# 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 Y $_i$ on $x_1$ and save the estimate for $\hat{\beta}_1$.

Given our simulated Data, we calculate the Probit model for our binary dependend variable $y_i$ given $x_1$

$$f(y_i|\mathbf{x_1}) = [\Phi(\mathbf{x}_1'\boldsymbol{\beta_1})]^{y_i} [1 - \Phi(\mathbf{x}_1'\boldsymbol{\beta_1})]^{1-y_i} \quad y_i = 0, 1$$

to get the probit estimator $\hat{\beta}_{1,j}$ for j = 1,2,3 through the maximization of the log likelihood of our density function $f(y_i|\mathbf{x_1})$.

The estimates for $\hat{\beta}_{1,1}, \hat{\beta}_{1,2}$ and $\hat{\beta}_{1,3}$ 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.

## Density Plot of Estimates
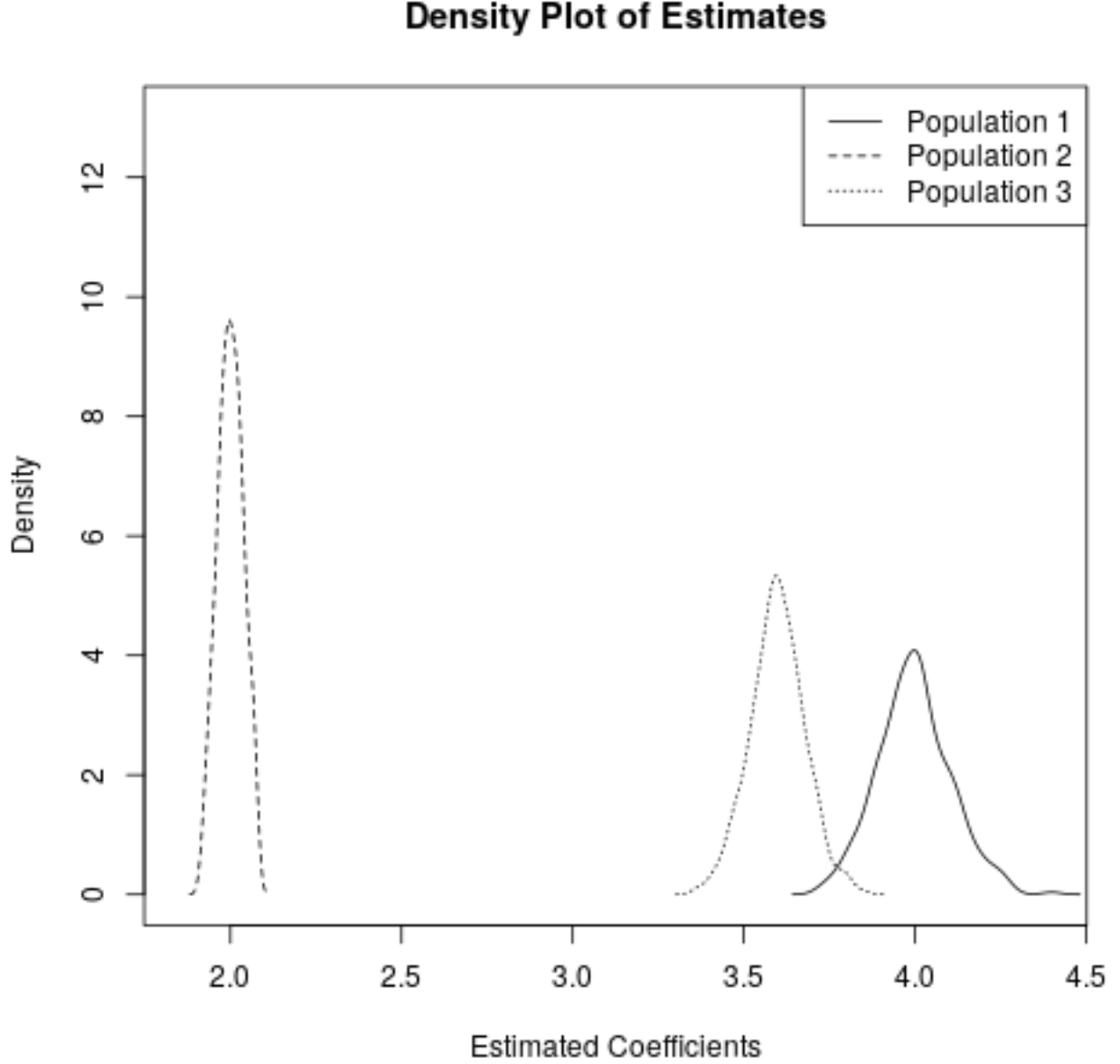


Based on our simulation, we estimated 400 random instances of $\hat{\beta}_{1,j}$ for each of the three Populations$(Y_j, X_1, \varepsilon_j^*)$. In order to be able to interprete our distributions, we need to find a smooth representation for the probability density function of our estimates. The typical non-parametric way to achieve a smooth function is to make use of a kernel 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 $K$ to be a *Gaussian Kernel*, which is the convential choice and produces a standard normal density function. That means that $K(x) = \phi(x)$, where $\phi$ is the standard normal density function. Given that we expect our estimate to be approximately normal (see below), this is the best choice for our kernel density estimation. The smoothing parameter $h$ is called the bandwidth

2

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 as big as necessary to represent the true density of our estimates. If the bandwith is to small, our function curve will be undersmoothed. which means that our function will contain jumps that are not part of our true density. On the other side, if our bandwidth is to big and our curve will be oversmoothed. The most extrem example for this case is a flat line. A rule-of-thumb for the bandwidth in R is "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 man page of function density() in R). 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).



**Population 1**
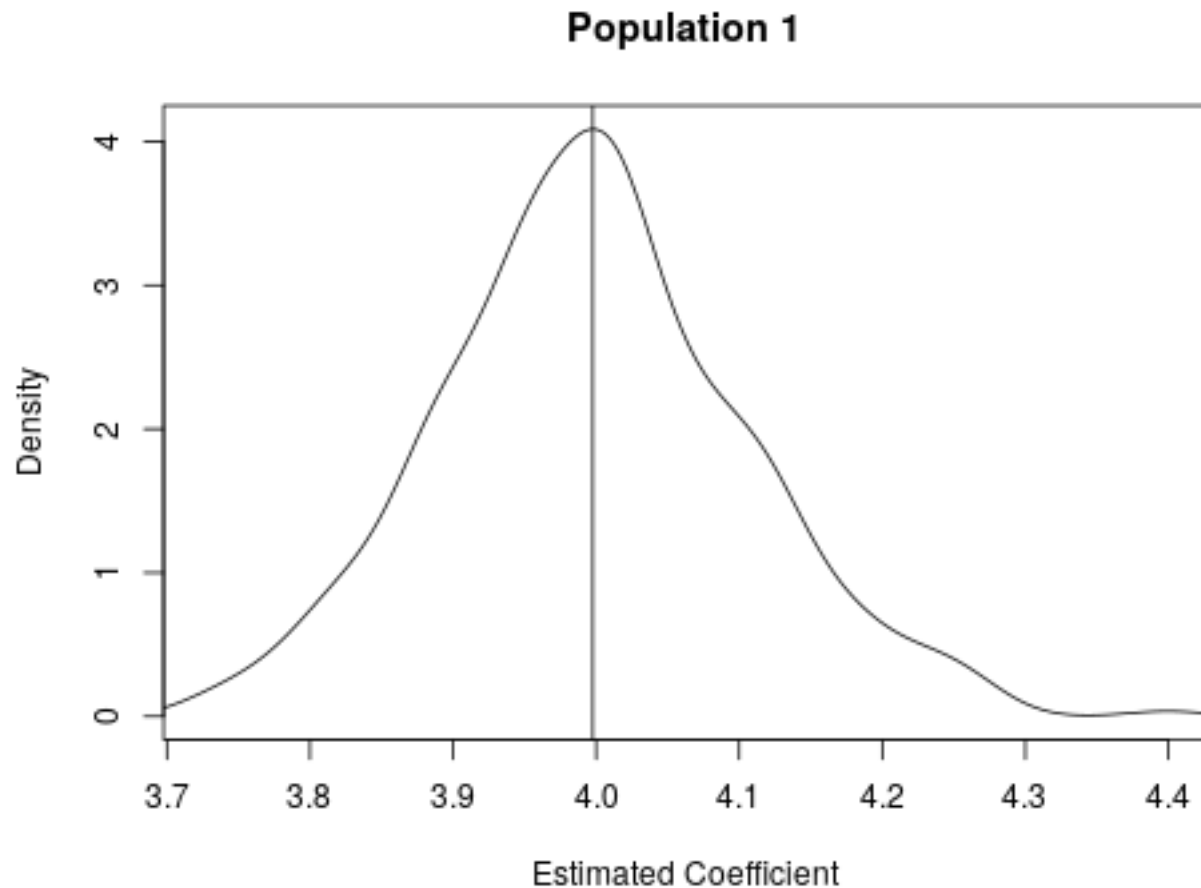
Given the basic properties of a ML-estimator we expect our estimates of $\hat{\beta}_{1,j}$ to be a random variable with probability distribution that is approximately normal and to converge in probability

to the true parameter $\beta_{1,j}$. Therefore we expect $\hat{\boldsymbol{\beta}}_{1,j}$ to follow

$$\hat{\boldsymbol{\beta}}_1 \sim \text{Normal}(\boldsymbol{\beta}_1, \mathbf{V})$$

with the asymptotic variance of $\hat{\beta}_{1,j}$

$$\text{Avar}(\hat{\boldsymbol{\beta}}_{1,j}) \equiv \left\{ \frac{\left[\phi\left(\mathbf{x}_1\hat{\boldsymbol{\beta}}_{1,j}\right)\right]^2 \mathbf{x}_1'\mathbf{x}_1}{\Phi\left(\mathbf{x}_1\hat{\boldsymbol{\beta}}_{1,j}\right)\left[1 - \Phi\left(\mathbf{x}_1\hat{\boldsymbol{\beta}}_{1,j}\right)\right]} \right\}^{-1} \equiv \hat{\mathbf{V}}$$

The plotted density confirms our expectation by first glance. The estimates of $\hat{\beta}_{1,j}$ for our first population concentrate around its expected value of 4, which is equivalent to the true parameter $\beta_1$ 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?

As we can can see in Plot 1 all 3 distributions are approximatly normally distributed, but differ in their means and variances. For the means we find that $\hat{\beta}_{1,1} = 4.12$, $\hat{\beta}_{1,2} = 2.03$ and $\hat{\beta}_{1,3} = 3.69$. As we explained earlier the simulated errors of $\hat{\beta}_1$ follow a standard normal distribution and we expect the estimate of $\hat{\beta}_1$ to be close to the true value of 4. In the second population our errors are not standard normal and intruduce a attenuation bias. The third population introduces a Covariance between the explanatory variable and the errorterm in our latent model. Therefore our error term is not independent of the explanatory variable in our latent model and the estimate of $\hat{\beta}_1$ from the population j = 3 is inconsistent.

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?

We know from the issue of Neglected Heterogenity, which has the structural model of

$$P(y = 1|\mathbf{x}, c) = \Phi(\mathbf{x}\beta + \gamma c),$$

where $\gamma c$ represents omitted variables which are independent of x and typically included in the errorterm. We assume that $c$ is independent of $\mathbf{x}$ and $c \sim$ Normal $(0, \tau^2)$. Now we know that probit of y on x consistently estimates plim $\hat{\boldsymbol{\beta}}/\sigma$. We call this the attentuation bias. Our error term is

$$\epsilon = u + \gamma c \sim N\left(0, \gamma^2\tau^2 + 1\right)$$

and we know from our distribution that $\epsilon \sim$ Normal $(0, 4)$. Therefore we know that $\sigma = 4^{1/2}$. Our estimated mean of $\hat{\beta}_{1,2}$ is approximately equal to the value of $\beta/\sigma = 2$.

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.

## Density Plot of AMPES
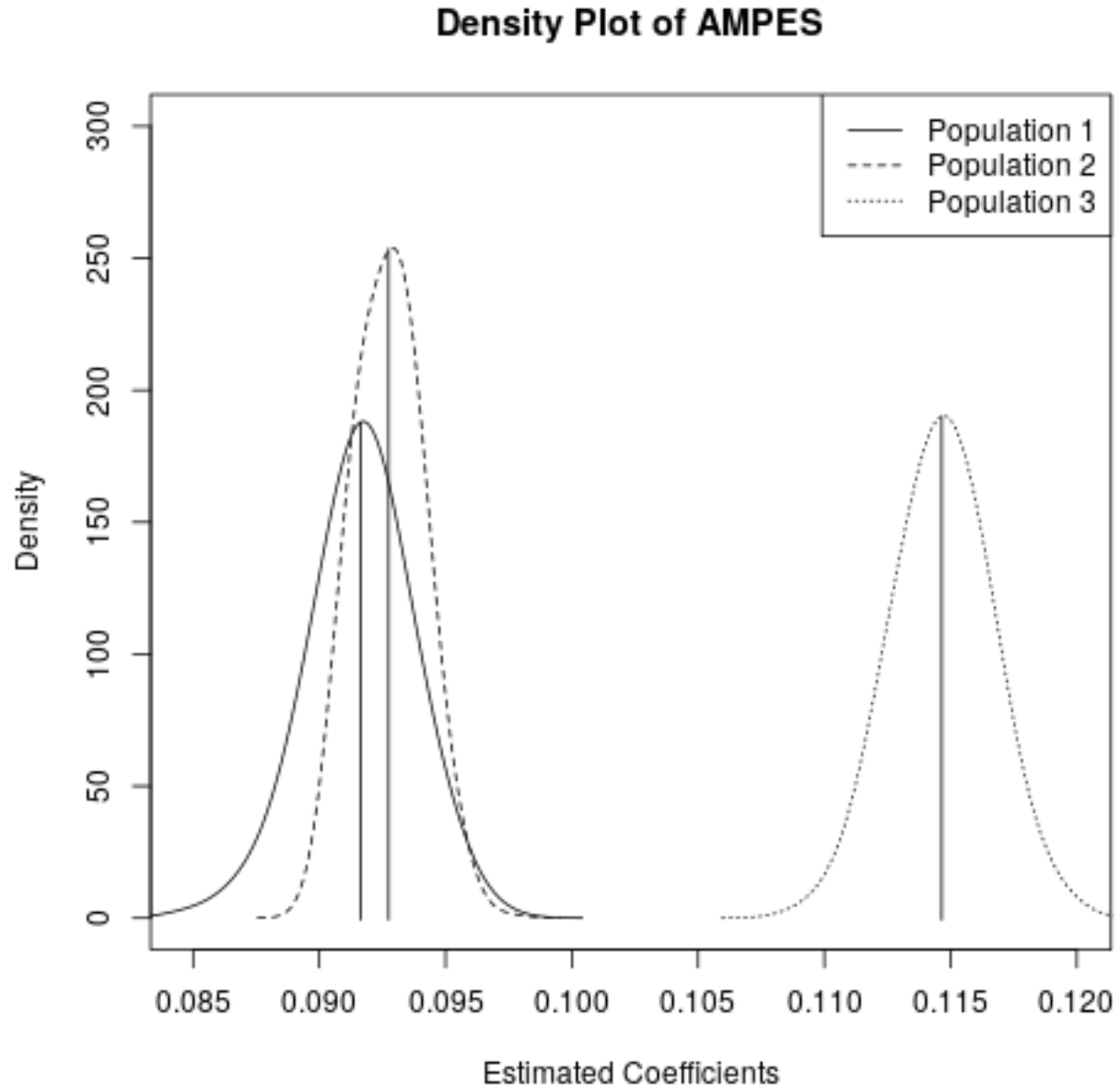


Again we can see that the sample distributions (with N = 400) are approximately normal and concentrated their means. For population j = 1 we get 0.0916318 for $\hat{E}[AMPE|j=1]$, for population j = 2 we get 0.0927414 for $\hat{E}[AMPE|j=2]$ and for population j = 3 we get 0.1146494 for $\hat{E}[AMPE|j=3]$.

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 $x_1$?

The relative difference between $\hat{E}[AMPE|j=1]$ and $\hat{E}[AMPE|j=3]$ is 25.1196038 percent. The relative difference of both estimates is rather big and we have to be careful, which one of

the two we would like to choose for a valid interpretation of the effect of variable $x_1$. In our case would definately choose the estimate of $\hat{E}[AMPE|j = 1]$ because we know from the distribution parameters of population 1 that our error terms are following a standard normal distribution and therefore that our estimate is consistent and unbiased. We also know, that in Population 3 the errorterms are correlated with our independent variable $x_1$ and we assume that the resulting estimate for $\hat{E}[AMPE|j = 3]$ is biased.

ii.) [0.5P] Calculate the relative difference (in percent) between $\hat{E}[AMPE|j = 1]$ and $\hat{E}[AMPE|j = 2]$.

The relative difference between $\hat{E}[AMPE|j = 1]$ and $\hat{E}[AMPE|j = 2]$ is 1.2109329 percent.

iii.) [3P] 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.

Based on our Plot 1, we can expect the differences in c-i) and c-ii). First of all, we know from the definition of average marginal probability effects that the attentuation bias or unobserved heterogenity average out. Therefore the probit of y on x consistently estimates the average partial effects even when we account for unobserved heterogenity. As a result we expect our estimates of poulation j = 1 and j = 2 to be very close, which is conformed in our Plot 3. We also know that we have to compare the magnitude of our estimates from the samples of population 3 with those from population 2, given the fact they have the same variance of error terms and the only difference in population parameters is the correlation of error terms with $x_1$ in poulation 3. Again, as we have seen in Plot 1, the magnitude for our estimate $\hat{E}[AMPE|j = 3]$ is bigger than the estimate of $\hat{E}[AMPE|j = 2]$, which follows the relation between $\hat{\beta}_{1,2}$ and $\hat{\beta}_{1,3}$ from Plot 1. This makes sense, since the average marginal probability effect $\beta_K E[\phi(\mathbf{x}\beta)]$ summarizes the estimated marginal effects of $x_1$ on $p(\mathbf{x})$, which depend on x through $(\phi\mathbf{x}\boldsymbol{\beta})$.

Therefore the size of $\hat{\beta}_1$ is reflected in the average of the estimates for the marginal probability effects. Since $\hat{\beta}_3$ is bigger than $\hat{\beta}_2$ it makes sense that $\hat{E}[AMPE|j = 3]$ is also bigger than $\hat{E}[AMPE|j = 2]$.

**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.