First Assignment

Malgorzata, Kevin, Johannes

December, 2018

First Problem: (Simulation: Latent Variable, Probit Model):

Consider the following JOINT distributions of random variables $X_0 \& \varepsilon_j^*$ (You must simulate it as a joint (multivariate) normal distributions!).

$$\begin{pmatrix} X_0 \\ \varepsilon_1^* \end{pmatrix} \sim N\left[\begin{pmatrix} 10 \\ 0 \end{pmatrix}, \begin{pmatrix} 2^2 & 0 \\ 0 & 1 \end{pmatrix}\right], \begin{pmatrix} X_0 \\ \varepsilon_2^* \end{pmatrix} \sim N\left[\begin{pmatrix} 10 \\ 0 \end{pmatrix}, \begin{pmatrix} 2^2 & 0 \\ 0 & 2^2 \end{pmatrix}\right], \begin{pmatrix} X_0 \\ \varepsilon_3^* \end{pmatrix} \sim N\left[\begin{pmatrix} 10 \\ 0 \end{pmatrix}, \begin{pmatrix} 2^2 & 3 \\ 3 & 2^2 \end{pmatrix}\right]$$

with

$$\beta_0 = -30$$
$$\beta_1 = 4$$

$$Y_j^* = \beta_0 + \beta_1 * X_1 + \varepsilon_j$$

$$Y_j = \begin{cases} 1 \text{ if } Y_j^* > 0 \\ 0, \text{ otherwise} \end{cases}$$

$$j \in \{1, 2, 3\}$$

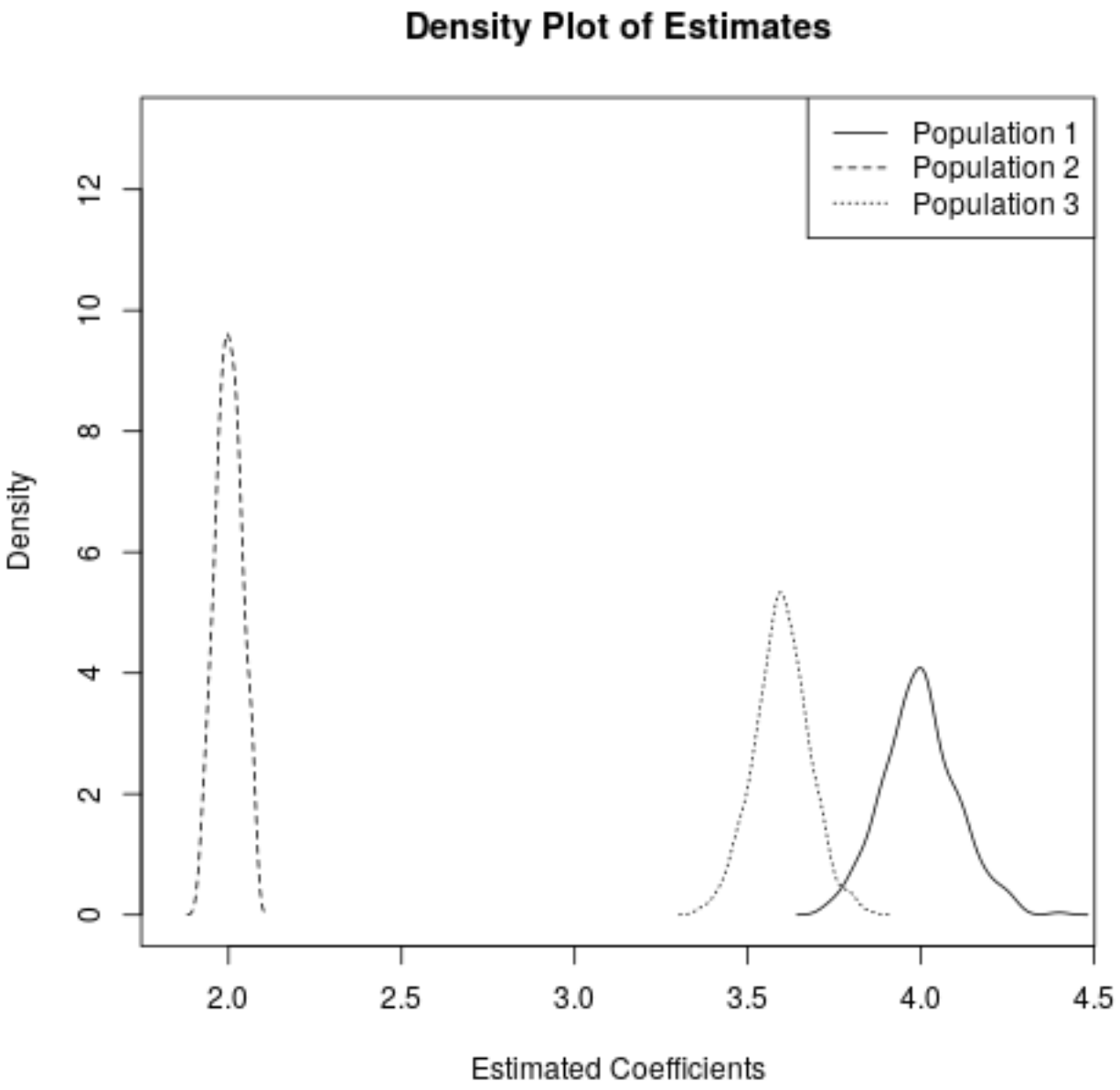
a) [1P] Please simulate the three datasets (i.e.: $Y_j, X_1, \varepsilon_j^*; j \in \{1, 2, 3\}$) with a sample size of 30000 observations. For each of the three datasets, estimate a probit model of $\mathrm{Y_i}$ on x_1 and save the estimate for $\hat{\beta}_1$.

Given our simulated Data, we calculate the Probit model $G(\cdot)$ to get the Maximum Likelihood estimator $\hat{\beta}_j$:

$$\pi = G\left(\mathbf{x}'\beta\right) = \Phi\left(\mathbf{x}'\beta\right) = \int_{-\infty}^{\mathbf{x}^*\beta} \frac{1}{\sqrt{2\pi}} \exp\left[-\left(z^2/2\right)\right] \mathrm{d}z$$

The estimates for $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$ are 4.12, 2.03 and 3.69 respectively.

b) [1.5P] Repeat a 400 times while saving all the different estimate in vectors for all three models (based on the three different populations). Plot the kernel density estimates for $\hat{\beta}_1$ based on the three populations next to each other. Describe shortly the choices you faced and made estimating the density functions.



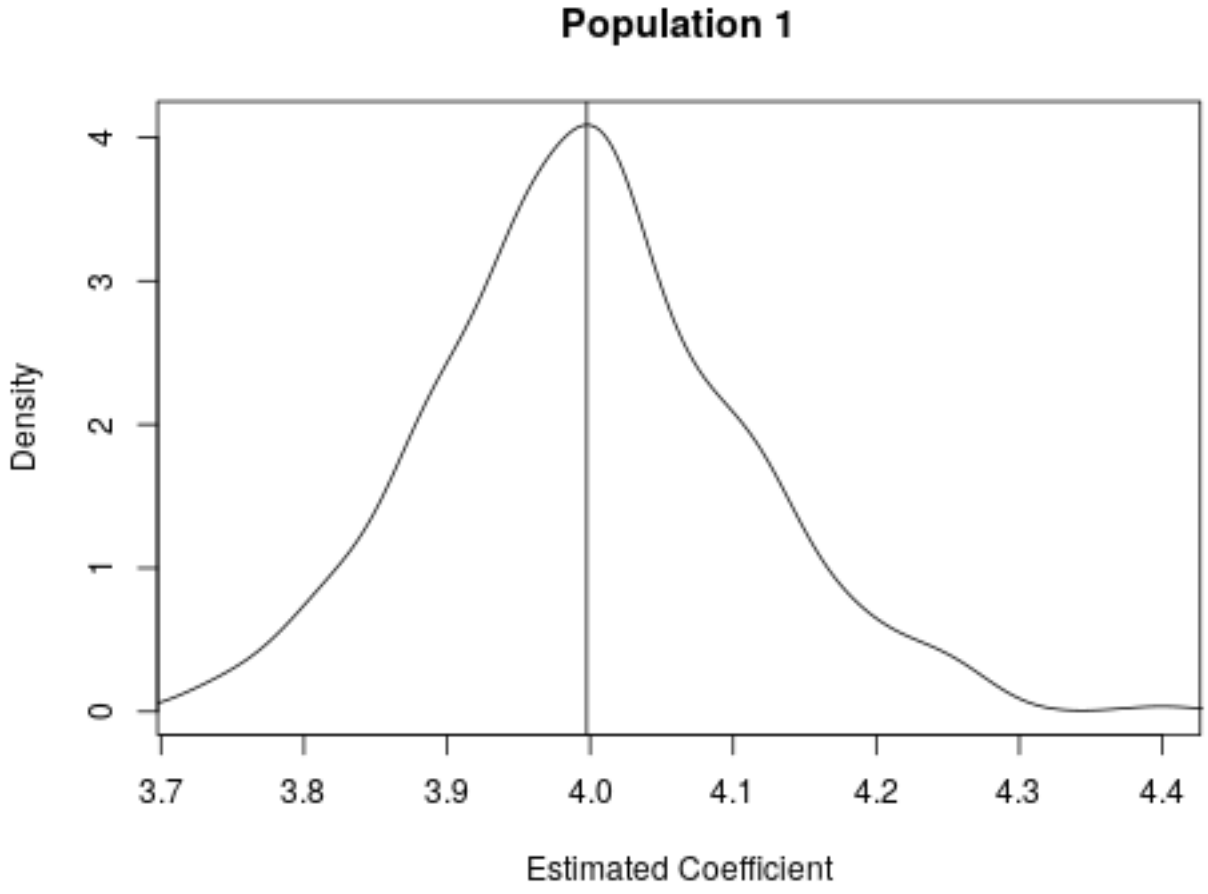
Based on our simulation, we estimated 400 random instances of β_j for each of the three Populations() $Y_1, X_1, \varepsilon_1^*$). To get three smooth Distributions for our estimates we used the kernel density estimator

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - x_i\right) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where we choose K as a *Gaussian Kernel*, which is the convential choice and produces a standard normal density function. Another option would have been the *Epanechnikov*, which is optimal in a mean square error sense. Given our Data - which is supoosed to follow a standard normal distribution - the Gaussian Kernel is optimal, though the loss of efficiency is generally small for other Kernels. In our formular h is the smoothing parameter called the bandwidth and had to be

choosen accordingly to the data. The goal is to find a bandwidth that is as small as possible to avoid loss of information and represent the approximatly true variance of the estimator. On the other side we want our bandwidth to smoothen the function in a way that we can interprete the true distribution of our estimator. If the bandwith is to small, we will fail to get an interpretable result. A usual rule-of-thumb for the bandwidth is the so called "Silvermann's *rule of thumb*" which is 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power (See Silvermann 1987, page 48, eqn (3.31)). It is used as a default for calculating the density in R and produced good result given our data. We also estimated the density plot with a slightly reduced bandwidth, which lead to a unstable distribution picture, especially for population 1. Therefore we remained with the default setting of R.

i.) [1.5P] Now concentrate on the kernel density plot based on the first population (i.e.: $Y_1, X_1, \varepsilon_1^*$). Does the distribution of $\hat{\beta}_1$ conform to your expectations? More specifically, explain which distribution you would expect (and why) and whether the plotted density conform to that expectation (no formal tests necessary).



Given the basic properties of a ML-estimator we expect or estimator $\hat{\boldsymbol{\theta}}_N$ with sample size N to converge in probabilty to the true parameter θ. Therefor we expect $\hat{\boldsymbol{\theta}}_N$ to be a consistent estimator and to be asypmtotically normally distributed with asymptotic variance of $\sqrt{N}\left(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}\right)$.

Therefore we expect $\hat{\boldsymbol{\theta}}_N$ to follow

$$\hat{\boldsymbol{\theta}}_N \sim \text{Normal}(\boldsymbol{\theta}, \mathbf{V}/N)$$

The plotted density confirms our expectation by first glance. The estimator is consistent around its expected value of 4, which is equivalent to the true parameter β_1 of our latent model.

ii.) [1P] Compare the distributions of $\widehat{\beta_1}$ from the population j = 1 to the $\widehat{\beta_1}$ from the population j = 2 and j = 3, respectively. Why do the means of the distributions differ?

ii.) [1P] Compute the mean of $\widehat{\beta_1}$ from the population j = 2. Can you explain, why the distribution of this particular $\widehat{\beta_1}$ concentrate approximately around this value?

Simulate again all three populations as you did in a. But this time, estimate and save the average marginal probability effect of x_1 Repeat the estimation 400 times with a sample size of each iteration equaling 30000 .

c.) [1P] Analogously to b), plot the kernel density estimates for all three AMPE's. Determine the values around which the sample distributions are concentrated.

i.) [1.5P] Calculate the relative difference (in percent) between $\hat{E}[AMPE|j = 1]$ and $\hat{E}[AMPE|j = 3]$. Which estimate would you use to ascertain the effect of the variable x_1?

ii.) [1.5P] Calculate the relative difference (in percent) between $\hat{E}[AMPE|j = 1]$ and $\hat{E}[AMPE|j = 2]$.

iii.) [1.5P] Based on the results in b-ii) would you expect the differences observed in c-i) and c-ii)? Please provide a detailed explanation for the observed results.

Problem 2: (Marginal effects estimation & Interpretation):

Load the dataset "south_african_heart_disease_data.dta" and estimate the effect of ldl-(bad)cholesterol (ldl) in blood on the probability of suffering from heart disease (chd equals 1 if one suffers from it).

a.) [0.5P] Can you learn anything from the estimated coefficients? Explain shortly.

b.) [0.5P] Are the S.E. valid, or do you need to adjust them for heteroscedasticity? Explain.

c.) [0.5P] Re-estimate the model from a) but this time include age in addition to ldl. You see that the estimated coefficient of ldl changes. Explain why? Additionally, show that your explanation is supported by the data.

d.) Finally, estimate the model from a) but include ldl squared next to ldl as a control variable.

i.) [1P] Based on the estimated coefficients from a) and d) draw the two resulting marginal probability effects of ldl as a function of ldl for ldl $\in [1; 15]$ next to each other.

ii.) [0.5P] Are any of the marginal probability effects linear in ldl? Explain why.

iii.) [1P] What is the advantage of the marginal probability effect based on the estimation in d) over the one based on a)? Explain shortly.

iv.) [1.5P] Calculate and properly interpret both marginal probability effects for the mean value of ldl in the sample (You do not need to compute standard errors).

iv.) [1P] Are any of the effects computed in iv), ceteris paribus effects? Explain shortly.