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x} \int_{-\beta_0 x/\sigma_0}^{\infty} t g\left(t\right) \mathrm{d}t &= 0 - \left(-\beta_0 x/\sigma_0\right) g\left(-\beta_0 x/\sigma_0\right)\left(-\beta_0/\sigma_0\right) + 0 \\
&= -\left(\beta_0^2 x/\sigma_0^2\right) g\left(-\beta_0 x/\sigma_0\right).
\end{aligned}$$

It follows that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{E}\left[Y|\, X=x\right] &= \beta_0\left[1-G\left(-\beta_0 x/\sigma_0\right)+\left(\beta_0 x/\sigma_0\right)g\left(-\beta_0 x/\sigma_0\right)\right]-\sigma_0\left(\beta_0^2 x/\sigma_0^2\right)g\left(-\beta_0 x/\sigma_0\right)\\ &= \beta_0\left[1-G\left(-\beta_0 x/\sigma_0\right)\right]+\left(\beta_0^2 x/\sigma_0\right)g\left(-\beta_0 x/\sigma_0\right)-\left(\beta_0^2 x/\sigma_0\right)g\left(-\beta_0 x/\sigma_0\right)\\ &= \beta_0\left[1-G\left(-\beta_0 x/\sigma_0\right)\right].\end{aligned}$$

[Strictly speaking, Leibniz rule does not apply without further justification with one or more limits at infinity. While it is possible to place sufficient conditions on g for Leibniz rule to apply even in our setting (e.g. g vanishing rapidly enough at $\pm\infty$), keeping in line with the textbook, we here simply proceed *as if* g is sufficiently regular.]

(6) Evaluate the claim: "Censoring leads to a reduction of the marginal effect of X relative to its marginal effect on the latent outcome."
Solution: The conditional expectation function of the latent outcome is

$$\begin{aligned}\mathrm{E}\left[Y^*|\, X=x\right] &= \mathrm{E}\left[\beta_0 X+\sigma_0\varepsilon|\, X=x\right]\\ &= \beta_0 x+\sigma_0\mathrm{E}\left[\varepsilon|\, X=x\right]\\ &= \beta_0 x+\sigma_0\mathrm{E}\left[\varepsilon\right]. \qquad (\varepsilon \text{ and } X \text{ independent})\end{aligned}$$

so the marginal effect on the latent outcome is

$$\mathrm{ME}^*\left(x\right)\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{E}\left[Y^*|\, X=x\right]=\beta_0,$$

a constant. Since $g>0$ everywhere, we must have $0<G<1$ everywhere. Hence, contrasting $\mathrm{ME}^*\left(x\right)$ with our previous finding of

$$\mathrm{ME}\left(x\right)=\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{E}\left[Y|\, X=x\right]=\beta_0\underbrace{\left[1-G\left(-\beta_0 x/\sigma_0\right)\right]}_{\in(0,1)},$$

we see that censoring indeed leads to an attenuation of the marginal effect of X relative to its marginal effect on the latent outcome. That is, the claim is true.

(7) Suppose that you have already established consistency of $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$ as $n\to\infty$. Suggest a consistent estimator $\widehat{\mathrm{ME}}\left(x\right)$ of the marginal effect $\mathrm{ME}\left(x\right)$ and argue its consistency at some point x.
Solution: The marginal effect is

$$\mathrm{ME}\left(x\right)=\beta_0\left[1-G\left(-\beta_0 x/\sigma_0\right)\right]$$

which we may view as the (nonlinear) function

$$h(\boldsymbol{\theta}) := \beta\left[1 - G(-\beta x/\sigma)\right]$$

at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. A natural estimator is the plug-in estimator

$$\widehat{\mathrm{ME}}(x) := h(\widehat{\boldsymbol{\theta}}) = \widehat{\beta}[1 - G(-\widehat{\beta}x/\widehat{\sigma})].$$

Given that both $\widehat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_0$ and h is continuous at $\boldsymbol{\theta}_0$, the continuous mapping theorem applies to show

$$\widehat{\mathrm{ME}}(x) = h(\widehat{\boldsymbol{\theta}}) \xrightarrow{p} h(\boldsymbol{\theta}_0) = \mathrm{ME}(x).$$

(8) Suppose now that you have already established that $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d \mathrm{N}(\mathbf{0}, \mathbf{V}_0)$ as $n \to \infty$ for some 2×2 variance matrix $\mathbf{V}_0$. What is the asymptotic distribution of the estimator $\widehat{\mathrm{ME}}(x)$ from your answer to the previous question?

Solution: The marginal effect is

$$\mathrm{ME}(x) = \beta_0\left[1 - G(-\beta_0 x/\sigma_0)\right]$$

which we may view as the (nonlinear) function

$$h(\boldsymbol{\theta}) := \beta\left[1 - G(-\beta x/\sigma)\right]$$

at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Differentiation shows that

$$\begin{aligned} \frac{\partial}{\partial\beta} h(\boldsymbol{\theta}) &= 1 \cdot \left[1 - G(-\beta x/\sigma)\right] + \beta\left[-G'(-\beta x/\sigma)(-x/\sigma)\right] \\ &= 1 - G(-\beta x/\sigma) + (\beta x/\sigma)\, g(-\beta x/\sigma), \end{aligned}$$

and

$$\begin{aligned} \frac{\partial}{\partial\sigma} h(\boldsymbol{\theta}) &= \beta\left[-G'(-\beta x/\sigma)(-\beta x)\left(-1/\sigma^2\right)\right] \\ &= -\left(\beta^2 x/\sigma^2\right) g(-\beta x/\sigma), \end{aligned}$$

so h is differentiable at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ with gradient $\nabla h(\boldsymbol{\theta}_0)$ given by

$$\nabla h(\boldsymbol{\theta}_0) = \left[\begin{array}{cc} 1 - G(-\beta_0 x/\sigma_0) + (\beta_0 x/\sigma_0)\, g(-\beta_0 x/\sigma_0), & -(\beta_0^2 x/\sigma_0^2)\, g(-\beta_0 x/\sigma_0) \end{array}\right].$$

(We have here used $\sigma_0 \in \mathbf{R}_{++}$.) Continuing with the estimator

$$\widehat{\mathrm{ME}}\,(x) := h(\widehat{\boldsymbol{\theta}}) = \widehat{\beta}[1 - G(-\widehat{\beta}x/\widehat{\sigma})],$$

it then follows from the delta method that

$$\sqrt{n}[\widehat{\mathrm{ME}}\,(x) - \mathrm{ME}\,(x)] \xrightarrow{d} \mathrm{N}\left(0, \nabla h\left(\boldsymbol{\theta}_0\right) \mathbf{V}_0 \nabla h\left(\boldsymbol{\theta}_0\right)'\right) \text{ as } n \to \infty.$$

(9) Discuss the components necessary to construct a 95% confidence interval for $\mathrm{ME}\,(x)$ and argue in what sense it is valid.

Solution: The previous part shows

$$\sqrt{n}[\widehat{\mathrm{ME}}\,(x) - \mathrm{ME}\,(x)] \xrightarrow{d} \mathrm{N}\left(0, v_0^2\right),$$

where

$$v_0^2 := \nabla h\left(\boldsymbol{\theta}_0\right) \mathbf{V}_0 \nabla h\left(\boldsymbol{\theta}_0\right)'.$$

An asymptotically valid 95% confidence interval for $\mathrm{ME}\,(x)$ therefore arises from

$$\widehat{\mathrm{ME}}\,(x) \pm 1.96 \frac{\widehat{v}}{\sqrt{n}},$$

where $\widehat{v}^2$ is any consistent estimator of v_0^2. To consistently estimate v_0^2 it suffices to consistently estimate $\mathbf{V}_0$ and $\nabla h\left(\boldsymbol{\theta}_0\right)$ and setting

$$\widehat{v}^2 := \widehat{\nabla h\left(\boldsymbol{\theta}_0\right)} \widehat{\mathbf{V}} \widehat{\nabla h\left(\boldsymbol{\theta}_0\right)}'$$

(cf. the continuous mapping theorem). A natural estimator of $\nabla h\left(\boldsymbol{\theta}_0\right)$ is the plug-in estimator $\widehat{\nabla h\left(\boldsymbol{\theta}_0\right)} := \nabla h(\widehat{\boldsymbol{\theta}})$, whose consistency follows from continuity of ∇h at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, which, in turn, follows from the continuity of g.

Part 2: COVID-19 and Temperature

(1) Pick an estimation sample and a set of additional covariates, $\mathbf{x}_{it}$, and justify your decision. You should keep this fixed throughout the rest of the questions.

- Students should discuss exogeneity of temperature, and regardless of whether additional controls are included or not, they should comment on whether they are also likely to be exogenous (for example, stringency may increase when deaths rise).

- Ultimately, the students should comment on this, but ultimately it should not be judged as better or worse to choose one way or the other.
- The estimates should be easily comparable in a table format.
- The coefficients should not be compared (although their sign can be), but the marginal effects should be. Differences should be related

(2) Estimate models of $\mathrm{E}(y_{it}|z_{it}, \mathbf{x}_{it})$ using respectively a Tobit model, and a Poisson regression model (see Cameron & Trivedi, 2005, ch. 5.2.1 and 20.2.1). Focus your comparison on the marginal effect of z_{it}.

- Even if the model did not converge, nearly full credit will be given if sufficient attempts were made to resolve these issues (e.g. switching the optimizer, trying different starting values, etc.) and demonstrate good command of numerical optimization.
- Marginal effects are required: Tobit is a special case of Part 1, which should be realized and used, and Poisson needs to be derived (or numerical differentiation used).
- Standard errors of the marginal effects are computed using the delta method, which were derived for the more general model in part 1, but for the Poisson model, the derivations are new.

(3) Assess the fit of the two models first in terms of $\mathrm{E}(y_{it}|z_{it}, \mathbf{x}_{it})$, and then in terms of other features of the distribution. Which model is the most suitable for understanding the development in Denmark?

- The R^2 is an obvious starting point for a comparison of model fit, but it is not the only or the best criterion always.
- The Poisson distribution has only a single parameter, implying a tight restriction on the shape of the distribution. For example, this implies that $\Pr(y = 0|\mathbf{x})$ may be lower than $\Pr(y = 1|\mathbf{x})$ if $\mathbf{x}'\boldsymbol{\beta}$ is large enough.
- Conversely, the Tobit implies a censored normal distribution for $y|\mathbf{x}$. So there is almost always lower mass for any $y > 0$ relative to the mass point at $y = 0$. This mass point in the histogram of y is present almost regardless of where in time or space you evaluate the distribution of y. In that particular sense, the Tobit model may be a better fit.

- A graphical illustration comparing is probably ideal, but it is also possible to compare $\Pr(y = 0|\mathbf{x})$ directly from the two models.

(4) Assess the robustness of your estimated marginal effects from the Tobit model with respect to the assumed distribution for the error term.

- Estimation of a non-Gaussian Tobit should be based on the derivations from part 1. The chosen distribution for the error term should be well-motivated (e.g. not discrete), and the student should compare the distribution to the normal and comment on pros and cons.
- It can be fine to estimate CLAD as a comparison, although the model assumptions (and objects of interest) are different so parameters cannot necessarily be compared directly. Moreover, marginal effects are complicated. CLAD is e.g. robust to heteroskedasticity.

(5) Is the effect of temperature on COVID-19 deaths constant across countries and over time? Are some countries likely to see sharper increases in fatalities over the coming months?

- This question is related to heterogenous marginal effects, which the students should note.
- Marginal effects change depending on the overall level of $\mathbf{x}'\beta$.
- But the model can be made more flexible by including interaction effects, which allows for a richer discussion.

Part II

New Assignment

1 Ordered Choice

(1) Derive the conditional probability of $Y = -1, 0$ and 1, respectively, given $X = x$.

Solution: Observe that

$$\begin{aligned}
\mathrm{P}(Y=-1|X=x) &= \mathrm{P}(Y^* \leqslant -a|X=x) \\
&= \mathrm{P}(\beta_0 X + \varepsilon \leqslant -a|X=x) \\
&= \mathrm{P}(\varepsilon \leqslant -(a+\beta_0 X)\,|X=x) \\
&= \mathrm{P}(\varepsilon \leqslant -(a+\beta_0 x)\,|X=x) \\
&= \mathrm{P}(\varepsilon \leqslant -(a+\beta_0 x)) && (\varepsilon \text{ and } X \text{ independent}) \\
&= G(-(a+\beta_0 x)) && (G \text{ CDF of } \varepsilon)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{P}(Y=0|X=x) &= \mathrm{P}(-a < Y^* < b|X=x) \\
&= \mathrm{P}(-a < \beta_0 X + \varepsilon < b|X=x) \\
&= \mathrm{P}(-(a+\beta_0 X) < \varepsilon < b - \beta_0 X|X=x) \\
&= \mathrm{P}(-(a+\beta_0 x) < \varepsilon < b - \beta_0 x|X=x) \\
&= \mathrm{P}(-(a+\beta_0 x) < \varepsilon < b - \beta_0 x) && (\text{independence}) \\
&= \mathrm{P}(\varepsilon < b - \beta_0 x) - \mathrm{P}(\varepsilon \leqslant -(a+\beta_0 x)) \\
&= G(b-\beta_0 x) - G(-(a+\beta_0 x)),
\end{aligned}$$

where the last step utilizes continuity of G to swap weak and strict inequalities. Thus, by the law of total probability,

$$\begin{aligned}
\mathrm{P}(Y=1|X=x) &= 1 - \mathrm{P}(Y \neq 1|X=x) \\
&= 1 - [\mathrm{P}(Y=-1|X=x) + \mathrm{P}(Y=0|X=x)] && (\text{trinary choice}) \\
&= 1 - [G(-(a+\beta_0 x)) + G(b-\beta_0 x) - G(-(a+\beta_0 x))] \\
&= 1 - G(b-\beta_0 x).
\end{aligned}$$

(2) Derive the likelihood contribution function of the ith observation and define the maximum likelihood estimator of β_0 based on $\{(Y_i, X_i)\}_1^n$.

Solution: Let $f(y|x)$ denote the conditional PDF of Y given $X = x$. Then, by our previous calculations,

$$f(y|x) = \begin{cases} G(-(a+\beta_0 x)), & y = -1, \\ G(b-\beta_0 x) - G(-(a+\beta_0 x)), & y = 0, \\ 1 - G(b - \beta_0 x), & y = 1. \end{cases}$$

It follows that

$$\ell_i(\beta) = \begin{cases} G(-(a+\beta X_i)), & Y_i = -1, \\ G(b-\beta X_i) - G(-(a+\beta X_i)), & Y_i = 0, \\ 1 - G(b - \beta X_i), & Y_i = 1, \end{cases}$$

for $\beta \in \mathbf{R}$, or, equivalently,

$$\begin{aligned} \ell_i(\beta) &= G(-(a+\beta X_i))^{\mathbf{1}(Y_i=-1)} \\ &\quad \times [G(b-\beta X_i) - G(-(a+\beta X_i))]^{\mathbf{1}(Y_i=0)} \\ &\quad \times [1 - G(b-\beta X_i)]^{\mathbf{1}(Y_i=1)}. \end{aligned}$$

The MLE $\widehat{\beta}$ is any maximizer of $\mathbf{R} \ni \beta \mapsto \sum_{i=1}^n \ln \ell_i(\beta)$.

(3) Derive the conditional mean of Y (not Y^*) given $X = x$.

Solution: By our previous calculations,

$$\begin{aligned} \mathrm{E}[Y|\, X = x] &= \sum_{y \in \{-1,0,1\}} y \mathrm{P}(Y = y | X = x) \qquad \text{(trinary choice)} \\ &= (-1)\,\mathrm{P}(Y = -1|X = x) + (0)\,\mathrm{P}(Y = 0|X = x) + (1)\,\mathrm{P}(Y = 1|X = x) \\ &= \mathrm{P}(Y = 1|X = x) - \mathrm{P}(Y = -1|X = x) \\ &= 1 - G(b - \beta_0 x) - G(-(a + \beta_0 x)). \end{aligned}$$

(4) Derive an expression for the marginal effect $\mathrm{ME}(x) := (\mathrm{d}/\mathrm{d}x)\,\mathrm{E}[Y|\, X = x]$ of X on the conditional mean of Y at x and comment on its dependence on x.

Solution: Here

$$\begin{aligned}
\mathrm{ME}(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\mathrm{E}[Y|X=x] \\
&= \frac{\mathrm{d}}{\mathrm{d}x}\left[1 - G(b-\beta_0 x) - G(-(a+\beta_0 x))\right] \\
&= -\frac{\mathrm{d}}{\mathrm{d}x}\left[G(b-\beta_0 x) + G(-(a+\beta_0 x))\right] \\
&= -\left[g(b-\beta_0 x)(-\beta_0) + g(-(a+\beta_0 x))(-\beta_0)\right] \\
&= \left[g(b-\beta_0 x) + g(-(a+\beta_0 x))\right]\beta_0,
\end{aligned}$$

which depends on x in a nonlinear manner, in general.

(5) Evaluate the claim: "Discretization leads to a change in sign of the marginal effect of X relative to its marginal effect on the latent outcome."
Solution: The conditional expectation function of the latent outcome is

$$\begin{aligned}
\mathrm{E}[Y^*|X=x] &= \mathrm{E}[\beta_0 X + \varepsilon | X=x] \\
&= \beta_0 x + \mathrm{E}[\varepsilon | X=x] \\
&= \beta_0 x + \mathrm{E}[\varepsilon], && (\varepsilon \text{ and } X \text{ independent})
\end{aligned}$$

so the marginal effect on the latent outcome is

$$\mathrm{ME}^*(x) = \frac{\mathrm{d}}{\mathrm{d}x}\mathrm{E}[Y^*|X=x] = \beta_0,$$

a constant. Since $g > 0$ everywhere, we must have

$$\begin{aligned}
\mathrm{sign}(\mathrm{ME}(x)) &= \mathrm{sign}\left(\left[g(b-\beta_0 x) + g(-(a+\beta_0 x))\right]\beta_0\right) \\
&= \mathrm{sign}(\beta_0) \\
&= \mathrm{sign}(\mathrm{ME}^*(x)),
\end{aligned}$$

where $\mathrm{sign}(\cdot)$ is the sign function

$$\mathrm{sign}(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0. \end{cases}$$

Hence the claim is <u>false</u>.

(6) Suppose that you have already established consistency of $\widehat{\beta}$ for β_0 as $n \to \infty$. Suggest a consistent estimator $\widehat{\mathrm{ME}}(x)$ of the marginal effect $\mathrm{ME}(x)$ and argue its consistency at any point x.

Solution: The marginal effect is

$$\mathrm{ME}(x) = [g(b - \beta_0 x) + g(-(a + \beta_0 x))]\,\beta_0$$

which we may view as the (nonlinear) function h defined by

$$h(\beta) := [g(b - \beta x) + g(-(a + \beta x))]\,\beta,$$

at the point $\beta = \beta_0$. A natural estimator is the plug-in estimator

$$\widehat{\mathrm{ME}}(x) := h(\widehat{\beta}) = [(b - \widehat{\beta}x) + g(-(a + \widehat{\beta}x))]\widehat{\beta}.$$

Given that both $\widehat{\beta}$ is consistent for β_0 and h is continuous at β_0 (using g continuous), the continuous mapping theorem applies to show

$$\widehat{\mathrm{ME}}(x) = h(\widehat{\beta}) \xrightarrow{p} h(\beta_0) = \mathrm{ME}(x).$$

(7) Suppose now that you have already established that $\sqrt{n}(\widehat{\beta} - \beta_0) \to_d \mathrm{N}(0, \sigma_0^2)$ as $n \to \infty$ for some (not necessarily known) variance $\sigma_0^2 \in \mathbf{R}_{++}$. What is the asymptotic distribution of the estimator $\widehat{\mathrm{ME}}(x)$ (appropriately centered and scaled) from your answer to (6)?

Solution: The marginal effect is

$$\mathrm{ME}(x) = [g(b - \beta_0 x) + g(-(a + \beta_0 x))]\,\beta_0$$

which we may view as the (nonlinear) function

$$h(\beta) := [g(b - \beta x) + g(-(a + \beta x))]\,\beta,$$

at $\beta = \beta_0$. Using g differentiable, differentiation of the right-hand side with respect to

β shows that

$$\begin{aligned}
h'(\beta) &= \left[g'(b-\beta x)(-x) + g'(-(a+\beta x))(-x)\right]\beta \\
&\quad + \left[g(b-\beta x) + g(-(a+\beta x))\right](1) \qquad \text{(product and chain rules)} \\
&= \left[g'(b-\beta x) + g'(-(a+\beta x))\right](-\beta x) \\
&\quad + \left[g(b-\beta x) + g(-(a+\beta x))\right], \qquad (1)
\end{aligned}$$

so h is differentiable at $\beta = \beta_0$ with derivative $h'(\beta_0)$ given by

$$\begin{aligned}
h'(\beta_0) &= \left[g'(b-\beta_0 x) + g'(-(a+\beta_0 x))\right](-\beta_0 x) \\
&\quad + \left[g(b-\beta_0 x) + g(-(a+\beta_0 x))\right]. \qquad (2)
\end{aligned}$$

Continuing with the estimator

$$\widehat{\mathrm{ME}}(x) := h(\widehat{\beta}) = [(b-\widehat{\beta}x) + g(-(a+\widehat{\beta}x))]\widehat{\beta},$$

it then follows from the delta method that

$$\sqrt{n}[\widehat{\mathrm{ME}}(x) - \mathrm{ME}(x)] \xrightarrow{d} \mathrm{N}\left(0, \left[h'(\beta_0)\right]^2\sigma_0^2\right) \text{ as } n \to \infty,$$

with $h'(\beta_0)$ given in (2).

(8) Continuing with the setup of (7), construct a 95% asymptotically valid (but not necessarily feasible) confidence interval for $\mathrm{ME}(x)$. What, if any, additional quantities do you need in order to make this confidence interval feasible in practice?
Solution: The previous part shows

$$\sqrt{n}[\widehat{\mathrm{ME}}(x) - \mathrm{ME}(x)] \xrightarrow{d} \mathrm{N}\left(0, v_0^2\right),$$

where

$$v_0^2 := \left[h'(\beta_0)\right]^2\sigma_0^2$$

and $h'(\beta_0)$ given in (2). An asymptotically (as $n \to \infty$) valid 95% confidence interval for $\mathrm{ME}(x)$ therefore arises from

$$\widehat{\mathrm{ME}}(x) \pm 1.96\frac{\widehat{v}}{\sqrt{n}},$$

where $\widehat{v}^2$ is any consistent estimator of v_0^2. To consistently estimate v_0^2 it suffices to consistently estimate σ_0^2 and $h'(\beta_0)$ and setting

$$\widehat{v}^2 := [\widehat{h'(\beta_0)}]^2\widehat{\sigma}^2$$

(cf. the continuous mapping theorem). A natural estimator of $h'(\beta_0)$ is the plug-in estimator $\widehat{h'(\beta_0)} := h'(\widehat{\beta})$ with h' given in (1), whose consistency follows from continuity of h' at $\beta = \beta_0$, which, in turn, follows from g being continuously differentiable.

2 Heteroskedastic Tobit

The numbers below are found using `fminunc` with default settings.

(1) The OLS estimates are $\hat{\boldsymbol{\beta}} = (.75, -.14)$ (se $= (.15, .10)$), and the Tobit estimates are $(-.47, .12)$ (se $= (.33, .21)$).

(1) The linear model is

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

and requires $\mathbb{E}(\varepsilon_i \mathbf{x}_i) = \mathbf{0}$ for consistency of OLS. The default standard errors require ε_i IID.

(2) The Tobit model is

$$y_i = \max\{\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, 0\},$$

and requires $\varepsilon_i \sim$ IID $\mathrm{N}(0, \sigma^2)$. The default standard errors are valid under this assumption

(2) A quantile regression for $\tau = 0.90$ reveals a coefficient on x_{i2} of $-.87$, implying a negative relationship. The same can be confirmed by plotting the 90th percentile within, say, 10 bins of x_{i2}. Similarly, the share of observations with $y_i > 0$ can be computed within 10 such bins to show a positive relationship. Alternatively, one could compute $\tilde{y}_i := \mathbf{1}_{\{y_i>0\}}$ and plot the results of a kernel regression estimator of $\tilde{y}_i$ on x_2 for a grid over $[1; 2]$.

(1) For computing the quantile regression estimator, minimization can be done using either a Newton-based (`fminunc`) or gradient-free (`fminsearch`) optimizer, although the latter is generally preferred as the criterion function is not smooth in finite samples.

(2) For quantile regression, standard errors are *not* required.

(3) The log-likelihood contribution for general h is derived precisely as it is for regular Tobit with the only difference that we normalize by σ_i rather than some homogeneous coefficient σ. This does not change derivations because conditional on $\mathbf{x}_i$, we have independence across cross-sectional observations.

$$\begin{aligned}\ell_i(\boldsymbol{\beta},\gamma) &= \mathbf{1}_{\{y_i>0\}}\left[-\frac{1}{2}\log(2\pi\sigma_i^2)-\frac{1}{2\sigma_i^2}(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2\right] \\ &\quad +\mathbf{1}_{\{y_i=0\}}\log\left[1-\Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma_i}\right)\right], \\ \text{where } \sigma_i &\equiv \gamma h(\mathbf{x}_i'\boldsymbol{\beta})\end{aligned}$$

and the log-likelihood function is $\mathcal{L}(\boldsymbol{\beta},\gamma)=\frac{1}{N}\sum_{i=1}^N \ell_i(\boldsymbol{\beta},\gamma)$.

(4) Using $h(z)=\exp(z)$ gives an average likelihood of 1.20, whereas the model with $h(z)=\exp(-z)$ gives a likelihood of 1.18. Based on this, $\exp(-z)$ gives a better fit of the data as measured by the likelihood of observing the data given the model. (And $h(z)=\exp(-z)$ is indeed the true specification). The two models give opposite implications

(5) Clearly, the results in (4) come from the correct DGP, so that estimate of $\boldsymbol{\beta}$ should be our preferred estimate.

Regarding question (2):

(1) Note that h is decreasing in x_{i2}. Hence, for small values of x_{i2}, σ_i is large, and therefore we tend to primarily observe $y_i>0$ due to large draws of the error term.

(2) Conversely, when x_{i2} is large (close to 2), y_i is more often positive because y_i^* itself is simply larger there.

(3) In conclusion, for low values of $\mathbf{x}_i'\boldsymbol{\beta}$, heteroskedasticity increases the share of observations with $y_i>0$. Conversely, when $\mathbf{x}_i'\boldsymbol{\beta}>0$, heteroskedasticity would instead reduce the share.

(4) It is important to note that we cannot alone judge which of the two h-functions is appropriate based on (2) alone.

Regarding question (1):

(1) OLS found the wrong sign on x_{i2}, whereas (homoskedastic) Tobit found a much too small (and insignificant) estimate. The negative estimate for OLS is due to a combination of censoring and the large "outliers" occurring for x_{i2} close to the

lower bound of the support, where σ_i is very large and draws are therefore very large.