Name: Bulat Gafarov

Student ID:

Midterm ARE256b W2022, February 7, 2022

- You have 90 minutes.
- The maximum number of points is 100.
- Put your name and student ID number on every sheet of paper.
- Do not write outside of the boxes. We will only grade what you write inside the boxes.

Name:

Student ID:

---

1. **(30 points total) Short theory questions**

   (a) (10) Discuss random utility interpretation of Probit or Logit models following the exposition in the class. You should explicitly define all components of the corresponding probabilistic model to get the full credit. You can use a decision to work as an example.

Individual decides between option 0 and 1 by comparing the difference between the corresponding utilities:

$\Delta u_i = \alpha + \beta x_i + \varepsilon_i$

x_i - regressors

ε_i is a utility shock that is distributed as $N(0,1)$ for Probit model and Logistic distribution for logit.

$P(y_i=1|x_i) = \Phi(\alpha+\beta x_i)$ for logit.

"$P(\Delta u_i \geq 0 | x_i)$" $= P(\varepsilon_i \geq -\alpha - \beta x_i | x_i)$

Example: Suppose that $y_i = 1$ if the person decides to work, $y_i = 0$ otherwise.

Suppose x_i is the wage.

ε_i - other factors that affect the decision.

Name: Bulat Gafarov

Student ID:

(b) (10) Suppose that you are dealing with a censored dependent variable, that is you only observe $Y^* = \alpha + \beta X + u$ only if $Y^* > 0$ (u is a zero mean r.v., independent of X). We don't observe data is if $Y^* \leq 0$. Explain briefly why the standard OLS estimator is inconsistent for β.

$y_i^* = \alpha + \beta x + u$.

$E(y_i^* | x, y_i^* \geq 0) = \alpha + \beta x + E(u_i | y_i^* \geq 0, x_i)$

Term $E(u_i | y_i \geq 0, x_i)$ is not 0 and depend on x_i. Hence a linear regression of y_i^* on x_i will have omitted variable bias from that term, $E(u_i | y_i^* \geq 0, x_i)$.

Name:

Student ID:

(c) (10) Discuss properties of maximum likelihood estimator in the linear regression model with homoscedastic errors that has Gaussian (i.e. normal) distribution. Explain why it remains consistent even if the errors are heteroscedastic?

$y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i | x_i \sim N(0, \sigma_\varepsilon^2)$.
MLE estimator coincides with the OLS estimator. By Gauss-Markov theorem, $\hat{\alpha}^{MLE}, \hat{\beta}^{MLE}$ are consistent, unbiased and efficient (has lowest variance)

If ε_i is heteroscedastic, OLS is still consistent. Hence MLE remains consistent as well (by equivalence with OLS)

Name: Boulat Gataerov

Student ID:

---

2. **(40 points total) Marginal effects in non-linear models.** Suppose that Y_i is a binary variable equal to 1 if individual i got hospitalized for with a specific disease, and equal to 0 if not. It is well known that the hospitalization risk grows with the age of the individual. A newly developed vaccine can reduce the risk of hospitalization. Our goal is to write a statistical model to estimate effectiveness of the vaccine, i.e. its effect on the chance of being hospitalized. Throughout the problem, assume that you have a random sample of individuals.

   (a) (10) Using the models from the class, propose a statistical model that relates hospitalizations, age (linearly) and the vaccination status.

$Y_i = 1$ if ind. is hospitalized

We can use Probit model

$$P(Y_i=1 \mid Age_i, Vac_i) = \Phi(\alpha + \beta Age_i + \gamma Vac_i)$$

   (b) (10) What is the formula for the average vaccine effectiveness using your model?

Vac. effectiveness conditional on age:

$$P(Y_i=1 \mid Age_i, Vac_i=1) - P(Y_i=1 \mid Age_i, Vac_i=0) =$$
$$= \Phi(\alpha + \beta Age_i + \gamma) - \Phi(\alpha + \beta Age_i)$$

Average effectiveness:

$$\frac{1}{N} \sum_{i=1}^{N} [\Phi(\alpha + \beta Age_i + \gamma) - \Phi(\alpha + \beta Age_i)]$$

Name:

Student ID:

(c) (20) Suppose that you estimated the model using a random sample of Californians. Next you would like to predict the vaccine effectiveness in Florida that has different demographic age distribution. Suppose that you know the distribution of Age, $P(Age_i = k)$, in Florida. Moreover, for your thought experiment suppose that every individual in Florida receives the vaccine irrespective of age. How would you compute the predicted average effectiveness of the vaccine in Florida? Please, explain and provide the corresponding formula.

$\hat{E}$ effectiveness $= E(\hat{E}(\text{eff.} \mid age_i)) =$

$$= \sum_{k=0}^{\infty} \left(\Phi(\hat{\alpha} + \hat{\beta} \cdot k + \hat{\gamma}) - \Phi(\hat{\alpha} + \hat{\beta} \cdot k) \right) \cdot P(Age_i = k)$$

We need to use age distribution of Florida's population to account for the distiction between the two states.

$\hat{\alpha}, \hat{\beta}$ are computed using californian data.

Name: Bulat Gafarov

Student ID:

---

3. **(30 points total) Maximum Likelihood Estimators.** Consider a random variable Y_i that takes values in $\{1, 2, 3\}$. The probablity distribution is parametrized as $P\{Y_i = 1\} = p$, $P\{Y_i = 2\} = q$, and $P\{Y_i = 3\} = 1 - p - q$. Suppose you observe a random sample $\{Y_i\}_{i=1}^n$.

   (a) (10) Write down the corresponding log likelihood function for a sample of n observations using short representation of the indicator functions $X_i = \mathbb{1}\{Y_i = 1\}$ and $Z_i = \mathbb{1}\{Y_i = 2\}$, which imply $W_i = \mathbb{1}\{Y_i = 3\} = (1 - X_i)(1 - Z_i)$. (Hint: use Bernoulli likelihood function from the class as an related example.)

$$\sum_{i=1}^{n} X_i \log p + Z_i \log q + W_i \log(1-p-q)$$

(For Bernoulli we had $\sum_{i=1}^{n} X_i \log p + (1-X_i)\log(1-p)$)

   (b) (10) Write down the first order optimum conditions that define MLE for parameters p, q.

$$\frac{\partial \log L}{\partial p} = \sum_{i=1}^{n} \left(X_i / p - W_i / (1-p-q) \right) = 0$$

$$\frac{\partial \log L}{\partial q} = \sum_{i=1}^{n} \left(Z_i / q - W_i / (1-p-q) \right) = 0.$$

or

$$\frac{\bar{X}}{p} = \frac{\bar{W}}{1-p-q} \quad \text{and} \quad \frac{\bar{Z}}{q} = \frac{\bar{W}}{1-p-q}$$

Name:

Student ID:

(c) (10) Find the MLE for p, q in the explicit form as functions of averages $\bar{X}, \bar{Z}, \bar{W}$.

We need to solve the F.O.C.

$\hat{p} = \frac{\bar{X}}{\bar{W}}(1-\hat{p}-\hat{q})$, $\hat{p} = \frac{\frac{\bar{X}}{\bar{W}}(1-\hat{q})}{1+\frac{\bar{X}}{\bar{W}}}$

also, $\hat{q}\cdot\bar{X} = \hat{p}\cdot\bar{Z}$, so $\hat{q} = \frac{\bar{Z}}{\bar{X}}\cdot\hat{p}$

so $\hat{p} = \frac{\frac{\bar{X}}{\bar{W}}(1-\frac{\bar{Z}}{\bar{X}}\hat{p})}{1+\frac{\bar{X}}{\bar{W}}}$

$\hat{p}\left(1+\frac{\bar{X}}{\bar{W}}+\frac{\bar{Z}}{\bar{W}}\right) = \frac{\bar{X}}{\bar{W}}$

$\hat{p} = \frac{\bar{X}}{\bar{W}} \Big/ \left(1+\frac{\bar{X}+\bar{Z}}{\bar{W}}\right) =$

$= \frac{\bar{X}}{\bar{W}+\bar{X}+\bar{Z}} = \frac{\bar{X}}{1} = \bar{X}$

here we used $X_i + Z_i + (1-X_i)(1-Z_i) = 1$

By symmetry,

$\hat{q} = \bar{Z}$

To summarize, $\hat{p}, \hat{q}$ are sample freq. of $\{Y_i = 0\}$ and $\{Y_i = 1\}$, respectively.