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

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

By the law of total probability,

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

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

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

and

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

(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

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

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

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

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

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

the marginal effect is

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

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

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

[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

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

so the marginal effect on the latent outcome is

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

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

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

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

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

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

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

which we may view as the (nonlinear) function

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

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

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

and

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

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

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

(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

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

(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

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

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 $\nabla h\,(\boldsymbol{\theta}_0)$ and setting

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

(cf. the continuous mapping theorem). A natural estimator of $\nabla h\,(\boldsymbol{\theta}_0)$ is the plug-in estimator $\widehat{\nabla h\,(\boldsymbol{\theta}_0)} := \nabla h(\widehat{\boldsymbol{\theta}})$, 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.

- Regarding the **choice of variables** to include, the good student should note that: Including stringency runs the risk of introducing reverse causality (governments shut down in response to high prevalence). Similarly, mobility is directly restricted

by stringency and thus runs the same risk. Many variables have zero variation over time within a country: e.g. diabetes prevalence, share of elderly, GDP, etc. Such variables cannot be included with fixed effects. However, if it is not feasible (for computational reasons) to use country fixed effects in non-linear models, then these time-invariant variables may capture the most important variation in country fixed effects (this can actually be examined!).

  - It is not important what variables are chosen, but the arguments behind the choices should be logical.

- Regarding **choice of sample**: it may be problematic to include developing and developed countries together as temperature may have a different effect (although that can be accounted for with an interaction by a very good student). However, it is important to include countries at different latitudes in order to have variation in temperature at the same month of year: if e.g. one only includes Europe, then there is a lot of disease early on and the temperature is also low, but it is because the pandemic started there. By including countries both on the northern and southern hemisphere, there are opposite winter/summer months and thus more variation in temperature within months.

  - If one includes only European countries, it is actually possible to get a positive estimate of temperature. This goes away when more countries are included, or if one focuses only on Denmark, say. It might be expected that the students have experimented with their choice of countries and had found a suitable set where the results make sense.

- (Technically speaking, the problem text does not ask for OLS estimates, so the students do not have to provide them... but the hint suggests it and it is a preferred way of exploring the effects of different choices of regressors, sample, fixed effects, etc.)

- Ideally, the results should be related to the coming work. Due to computational constraints, it may not be feasible to work with country (or time) fixed effects in the non-linear models, so the regression results can be used to guide whether one should be concerned about their omission. For example, one could compare a model with country fixed effects to one that just has continent fixed effects, which may be more tractable to use in the non-linear models.

(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 the student should realize and use, and Poisson needs to be derived (or numerical differentiation used).
- Standard errors of the marginal effects should be 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 and should be provided.

(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 (e.g. with calendar time on the x-axis) 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 not taking only positive values), and the student should compare the distribution to the normal and comment on pros and cons. The astute student should verify that the distribution satisfies the requirements used in the derivation.

- If the distribution contains additional parameters (like the degrees of freedom for the Student's t distribution), then some reasoning for the choice of df should be provided. If the df for example are estimated, the student should argue carefully why that is valid to do and whether it is even a continuous variable.
- It would be excellent to estimate a model which allows for heteroskedasticity and censoring:
  - CLAD: Very robust, but numerically difficult to work with. 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, but variance estimation does *not* follow standard M-estimator theory (bootstrapping can be used).
  - Alternatively, one can estimate the heteroskedastic Tobit from the lecture slides (the one which is also in the exam project, not that students would have had time to complete that).

(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?

- The clearest way of addressing the question is to allow for interaction effects between temperature and covariates. For example, one could interact a dummy for "spring" with temperature, implying a different effect of temperature in the first months and in the last months. The relationship is probably clearest from summer when protective measures were enacted. Another example is to interact GDP, population density, or health indicators (e.g. share of smokers) with temperature. One might imagine that poor countries are hit harder when the cold months come.
- Alternatively, one can estimate the model on subsamples of the data (e.g. only including certain countries or sub-periods). This has two primary disadvantages: fewer observations (i.e. loss in efficiency), and a loss in variation in temperature. For example, if one chooses only Scandinavian countries in the summer there is very little variation in temperature. However, it may be an advantage to try to estimate without the earliest part of the sample where the pandemic was first spreading as behavior might have been quite different then.
- Another approach is to look at the marginal effect of temperature: since the models are non-linear, they depend on the values of all other regressors. Hence, if the country is fully locked down, it might be able to reduce deaths completely

even though it is very cold. This discussion should emphasize some other regressor than temperature (e.g. a dummy for a continent with fewer deaths in general, like Oceania) and look at the marginal effects evaluated in countries that have very high risk from that variable and compare it to a country that does not, showing how the effect of temperature depends on the chosen regressor.

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

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

and

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

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

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

(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 $f\left(y|x\right)$ denote the conditional PDF of Y given $X=x$. Then, by our previous calculations,

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

It follows that

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

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

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

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

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

Solution: By our previous calculations,

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

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

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

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

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

so the marginal effect on the latent outcome is

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 1: Estimates

	OLS	Tobit	Quantile	$h(z) = \exp(-z)$	$h(z) = \exp(z)$
β_1	.751	-.472	3.257	-1.557	.221
	(.15)	(.33)	n.a.	(.22)	(.17)
β_2	-.142	.129	-.868	.836	-.303
	(.10)	(.21)	n.a.	(.13)	(.10)
σ or γ	n.a.	1.695	n.a.	1.225	2.105
	n.a.	(.07)	n.a.	(.07)	(.20)
$\mathcal{L}(\theta)$		-1.2089		-1.1886	-1.2026
N	1000	1000	1000	1000	1000

where

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

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 $h'\left(\beta_0\right)$ and setting

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

(cf. the continuous mapping theorem). A natural estimator of $h'\left(\beta_0\right)$ is the plug-in estimator $\widehat{h'\left(\beta_0\right)} := h'(\widehat{\beta})$ 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 estimates are shown in Table 1.

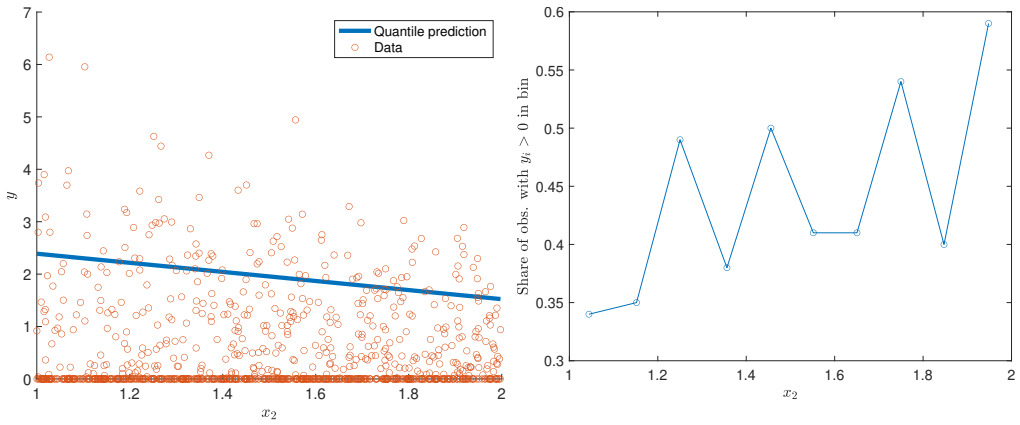
(1) The linear model is

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

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

Figure 1: Part 2, question 2: descriptive pictures

(a) Quantile prediction, $\tau = 0.9$ (b) Share of non-censored obs. in bins



Note: Panel (a) shows the predicted values from a quantile regression at $\tau = 0.90$. Panel (b) shows the fraction of observations with $y_i > 0$ within deciles of x_{i2}. Each dot has x value equal to the average x_{i2} in that bin and y equal to the fraction of non-censored.

(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 M-estimator standard errors are valid under this assumption, and the Sandwich formula collapses (due to the information matrix equality), so that either the outer product of the scores or the Hessian can be used as the basis for the covariance matrix estimator.

(3) If the error term is heteroskedastic, then OLS is still consistent, but robust standard errors must be used; Tobit, on the other hand, is inconsistent (and hence inference is not interesting). If the true model has censoring, then OLS is additionally inconsistent (attenuation bias, i.e. bias towards zero of the slope coefficient).

(2) A quantile regression for $\tau = 0.90$ reveals a coefficient on x_{i2} of $-.87$ (see Table 1), 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 (the excellent student will consider various numbers of bins and realize that there is a bias-variance tradeoff in this choice, similarly to the bandwidth choice in kernel regression). 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 $[1; 2]$. See Figure 1.

(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 but the students should know that the usual M-estimator asymptotics are *invalid* (the criterion function is not smooth), but bootstrap methods can be used.

(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) The estimates from the two models are shown in Table 1. Two ways of arguing this:

(1) **Likelihood:** 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).

(2) **Intuitively:** With $h(z) = \exp(-z)$, heteroskedasticity is decreasing in x_{i2}, while the latent index is increasing y_i^*. Conversely, with $h(z) = \exp(z)$, the estimates imply that both heteroskedasticity and the latent index are decreasing in x_{i2}. Both specifications have decreasing heteroskedasticity, which is consistent with the quantile regression showing that the 90th percentile of the distribution $y_i|x_{i2}$ is decreasing in x_{i2}. However, the finding from earlier that the share of observations with $y_i > 0$ is increasing in x_{i2} can only be explained by the latent index increasing. The astute student may note the fact that the top decile bin has 59% observations non-censored – since the error term is symmetric, there will only be more than 50% non-censored (asymptotically) if the latent index is positive at that value of x_{i2}.

(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.