UNIVERSITY OF GOTHENBURG
School of Business, Economics and Law

June 15, 2016

Bachelor thesis

Logistic regression: effect of unobserved heterogeneity on estimator's bias and variance

Author: Pavle Pazanin
Supervisor: Kristofer Månsson

Contents

Abstract

In this paper we study unobserved heterogeneity in logistic regression, which occurs as a result of omitted variables. Unlike linear regression, logistic regression estimates are affected by model misspecification even if omitted variables are not correlated to the explanatory variables in the model. As a result, interpretation of log-odds ratios and odds ratios is not straight forward and similar models with different independent variables can not be compared. We study average marginal effect as a possible measure of overcoming the unobserved heterogeneity problem.

1 Introduction

Throughout various fields in research one might be interested in studying qualitative (or categorical) outcome that is represented by binary variables. Such practice may often arise in medical research (for example the presence or absence of a particular disease), economics (being classified as unemployed, or within some credit-risky group), demographic research (married - divorced) or political research (voting preferences). Logistic regression has become one of the most commonly-used statistical methods for modeling such dichotomous outcome (Harrel and Frank, 2001). Popularity aside, a care should be taken when using logistic regression since recent research shows that our usual way of interpreting results from logistic regression might lead to several issues that arise from the fact that we are often not able to include all relevant variables in our model. However, unobserved variables still affect the outcome. Variations in the outcome, or the dependent variable, due to omitted variables is called unobserved heterogeneity (Mood, 2009). In this paper we investigate the effect of unobserved heterogeneity on logistic regression coefficients. Further, we study one of several methods designed to address the problem of unobserved heterogeneity.

As noted above, some researchers have shed light on the problem caused by unobserved heterogeneity in logistic regression. Cramer (2006) writes about logit and probit models, having omitted variables in focus. He refers to Wooldridge's analysis of downward bias on probit coefficients that occurs when omitting relevant variable from the model and widens the discussion on logistic regression. Using Wooldridge's average partial effect estimator (APE) as a starting point, Cramer defines a similar estimator, average sample effect (ASE), and investigates the effects of unobserved heterogeneity on it. ASE is tested in a simulation using a latent variable framework. He studies the effect of omitted independent variable, even as a result of using different distributions. The simulation shows no bias in ASE when a relevant independent variable is omitted.

Mood (2009) studies the unobserved heterogeneity from the aspect of our usual interpretation of logistic regression results. She shows how omitted variables can lead to misleading interpretation even when they are not related to the independent variables in the model. Her article describes several estimators that are used to overcome this issue. Marginal effect at mean is obtained by taking the value of logistic probability density function and plugging in the predicted logit of all independent variables being set to their means. This value is multiplied by the estimated coefficient for the independent variable of interest. Average marginal effect is obtained by taking average of logistic probability density function of the predicted logit, evaluated for all values of independent variables in the sample. Resulting value is multiplied by logistic regression coefficient. Average partial effects is calculated in the same way as average marginal effect, but within a specific range of values on the independent variable related to the aforementioned coefficient. In addition, she also investigates linear probability models (LPM) as a possible solution in a certain contexts, and compares the LPM estimators to average marginal effects from logistic regression.

Karlson,Holm and Breen (2010) compare regression coefficients between logit and probit models. They suggest that comparison of coefficients of a certain variable across different probit and logit models should be avoided, since the change in the coefficient of interest cannot be directly attributed to the inclusion of confounding variables. Using the latent variable framework they show that average partial effect estimator, as defined by Wooldridge, can be sensitive to rescaling. They claim to take even a step further and develop another estimation method that allows unbiased comparisons of logit or probit coefficients of the same variable (x) across nested models successively including control variables (z). According to them, the method extends all the decomposition features of linear models to logit and probit models.

Wooldridge (2009) writes about interpretation of logit and probit models. Although the unobserved heterogeneity is not analysed in depth, he does note the rescaling problem and inability to directly compare magnitudes of the coefficients estimates across models. For this purpose he derives two types of rescaling factors - partial effect at the average (PEA) and average partial effect (APE). They are calculated in the same way as Mood (2009) describes (marginal effects and average partial effects). Wooldridge shows that one can make a consistent comparison across logit, probit and LPM models coefficients using these estimates.

In this paper we focus on one of the solutions defined by Wooldridge - average marginal effect (AME). Since this estimator is already well studied in the literature in terms of bias we will simulate and study both it's bias and variance. However, while this estimator is often studied in an asymptotical manner, our main focus is on empirically relevant sample sizes. After introduction we give a short overview of logistic regression, explaining the latent variable model and how omitted variable can affect the outcome. After that we will describe three types of estimates relevant for our discussion. These include logistic regression coefficients, average marginal effects and mean squared error for the average marginal effect, as a variance estimator. Next part is empirical analysis where we conduct our simulation experiment and test behaviour of average marginal effect. In addition to the simulation study, we illustrate the effect of unobserved heterogeneity on a model built on MROZ dataset. After that we present and discuss the results. The final remarks and thoughts are given in conclusion.

Since equations appear in the text quite often, it is good to describe how they are organized. Throughout this paper we will use the following notation unless otherwise stated:

n the number of observed cases

p the number of parameters

y $n \times 1$ vector with ith element y_i

X $n \times p$ design matrix where the first column is always a vector of 1's, and the other columns represent variables

X_i ith row of the matrix X, that contains all of it's columns: $x_0, x_1, ...x_p$. As noted above, the first column x_0 has value 1 through all the rows

x_j jth column of the matrix X. In equations for the ith case (y_i) we avoid writing explicitly the first column of X (since it's value is 1). As a result, the intercept is simply β_0 and x_{ij} represents the ith observation in the variable j

β $p \times 1$ vector with element β_j, the coefficient for the jth parameter

2 Background

2.1 Logistic regression

Binary outcome is often modeled with logistic regression, but it can also be modeled with linear probability models that are based on OLS. Due to importance of OLS, we will compare these two methods in few words. Logistic regression belongs to the family of generalized linear models, and linearity in parameters can make it similar to OLS. Further, since probability is measured in a range from 0 to 1 we need a model with a response variable that falls within this interval.In OLS solution to this problem is achieved with linear probability model (LPM), where the linear regression is adjusted so that we focus specifically on the binary outcome of response variable. In practice this has at least two shortcomings - for certain values of predictor variable it is still possible to get predictions less than 0 or larger than 1. Another problem is linearity itself, which does not hold in many contexts. In logistic regression these problems are addressed by introducing a link function that is more appropriate for modelling probabilities (since it returns output between 0 and 1). Unlike linear regression where we model the response $E[y]$ directly, in logistic regression we estimate probability that the response variable falls in a particular class by modeling the link function - $logit(E[y])$.

Let us briefly motivate the use of link function in logistic regression. When dealing with binary dependent variable it is reasonable to assume that it follows a Bernoulli distribution where $E[y] = P(y = 1)$ and the mean must be bounded between zero and one. In order to satisfy this boundary we model $E[y]$ using logistic cumulative distribution function. Assume that the ith response of the variable y is a linear function of predictors X_i such as

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \tag{1}$$

then we can model the ith outcome of the variable y with a cumulative logistic distribution as

$$E[y_i] = P(y_i = 1|X_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}}. \tag{2}$$

After re-arranging terms we can write the expression above as

$$\frac{p(X_i)}{1 - p(X_i)} = e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}, \tag{3}$$

where the left side of the equation denotes odd-ratio (McCulloh and Searle, 2001). In order to have the model linear in parameters we take a natural logarithm of both sides

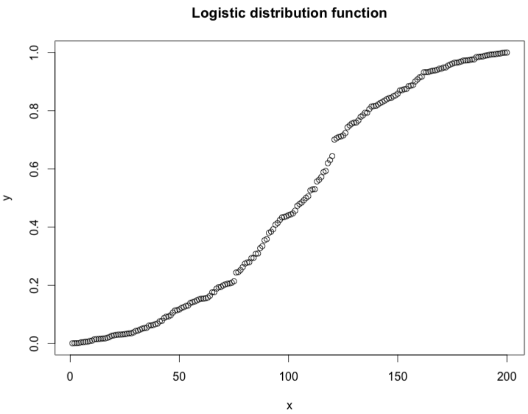
$$ln\Big[\frac{p(X_i)}{1 - p(X_i)}\Big] = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}. \tag{4}$$

The right side of the equation above is our linear model, and the left side is $logit(E[y_i])$ or log-odds ratio. Figure 1 shows the graph of the realization of the empirical logistic distribution function, with characteristic "S"- shaped (sigmoid) curve that ranges between zero and one.

2.2 Latent variable model

In this paper we study logistic regression model that is derived from an underlying latent variable. Latent variable model represents an underlying process that we can not observe or quantify, but

Figure 1: Realization of the empirical logistic distribution function



we may still be able to model it's eventual presence or absence with logistic regression. Let us think of y^* as an unobserved latent process and define it's ith outcome as

$$y_i^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \tag{5}$$

where β_0,β_1,β_2 are parameters, x_{i1}, x_{i2} are independent variables for the ith case and e_i is corresponding error, unobserved but assumed independent of x_{i1} and x_{i2} (Karlson,Holm & Breen, 2010). The number of explanatory variables can be larger, we use two for simplicity. Since the process y^* is not observed, we can not model it directly. Instead, let us assume that we can observe a binary variable y that takes values.

$$y_i = \begin{cases} 1 \text{ for } y_i^* > 0 \\ 0 \text{ for } y_i^* \leq 0. \end{cases} \tag{6}$$

We can further explain the latent variable framework through some practical examples. For instance, we can't observe a customer's utility of buying a car - all we can see is whether a customer has bought a car or not. Similarly, it would probably be hard to observe a feeling of safety of a personal property. What we can observe is whether a person buys an extra insurance or not. So, although y^* is not possible to estimate directly, we can model the binary variable y that reflects behaviour of y^*.

In further discussion we are interested in how changes in the latent, unobserved model affect it's variance and therefore the estimated, logistic regression model. In Equation 5 denote standard deviation of the error term $sd(e) = \sigma^*$. The total variance in the latent variable y^* consists of the explained variance, $Var(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$, and residual variance $Var(e)$.

If we estimated this model with OLS, adding new variables would not change the total variance; explained variance would increase and unexplained variance would decrease. However, since we use logistic regression to analyze binary outcome y of the latent process y^*, we assume that the error e_i from Equation 5 is independent of x_{ij} and has the standard logistic distribution. As a result of this assumption we can rewrite the latent variable for the ith case as

$$y_i^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \sigma^* u_i, \tag{7}$$

where σ^* is a scaling parameter defined as a ratio of the true and assumed standard deviation for e, and u_i is random variable that follows standard logistic distribution with the mean $\mu = 0$ and variance $\sigma^2 = \frac{\pi^2}{3}$ (which we round to 3.29). In other words, $\sigma^* = sd(e)/sd(u)$.
Next, we define the response probability for y_i as

$$\begin{aligned} P(y_i = 1|X_i) &= P(y_i^* > 0|X_i) \\ &= P[e_i > -(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})|X_i] \\ &= P\left[u_i < \frac{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}{\sigma^*}|X_i\right] \end{aligned} \tag{8}$$

which we write in terms of the equation (3) as

$$P\left[u_i < \frac{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}}{\sigma^*}|X_i\right] = \frac{exp(\frac{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}{\sigma^*})}{1 + exp(\frac{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}{\sigma^*})}. \tag{9}$$

In other words, after estimating the latent variable y^* with logistic regression the result is

$$logit[P(y_i = 1|X_i)] = \beta_0^L + \beta_1^L x_{i1} + \beta_2^L x_{i2}, \tag{10}$$

where

$$\beta_0^L + \beta_1^L x_{i1} + \beta_2^L x_{i2} = \frac{\beta_0}{\sigma^*} + \frac{\beta_1}{\sigma^*} x_{i1} + \frac{\beta_2}{\sigma^*} x_{i2}. \tag{11}$$

Equations 7,8 and 9 reveal the mechanism of logistic regression. Since we assume that the variance for e_i in Equation 5 follows logistic distribution (meaning that it becomes fixed to 3.29), and can not observe σ^* from the re-written latent model in Equation 7, using logistic regression we standardize the coefficients β_0, β_1 and β_2 from Equation 7. In equations 10 and 11 we see this standardization, where the estimated coefficients are ratio of the coefficients in the latent model and the scaling parameter σ^* (Karlson, Holm & Breen, 2010). We write the estimated standardized coefficients using notation β^L in order to avoid confusion with non-standardized coefficients further in the text.

2.3 The effect of omitted variable in the latent model

Let us further investigate what happens in case of omitting a variable that is relevant and present in the latent model. In Equation 5, the total variance in in y^* can be decomposed to explained and unexplained, or residual variance. By fixing the residual variance from Equation 5 to the variance of standard logistic random variable, any change in amount of the explained variance leads to change in amount of the total variance in y^* (and hence it's scale), since the variance in e is fixed to 3.29 (Cramer,2007) . With next example we show what happens if we omit a variable from a latent model. Let us assume that the full equation of the latent model is

$$y_i^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \tag{12}$$

where e_i, with zero mean and variance σ_F^2, is uncorrelated with both predictor variables and the relationship between x_2 and x_1 for the ith case can be described as

$$x_{i2} = \delta_0 + \delta_1 x_{i1} + v_i, \tag{13}$$

where δ_0 and δ_1 are parameters and v_i is the error term, uncorrelated to the error e_i from Equation 12. Omitting the variable x_2 affects our analysis in two ways. First, it leads to confounding results in a reduced model where the effect of x_1 is confounded with the effect of x_2. We show it by substituting Equation 13 into Equation 12

$$\begin{aligned} y_i^* &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \\ &= \beta_0 + \beta_1 x_{i1} + (\delta_0 + \delta_1 x_{i1} + v_i)\beta_2 + e_i \\ &= \beta_0 + \delta_0\beta_2 + (\beta_1 + \beta_2\delta_1)x_{i1} + \beta_2 v_i + e_i. \end{aligned} \tag{14}$$

From the Equation 14 we see that the effect that x_{i1} has on the response variable is now $\beta_1 + \beta_2\delta_1$. On the other hand, the error of the latent model increased to $\beta_2 v_i + e_i$.

As we saw in the Equation 11 from the previous section, estimating y^* in Equation 14 with a logistic regression will result in ratio of it's coefficients and the scale parameter. Since we defined the scale parameter as a ratio of the true and assumed residual standard deviation, after omitting the column x_2 the residual variance is $Var(\beta_2 v + e) = \beta_2^2 Var(v) + 3.29$ (Mood,2010). This means that the scale parameter becomes

$$\sigma^* = \frac{sd(\beta_2 v + e)}{sd(u)} = \sqrt{\frac{Var(\beta_2 v + e)}{Var(u)}} = \sqrt{\frac{\beta_2^2 Var(v) + 3.29}{3.29}}. \tag{15}$$

Using notation from Equation 10, the estimated coefficient β_1^L for the model y^* in Equation 14 is

$$\beta_1^L = (\beta_1 + \beta_2\delta_1)\frac{\sqrt{3.29}}{\sqrt{\beta_2^2 Var(v) + 3.29}}. \tag{16}$$

In Equation 13 we defined the relationship between x_1 and x_2. Assuming that they are uncorrelated means that the parameter δ_1 equals zero, and the Equation 16 above becomes

$$\beta_1^L = \beta_1 \frac{\sqrt{3.29}}{\sqrt{\beta_2^2 Var(x_2) + 3.29}}, \tag{17}$$

which shows that the size of unobserved heterogeneity will depend on the variance of the omitted variable. In other words, even when dealing with uncorrelated variables, omission of one of them will cause the logit estimate β_j^L to vary inversely with the extent of unobserved heterogeneity (Cramer,2007). This is different from OLS where an omitted and uncorrelated variable does not lead to bias.

3 Method and theoretical framework

3.1 Estimating coefficients in logistic regression

In this section we derive the estimators for parameters in a logit model. Logistic regression belongs to the family of generalized linear models so parameter estimation differs a lot in comparison to simple (and multiple) linear regression, where OLS is sufficient. Instead, a maximum likelihood estimation (MLE) is applied. Since it does not lead to a closed-form solutions for the estimators, another procedures have to be applied after MLE. In this paper we describe the Newton-Raphson method which in this case is the same as Iterative Re-weighted Least Squares framework (James, Witten, Hastie and Tibshirani, 2013).

In logistic regression response variable can take on two values, 0 or 1, which we denote as "success" and "failure". This means that we can relate the maximum likelihood estimation to the Bernoulli distribution. Assume that Y_i are $Bernoulli(p_i)$ independent distributed random variables where a probability of observing success y_i is given by p_i. Then we describe this relationship with probability mass function of Bernoulli distribution (McCulloh and Searle, 2001)

$$P(Y_i = y_i | p_i) = p_i^{y_i}(1 - p_i)^{1-y_i}. \tag{18}$$

For the maximum likelihood estimation we need the joint probability distribution of variables $Y_1, Y_2, ..., Y_n$, which for Bernoulli distribution is

$$P(Y_i = y_i | p_i) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \tag{19}$$

and we formulate the log-likelihood equation as

$$\begin{aligned} L(\beta) &= log\Big(\prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}\Big) \\ &= \sum_{i=1}^{n} \big(log(1 - p_i) + y_i log(\frac{p_i}{1 - p_i})\big) \end{aligned} \tag{20}$$

The terms from the equation above are already defined in equations (2), (3) and (4). We can express them in the matrix notation

$$\begin{aligned} p_i &= \frac{exp(X_i^T \beta)}{1 + exp(X_i^T \beta)} \\ 1 - p_i &= \frac{1}{1 + exp(X_i^T \beta)} \\ log(\frac{p_i}{1 - p_i}) &= X_i^T \beta \end{aligned}$$

and rewrite the log-likelihood equation as

$$L(\beta) = \sum_{i=1}^{n} \big(y_i(X_i^T \beta) - log(1 + exp(X_i^T \beta)\big) \tag{21}$$

In order to maximize the equation above we compute the score function which we will denote as $U(\beta)$

$$U(\beta) = \begin{pmatrix} \partial L(\beta)/\partial \beta_0 \\ \partial L(\beta)/\partial \beta_1 \\ \vdots \\ \partial L(\beta)/\partial \beta_j \end{pmatrix}$$

and proceed solving the equations by setting

$$U(\beta) = 0.$$

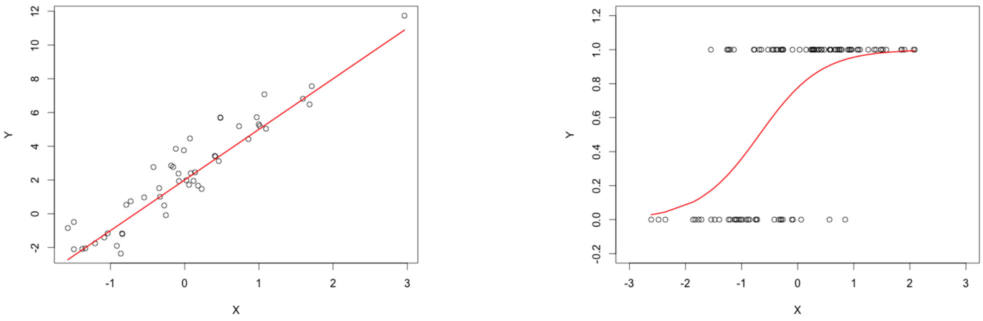
There is no closed-form solution to this system which means that we apply another technique in order to estimate coefficients β. MLE estimators are found through an iterative procedure such as Newton-Raphson algorithm, which we describe in the APPENDIX (8.2).

3.2 Estimation of average marginal effects

We have already showed that a logistic regression estimator is affected by omitted variable from the latent model and unobserved variance (heterogeneity). The fact that the variance is fixed in the latent variable causes a problem in interpretation of logit coefficients across different samples, or groups within samples, without making assumption of constant unobserved heterogeneity. Further, given the non-linear nature of logistic regression, the estimated coefficients can not be interpreted as marginal effects as it is the case in OLS. For various analytical purposes we would like to be able to find a way around these problems. In literature we find several kinds of approach to this problem, some of them considering probability changes (Mood, 2010).

We focus on the latter approach, that is, measure of probability changes per unit-change in a predictor variable. As noted before, logistic regression yields coefficients that are more complex in comparison to OLS, due to non-linear nature of logistic distribution. Figure 2 illustrates this difference in a very simple manner. Panel (a) shows straight line coming from a simple regression model and panel (b) shows sigmoid line which is result of a logistic regression.

Figure 2: Difference between simple and logistic regression lines



(a) Simple regression

(b) Logistic regression

While the line of in the panel (a) has a constant slope, which allows us to interpret coefficients from a linear regression as a marginal (and partial) effects, slope of the line in the panel (b) is changing along the x-axis.Similarity with OLS is that we have to rely on calculus in order to find partial effects. We find partial effect of continuous variables in logistic regression by taking the partial derivative (Wooldridge, 2009)

$$\frac{\partial p(X_i)}{\partial x_j} = f(X_i^T \beta)\beta_j, \quad \text{where} \;\; f(z) = \frac{dF}{dz}(z). \tag{22}$$

Since we defined $p(X)$ in Equation 2 as a logistic cumulative distribution $F(X)$, Equation 22 shows that we can obtain partial derivative of a continuous variable in the logistic regression by multiplying it's coefficient with a probability density function derived from our model equation. Figure 2 (b) shows graph of $F(X)$ and as we see it is strictly increasing. Difference in comparison to OLS coefficient estimates is that the partial effect of x_j on $p(X_i)$ depends on X_i throughout calculated $F(X)$ (but always has the same sign as β_j). In other words, since the calculation of the derivative (partial effect) results in different value for each X_i, logit coefficients do not represent marginal effects. According to Wooldridge (2009), for continuous variable x_j it can be shown that

$$\Delta\hat{P}(y=1|X) \approx [f(X\hat{\beta})\hat{\beta}_j]\Delta x_j, \tag{23}$$

for small changes in x_j. If $\Delta x_j = 1$, the change in estimated probability will be approximately $f(X\hat{\beta})\hat{\beta}_j$. In literature we find several operations that can be done on logit coefficients in order to obtain marginal effects. In this paper we focus on average marginal effect (AME) of continuous variables. If x_j is a continuous variable, it's average marginal effect is defined as (Wooldridge, 2009)

$$AME_j = \frac{1}{n}\sum_{i=1}^{n}[f(X_i^T\hat{\beta})\hat{\beta}_j] = \Big[\frac{1}{n}\sum_{i=1}^{n}f(X_i^T\hat{\beta})\Big]\hat{\beta}_j, \tag{24}$$

where $f(X_i^T\hat{\beta})$ is i:th evaluation of logistic probability density function of our model,

$$f(X_i^T\hat{\beta}) = \frac{exp(X_i^T\hat{\beta}))}{(1+exp(X_i^T\hat{\beta}))^2}. \tag{25}$$

Simply, Equation 25 shows that we obtain the average marginal effect by plugging our model equation into logistic probability density function and evaluating it across the sample. The outcome is then divided by the sample size n and multiplied by a coefficient β_j. The final result is a single value (for the coefficient j of interest) and represents the average marginal change in probability $p(X)$ for a small unit change in x_j.

3.3 Variance of average marginal effect

In this paper we investigate variance of the average marginal effects for the full and the curtailed models, throughout different sample sizes. For that purpose we estimate the mean squared error for average marginal effect, which we call MSE_{AME}. Our estimate is derived from Cramer's (2007) average sample effect, ASE, defined as

$$ASE_j = \frac{1}{n}\sum_{i=1}^{n}\hat{P}_i(1-\hat{P}_i)\hat{\beta}_j, \tag{26}$$

which represents the partial derivative of the expected sample frequency with respect to explanatory variable x_j. $\hat{P}_i$ denotes the probability of observing some value of y_i given predictors X_i and x_j refers to the variable whose coefficient $\hat{\beta}_j$ is plugged into Equation 26). This probability is obtained using logistic regression with the latent variable as a response. Equation 26 is actually another way of writing Equation 28, therefore $ASE_j = AME_j$.

In order to calculate variance for AME we need to find a reference value for comparison. We use Equation 26 to derive what we call AME_T, the true value for average marginal effect. AME_T is obtained by plugging in the theoretical true values for $\hat{P}_i$ and $\hat{\beta}_j$. Instead of $\hat{\beta}_j$ we simply use the coefficient β that we defined as true, while $\hat{P}_i$ values are replaced by the expectation of P, which is $\bar{P} = 0.5$. In other words, we calculate the AME_T as follows

$$AME_T = 0.5(1 - 0.5)\beta = 0.25\beta. \tag{27}$$

Finally, we estimate MSE_{AME} by calculating deviation of the obtained AME estimates from their *true* value

$$MSE_{AME} = \frac{1}{R}\sum_{k=1}^{R}(A\hat{M}E_k - AME_T)^2, \tag{28}$$

where R denotes number of replications in the simulation, $A\hat{M}E_k$ is an estimate obtained using the Equation 24 and AME_T is obtained using Equation 27.

4 Empirical analysis

4.1 Design of the experiment

In this section we give a brief description of our simulation of the latent variable model. In order to make explanations as clear as possible, we divide the simulation into four main parts and describe them in the following steps.

Step 1:

We generate random data and specify the latent variable as

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \tag{29}$$

where we set $\beta_0 = 0$, $\beta_1 = 1$ and for β_2 we choose values $0.5, 1, 2$ throughout simulations. Predictors x_1, x_2 are random variables normally distributed with mean $\mu = 0$ and variance $\sigma^2 = 1$, while the error term ϵ is a vector of random variables that follow logistic distribution with mean $\mu = 0$ and variance $\sigma^2 = \frac{\pi^2}{3}$.

Step 2:

Using Equation 6 we create binary [0,1] vector from the latent variable model in Equation 29. Next, we set this binary vector as a response variable in two logistic regression models, full and curtailed, that we use to estimate the latent variable. The full model is

$$y^F = \beta_0^F + \beta_1^F x_1 + \beta_2^F x_2 + \epsilon, \tag{30}$$

where we estimate all parameters from the latent variable so there is no miss-specification. The curtailed model is

$$y^C = \beta_0^C + \beta_1^C x_1 + \epsilon. \tag{31}$$

where we omit the second predictor, x_2, from Equation 30. This simulates miss-specification and enables us to study heterogeneity.

Step 3:

We extract coefficients from the models in Equations 30 and 31 and by plugging them into Equation 24 we obtain AME estimators, both for the full and the curtailed model. Steps 1,2, and 3 are repeated in $R = 1000$ replicates, resulting in 1000 estimates of extracted β_1 and AME for the full and the curtailed model. In addition, we calculate the theoretically expected value of the $\hat{\beta_1}$ from the curtailed model, as defined in Equation 17.

Step 4:

Having obtained 1000 estimates $A\hat{M}E_j$ we can calculate MSE_{AME} as we described in Equation 25, both for the full and the curtailed model.

The four steps above describe how our simulation works for given sample size and value of β_2 in Equation 29. As noted in the Step 1, for this paper we investigate three values of the true coefficient β_2; 0.5,1,2. For each of these values we run six simulations (Steps 1 to 4), using sample sizes $n = 100, 200, 500, 1000, 1500, 3000$. The simulation is written in the statistical software R and the description above is closely related to our algorithm. The complete code is in the APPENDIX.

4.2 Application on dataset

Besides simulation study described in the previous section, we study the effect of unobserved heterogeneity on MROZ dataset. Wooldridge (2014) uses this dataset for the purpose of studying logit models. We take the model studied by Wooldridge as a hypothetical true underlying model, and consider it's coefficients as true ones. Using this model as an unobserved latent process with it's true coefficients, we obtain the outcome as described in Equation 5. Using Equation 6 we transform this outcome into binary response variable which we later use as the dependent variable in the logistic regression, with purpose of estimating the initial outcome and coefficients of the model that we described as true. The result is presented in the section 5.2.

5 Results and discussion

5.1 Experiment results

Here we describe the result of the experiment. Table 1 shows the result from simulation of the Model 1 where the true coefficients in the latent variable are $\beta_0 = 0, \beta_1 = 1, \beta_2 = 0.5$.

We notice a slightly decreased $\hat{\beta_1}$ coefficient (on average) in the reduced model. Section in the table where we measure difference among estimates tells us that on average $\hat{\beta_1}$ in the curtailed model decreased by approximately 0.05 (from ca 1.01 to ca 0.96). This is quite close to the theoretically predicted value of $\hat{\beta_1^L}$ (0.964) calculated using Equation 17. On the other hand the AME estimators don't show a substantial change. In the both models they are around 0.199 regardless of a sample size. While there is no change in MSE_{AME} between the full and the

Table 1: Model with true coefficients $\beta_0 = 0, \beta_1 = 1, \beta_2 = 0.5$

Sample size	100	200	500	1000	1500	3000
Full model						
$\hat{\beta}_1$	1.0593	1.0161	1.0092	1.0099	1.0000	1.0013
AME_{β_1}	0.2002	0.1983	0.1992	0.1999	0.1986	0.1992
MSE_{AME}	0.0040	0.0034	0.0028	0.0026	0.0027	0.0026
Reduced model						
$\hat{\beta}_1$	0.9936	0.9623	0.9566	0.9584	0.9487	0.9504
AME_{β_1}	0.1998	0.1984	0.1992	0.1999	0.1986	0.1991
MSE_{AME}	0.0040	0.0034	0.0028	0.0026	0.0027	0.0026
Difference among estimates						
$\Delta\hat{\beta}_1$	-0.0657	-0.0538	-0.0526	-0.0515	-0.0513	-0.0508
ΔAME_{β_1}	-0.0004	0.0001	0.0000	0.0000	0.0000	0.0001
ΔMSE_{AME}	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Theoretical unbiased estimator						
$\hat{\beta}_1^L$	0.9641	0.9641	0.9641	0.9640	0.9640	0.9640

Presented result is the mean value after 1000 replications

curtailed model, we see that the variance tends to decrease with larger sample size. Scale of this decrease is largest between sample size of $n = 100$ and $n = 500$.

In the Table 2 we analyze the Model 2, the true coefficients in the latent variable are $\beta_0 = 0, \beta_1 = 1, \beta_2 = 1$. We notice that the change in the magnitude of $\hat{\beta}_1$ in the curtailed model has increased. On average $\hat{\beta}_1$ in the curtailed model is now around 0.84, which is quite smaller in comparison to the Table 1. This difference reduces with a larger sample size, most obvious from $n = 100$ to $n = 500$. Value of the theoretical estimator $\hat{\beta}_1^L$ is now about 0.87 for all sample sizes. AME ad MSE_{AME} show no change between the full and the curtailed model, however MSE_{AME} decreases with larger sample size.

Table 3 shows result from the third simulation, where we analyze Model 3. The true coefficients in the latent variable are $\beta_0 = 0, \beta_1 = 1, \beta_2 = 2$. On average $\hat{\beta}_1$ is much smaller in the curtailed model, with the value of 0.64 when the sample size is $n = 100$ and around 0.62 for larger sample sizes. It is somewhat lower than it's calculated theoretical value, $\hat{\beta}_1^L = 0.671$, but it follows decreasing pattern. On the other hand AME estimators don't show substantial change between the models, with estimated value of about 0.141. Just as in previous tables, there is no change in MSE_{AME} between the full and the curtailed model, but it decreases with a larger sample size.

Comparing these estimators across the tables reveals some interesting facts. First, the most obvious is the change in estimated $\hat{\beta}_1$ which in the curtailed models decreases from roughly 0.95 to 0.62. This clearly shows that the omitted variable in the latent model causes bias which increases with larger magnitude of the omitted variable coefficient. Second, although the AME estimators do not seem to be affected by omitted variable, since we do not see any substantial difference between the full and the curtailed equation within a particular model, they do change

Table 2: Model with true coefficients $\beta_0 = 0, \beta_1 = 1, \beta_2 = 1$

Sample size	100	200	500	1000	1500	3000
Full model						
$\hat{\beta_1}$	1.0511	1.0279	1.0123	1.0012	1.0052	1.0003
AME_{β_1}	0.1809	0.1806	0.1816	0.1811	0.1818	0.1813
MSE_{AME}	0.0063	0.0055	0.0049	0.0048	0.0047	0.0047
Reduced model						
$\hat{\beta_1}$	0.8661	0.8536	0.8450	0.8381	0.8423	0.8367
AME_{β_1}	0.1804	0.1813	0.1817	0.1810	0.1819	0.1812
MSE_{AME}	0.0065	0.0055	0.0050	0.0049	0.0047	0.0047
Difference among estimates						
$\Delta\hat{\beta_1}$	-0.1850	-0.1742	-0.1673	-0.1631	-0.1629	-0.1635
ΔAME_{β_1}	-0.0005	0.0007	0.0001	-0.0001	0.0000	-0.0001
ΔMSE_{AME}	0.0002	0.0000	0.0001	0.0000	0.0000	0.0000
Theoretical unbiased estimator						
$\hat{\beta_1^L}$	0.8758	0.8763	0.8757	0.8759	0.8756	0.8757

Presented result is the mean value after 1000 replications

as the magnitude of the omitted variable coefficient changes. In the first model we have a true $\beta_2 = 0.5$ and $AME = 0.19$, in the second model the $\beta_2 = 1$ and AME decreases to 0.18, while in the third model the true $\beta_2 = 2$ and AME is around 0.14. However, this difference is much smaller than the one that we observe for $\hat{\beta_1}$. Third, variance of the average marginal effect,MSE_{AME}, also does not appear to be effect by omitted variable within a particular and it decreases with a larger sample. However, just as AME - it generally increases with larger magnitude of the omitted variable, being substantially higher in the third model, presented in Table 3.

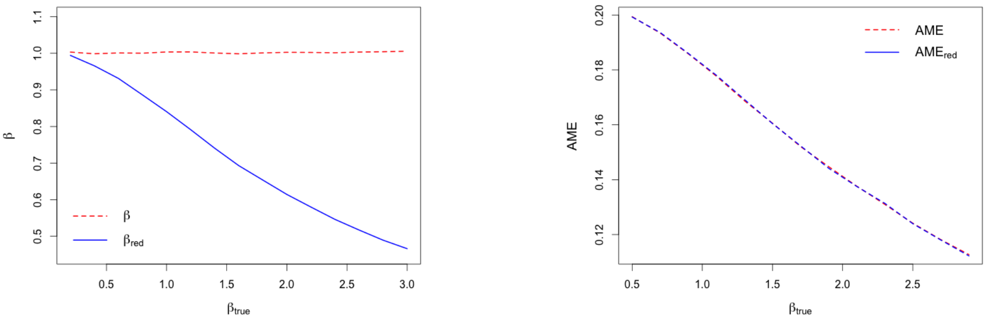
In order to have better understanding of changes in MSE_{AME}, we run more simulations with additional magnitudes of the true β_2 in the latent variable. We set the true β_2 to 0.5 and with each new simulation we increase it by 0.1 until we reach $\beta_2 = 3$. Result is the same type of data that we presented in Tables that covers larger range of magnitudes of β_2. This should further clarify certain relationships among estimators presented in the tables. We display the result graphically in Figure 3 and 4. Panel (a) in the Figure 3 shows estimated β_1 coefficients as a function of the true β_2 in the latent variable. The red dashed line shows the logit coefficient in the full model, which is constantly around 1. The blue line displays behaviour of the β_1 in the reduced model. As we saw in the Tables, there is a downward bias on this coefficient as the magnitude of the true β_2 in the latent variable increases. Panel (b) shows that there is no difference in AME for the full and reduced (AME_{red}) model. It also decreases with increased magnitude of the true latent β_2. While it is no surprise that AME_{red} decreases (since we see decrease in the estimated β_1 for the reduced model) - it could seem counter intuitive that the AME for the full model decreases despite the practically constant β_1 ifor the full model. This means that the averaged logistic density function (evaluated for all observations in the sample)

Table 3: Model with true coefficients $\beta_0 = 0, \beta_1 = 1, \beta_2 = 2$

Sample size	100	200	500	1000	1500	3000
Full model						
$\hat{\beta}_1$	1.0625	1.0302	1.0126	1.008	1.0027	1.0002
AME_{β_1}	0.1399	0.1416	0.1415	0.1414	0.1412	0.1409
MSE_{AME}	0.0135	0.0124	0.0120	0.0119	0.0119	0.0119
Reduced model						
$\hat{\beta}_1$	0.6456	0.6288	0.6180	0.6197	0.6135	0.6137
AME_{β_1}	0.1424	0.1415	0.1411	0.1418	0.1407	0.1410
MSE_{AME}	0.0136	0.0127	0.0121	0.0119	0.0120	0.0119
Difference among estimates						
$\Delta\hat{\beta}_1$	-0.4168	-0.4014	-0.3946	-0.3886	-0.3891	-0.3865
ΔAME_{β_1}	0.0024	-0.0001	-0.0003	0.0004	-0.0004	0.0001
ΔMSE_{AME}	0.0001	0.0003	0.0001	0.0000	0.0001	0.0000
Theoretical unbiased estimator						
$\hat{\beta}_1^L$	0.6723	0.6728	0.6722	0.6717	0.6715	0.6719

Presented result is the mean value after 1000 replications

Figure 3: estimated logit coefficients and AME as a function of the true β_2



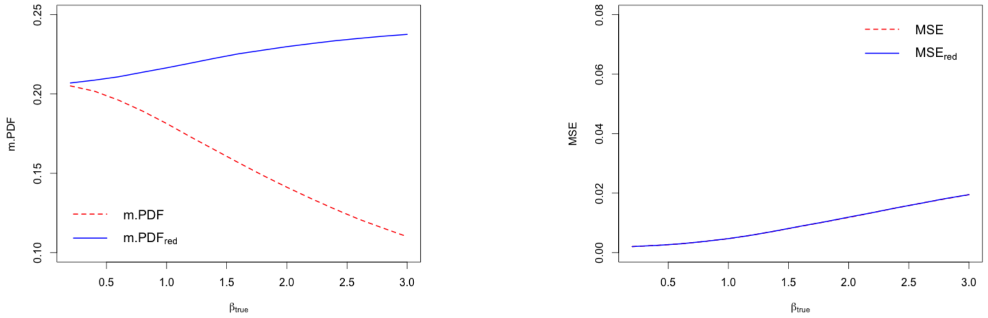
(a) estimated β_1 for the full and reduced model

(b) AME for the full and reduced model

has to differ between full and reduced model. This is shown in the Figure 4.

In the panel (a) we see mean of evaluated logistic function, denoted as $m.PDF$ where $m.PDF_{red}$ stands for the curtailed model. The red dashed line represents behaviour of the $m.PDF$ for the full model, and we see that it decreases, as opposed to increasing $m.PDF_{red}$, along the different magnitudes of the true latent β_2. Since the AME consists of both $m.PDF$ and logit coefficient β_1, as defined in the Equation 24, downward bias in the β_1 for the reduced model is affected by increased $m.PDF_{red}$ while the constant β_1 for the reduced model is affected

Figure 4: Mean PDF and MSE as a function of the true β_2



(a) mean PDF for the full and the curtailed model

(b) MSE for the full and the curtailed model

by decreasing $m.PDF$. It occurs at the same rate, which results in the practically same value of AME for the full and the curtailed model. This also explains their practically identical variance, MSE_{AME}, shown in the panel (b). Red dashed line denotes MSE_{AME} for the full model, and blue line is MSE_{AMEred} for the curtailed model. We see that there is no change between MSE_{AME} and MSE_{AMEred}, and they seem to increase along increasing true latent β_2.

5.2 Real data study

As noted in the section 4.2., for this experiment we use the MROZ dataset which is collected with purpose of investigating women's participation in labor force. We take the model studied by Wooldridge and use it as a true underlying model. Wooldridge uses logistic regression to predict the outcome of binary variable *inlf*, where 1 denotes woman's participation in the labor force. In Table 4 we show explanatory variables with corresponding coefficients, obtained from the Wooldridge's logit model. Using them we build a linear model, as in Equation 5, and obtain an outcome which is transformed into binary variable as described in Equation 6. Knowing the hypothetical true model, we introduce different levels of misspecification and study the effect of introduced unobserved heterogeneity. The purpose of this modeling is just to give an illustration of how the unobserved heterogeneity affects logit coefficients, not to investigate or replicate the original study on women's participation in labor force.

Table 4: Hypothetical true model

Variables	*nwifeinc*	*educ*	*exper*	*expersq*	*age*	*kidslt6*	*kidsge6*
Coefficients	-0.021	0.221	0.206	-0.0031	-0.088	-1.443	0.06

As we saw earlier, unobserved heterogeneity affects the misclassified model even when the omitted variable is uncorrelated with the ones that are present in a model. It is therefore good to examine correlation among the variables in the hypothetical true model. Table 5 displays correlation coefficients for all the variables. We see that there is no strong correlation among any pair of variables (with exception of *exper* and *expersq*, since the latter variable is a square of the former).

Table 5: Hypothetical true model: correlation table

Variables	*nwifeinc*	*educ*	*exper*	*expersq*	*age*	*kidslt6*	*kidsge6*
nwifeinc	1.000	0.278	-0.172	-0.165	0.059	0.038	0.025
educ	0.278	1.000	0.066	0.024	-0.120	0.109	-0.059
exper	-0.172	0.066	1.000	0.938	0.334	-0.194	-0.300
expersq	-0.165	0.024	0.938	1.000	0.380	-0.184	-0.300
age	0.059	-0.120	0.334	0.380	1.000	-0.434	-0.385
kidslt6	0.038	0.109	-0.194	-0.184	-0.434	1.000	0.084
kidsge6	0.025	-0.059	-0.300	-0.300	-0.385	0.084	1.000

Let us assume that we know all the relevant variables and we are able to model logistic regression without misspecification. Table 6 shows the result of such model and we see that the coefficients are quite similar to the "true" ones. In addition, under the coefficients we show average marginal effect, calculated as described in Equation 24. We will use these coefficients and AME estimates as a reference and compare them to misspecified models.

Table 6: Estimation of the true model, no misspecification

Variables	*nwifeinc*	*educ*	*exper*	*expersq*	*age*	*kidslt6*	*kidsge6*
Coefficients	-0.023	0.221	0.214	-0.003	-0.089	-1.363	0.118
AME	-0.004	0.038	0.037	0.00	-0.015	-0.237	0.021

Table 7 shows coefficients and AME estimates for our first misspecified model, where we exclude variables *kidslt6* and *kidsge6*. While the coefficient estimates for the variables *nwifeinc*, *exper* and *expersq* do not change much, we see more noticable change for variables *educ* and *age*. On the other hand, just as we saw in the simulation study, AME estimates do not show substantial change.

Table 7: Misspecified model 1: *kidslt6* and *kidsge6* excluded

Variables	*nwifeinc*	*educ*	*exper*	*expersq*	*age*
Coefficients	-0.02	0.177	0.21	-0.0029	-0.053
Δ Coefficients	0.003	-0.044	-0.004	0.000	0.036
AME	-0.004	0.034	0.04	-0.001	-0.01
ΔAME	0.000	-0.004	0.003	-0.001	0.005

We try another misspecified model, this time excluding variables *exper* and *expersq*. Table 8 displays the estimated coefficients and we see noticable change in all other estimates, the largest being for the variable *kidsge6* whose coefficient even changed a sign. On the other hand, change in the AME estimates is not so large.

Table 8: Misspecified model 2: *exper* and *expersq* excluded

Variables	*nwifeinc*	*educ*	*age*	*kidslt6*	*kidsge6*
Coefficients	-0.038	0.257	-0.051	-1.393	-0.036
Δ Coefficients	-0.015	0.036	0.038	-0.030	-0.154
AME	-0.008	0.054	-0.011	-0.294	-0.008
ΔAME	-0.004	0.016	0.004	-0.057	-0.029

In our last misspecified model we omit all the variables except for *nwifeinc*, *educ* and *age*. Table 9 shows coefficient and AME estimates. While there is not much change in the coefficients for *nwifeinc* and *educ*, the one for *age* deviates much more from the true coefficient. Again, AME estimates do not show any substantial change from the true ones.

Table 9: Misspecified model 3: *kidslt6*, *kidsge6*,*exper* and *expersq* excluded

Variables	*nwifeinc*	*educ*	*age*
Coefficients	-0.037	0.225	-0.010
Δ Coefficients	-0.014	0.004	0.079
AME	-0.008	0.052	-0.002
ΔAME	-0.004	0.014	0.013

6 Conclusion

Logistic regression, although robust in terms of assumptions, is a complex method that should be used with a care. If the omitted variable effect is ignored, it is easy to misunderstand and misreport logistic regression estimates. In Equation 17 we see that logit coefficients depend both on effect size and the magnitude of unobserved heterogeneity, meaning that we can not interpret and compare them as we usually do in linear regression. There is no unique solution to the unobserved heterogeneity problem in logistic regression. It is therefore important to be aware of these problems and already at data collection stage find enough information on variables that could affect the outcome - even if they are weakly related to the existing independent variables in our model (Mood,2009). Our simulation shows that the AME estimates in the curtailed model, although to some extent affected by the magnitude of the omitted variable, are not biased and within the same models their variance is not much affected. They provide good estimator of average marginal effect of continuous variables, and can be used in addition to logistic regression coefficients especially if a researcher is mainly interested in the sign instead of size of an effect. Our experiment with the MROZ dataset closely follows conclusions derived from the simulation - estimated logit coefficients among misspecified models differ enough to be incomparable across the different models. AME estimates mostly change by substantially lower magnitude and their comparison is therefore more consistent.

7 References

Literature

Harrel,Jr., Frank E. (2001) *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis.* New York: Springer.

Hastie,T., Tibshirani, R. and Friedman, J. (2003) *The elements of statistical learning; data mining, inference and prediction.* New York: Springer.

James, G., Witten, D., Hastie,T. and Tibshirani, R. (2013) *An introduction to statistical learning; with applications in R.* New York: Springer.

McCulloh,C.E. and Searle,S.R.(2001) *Generalized,linear and mixed models.* New York: John Wiley & Sons

Wooldridge, J.M. (2009) *Introductory econometrics.* 4.ed. Mason,Ohio: South-Western Cengage Learning.

Articles

Cramer, J.S. (2006) *Robustness of logit analysis: unobserved heterogeneity and misspecified disturbances.* Discussion paper, Amsterdam School of Economics

Karlson, K. B., Holm, A. and Breen, R. (2010) *Comparing regression coefficients between models using logit and probit: a new method.* Working paper series, Centre for Strategic Research in Education.

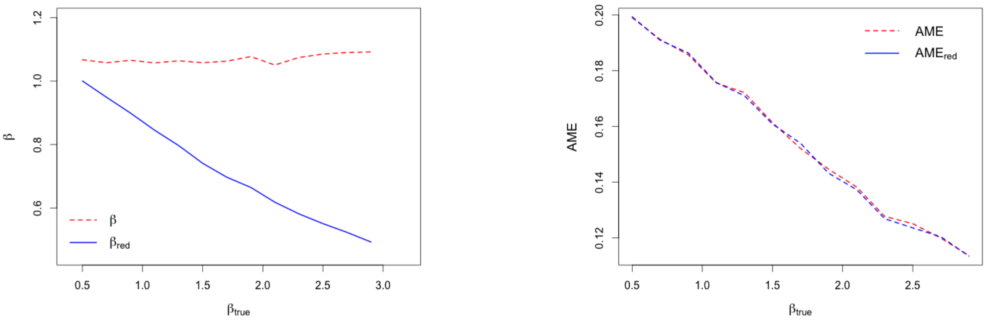
Mood,Carina (2009) *Logistic regression: why we cannot do what we think we can do, and what we can do about it.* European Sociological Review vol.26, p.67-82.

8 Appendix

8.1 Simulation on different sample sizes

Figures 4 and 5 from the section 4 show the result of repeated simulation, where we plot logistic regression β_1 coefficients, APE, MSE_{APE} and estimated logistic PDF as a function of the omitted true β_2 latent coefficient, for both full and reduced model. In this section we repeat the simulation using different sample sizes and investigate changes in these estimates. We perform four additional experiments, using sample sizes $n = 100$,200,500 and 1000.

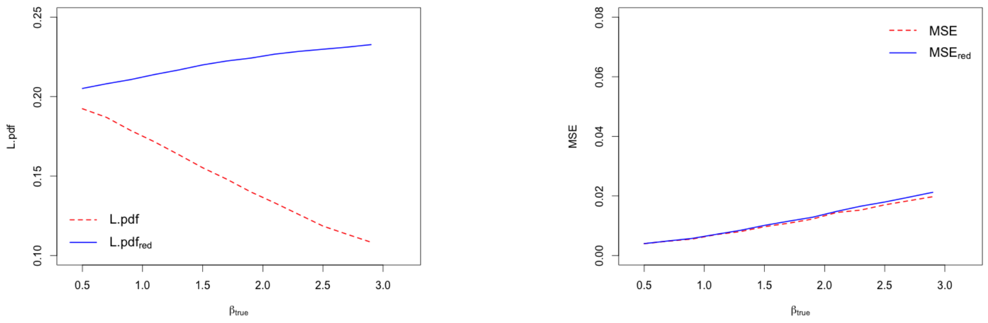
Figure 5: Estimated logit coefficients and AME as a function of the omitted true latent β_2, sample size is $n = 100$



(a) Estimated β_1 for the full and reduced model

(b) AME for the full and reduced model

Figure 6: Logistic PDF and estimated MSE as a function of the omitted true latent β_2, sample size is $n = 100$

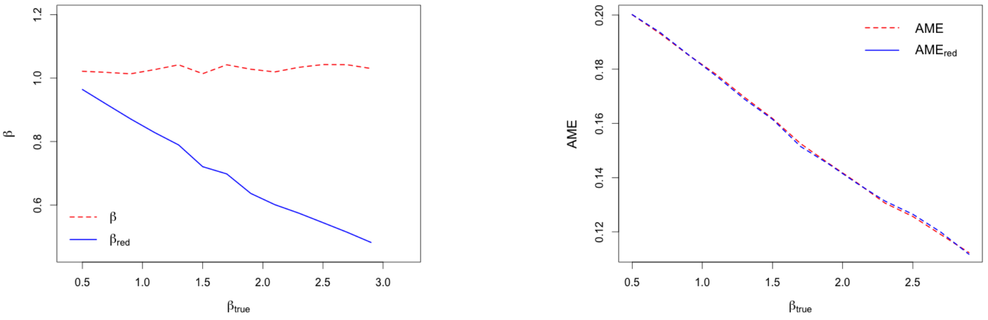


(a) Logistic PDF for the full and reduced model

(b) MSE for the full and reduced model

Figures 5 and 6 look very similar to the first test where we generated a sample size $n = 3000$. We proceed with the sample size $n = 200$.

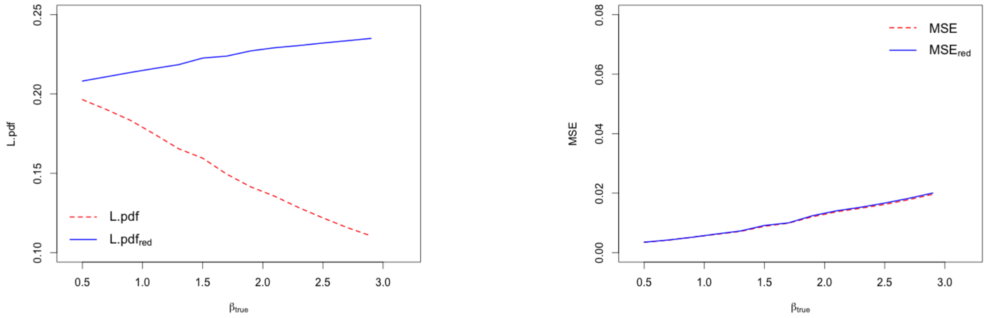
Figure 7: Estimated logit coefficients and AME as a function of the omitted true latent β_2, sample size is $n = 200$



(a) Estimated β_1 for the full and reduced model

(b) AME for the full and reduced model

Figure 8: Logistic PDF and estimated MSE as a function of the omitted true latent β_2, sample size is $n = 200$

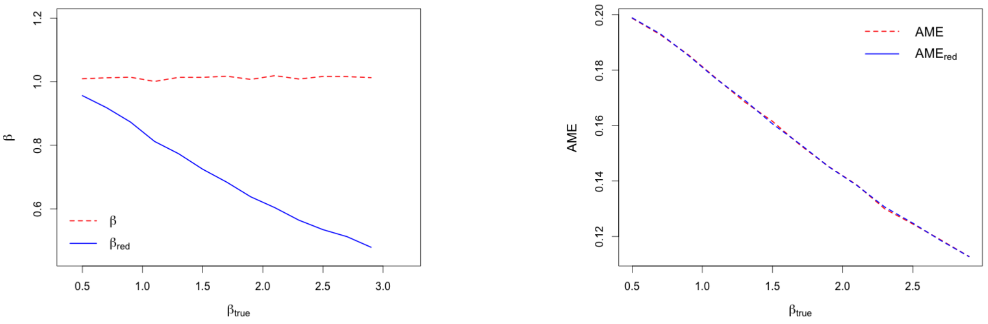


(a) Logistic PDF for the full and reduced model

(b) MSE for the full and reduced model

We can say that the result in the Figures 7 and 8 is the same, so far we haven't seen a change in the estimators' behaviour given a larger sample size.

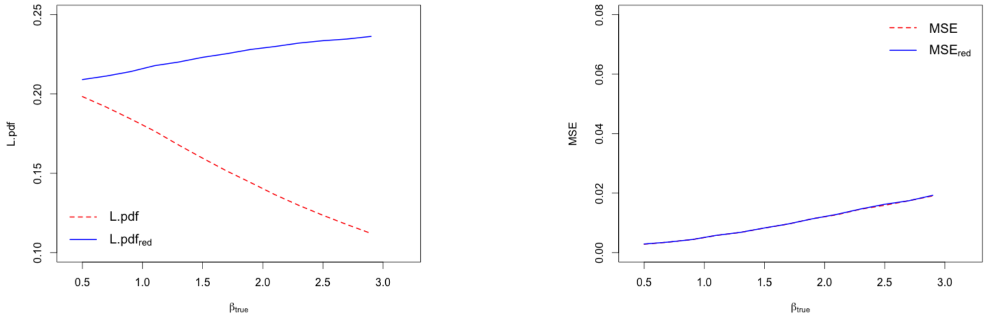
Figure 9: Estimated logit coefficients and AME as a function of the omitted true latent β_2, sample size is $n = 500$



(a) Estimated β_1 for the full and reduced model

(b) AME for the full and reduced model

Figure 10: Logistic PDF and estimated MSE as a function of the omitted true latent β_2, sample size is $n = 500$



(a) Logistic PDF for the full and reduced model

(b) MSE for the full and reduced model

As we see, there is no essential change in the estimators of our interest, they seem to keep the same behaviour no matter what sample size we choose.

8.2 Newton-Raphson algorithm

In this section we explain the Newton-Raphson algorithm in greater detail. One way of illustrating the idea behind this algorithm is Taylor expansions. We assume that a function f is smooth with an interior minimum w^* (derivative at w^* is zero and the second derivative is positive). Further, we can assume a start point w_0 near the minimum and take a second order Taylor expansion around w_0 :

$$f(w) \approx f(w_0) + (w - w_0)\frac{df}{dw}\bigg|_{w=w_0} + \frac{1}{2}(w - w_0)^2\frac{d^2f}{dw^2}\bigg|_{w=w_0} \tag{32}$$

The Newton-Rhapson method is relying on minimization of the equation above, by taking the derivative with respect to w and setting it equal to zero at a point w_1

$$0 = f'(w_0) + \frac{1}{2}f''(w_0)2(w_1 - w_0) \tag{33}$$

$$w_1 = w_0 - \frac{f'(w_0)}{f''(w_0)} \tag{34}$$

where the point w_1 should be a better approximation of the minimum w^* than the initial point w_0. By iterating this procedure each minimized approximation is used to get a new approximation until the true minimum is reached.

Let us see how the Newton-Rhapson method is applied to our problem. If we take a look at the equation (21) we can see that $L(\beta)$ depends on the elements of β only through the values of X which is linear, so each of the partial derivatives in the $U(\beta)$ will have the same form.

First we define all terms needed for performing the Newton-Raphson iterative algorithm:

$$\partial L(\beta)/\partial\beta_j = \sum_{i=1}^{n}\big(y_i\frac{\partial}{\partial\beta_j}(X_i^T\beta) - \frac{\partial}{\partial\beta_j}log(1 + exp(X_i^T\beta)\big) \tag{35}$$

where

$$\frac{\partial}{\partial\beta_j}(X_i^T\beta) = x_{ij}$$

and

$$\frac{\partial}{\partial\beta_j}log(1 + exp(X_i^T\beta)) = \frac{exp(X_i^T\beta)}{1 + exp(X_i^T\beta)}\frac{\partial}{\partial\beta_j}X_i^T\beta = p_i x_{ij}$$

so that

$$\frac{\partial L(\beta)}{\partial\beta_j} = \sum_{i=1}^{n}\big(x_{ij}(y_i - p_i)\big).$$

For Newton-Raphson method we will also need the second partial derivatives

$$\frac{\partial^2 l}{\partial\beta_j\partial\beta_k} = \sum_{i=1}^{n}\Big[x_{ij}(y_i - \frac{\partial p_i(\beta)}{\partial\beta_k})\Big].$$

Since

$$\frac{\partial p_i(\beta)}{\partial \beta_k}) = x_{ik} p_i(\beta)(1 - p_i(\beta)).$$

the second derivative is

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} \big[x_{ij} x_{ik} p_i (1 - p_i)\big].$$

In Bernoulli distribution variance is defined as $Var(y) = p(y)(1 - p(y))$ so we write the expresion above as

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} \big[x_{ij} x_{ik} v_i(\beta)\big],$$

which means that in matrix notation we can write the second derivative of the log-likelihood function with respect to coefficients as

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_k} = -X^T V(\beta) X,$$

where $V(\beta)$ denotes a diagonal matrix with coefficient variances. The term above is also called the information matrix.

Following our discussion where we explained principles of the Newton-Raphson method using Taylor expansions, the ith iteration will be

$$\hat{\beta}_{i+1} = \beta_i - (X^T V(\beta) X)^{-1} X^T (y - p),$$

until it converges to MLE$(\hat{\beta})$.

8.3 R-code

Note: the R-codes listed here are complete and can reproduce all simulations described in the paper. However, care should be taken in case of copy - pasting the code due to different font formats. If you use the code, make sure that it is carefully re-written (or copied) to a text format that is readable in R.

Function used to perform the first set of simulations

The "mySim" function generates the first set of simulations. The data in the Tables is obtained using it.

```
mySim <- function(Betas,samplesize,replicates,xout){

  #-------  re-define parameters -----
  R <- replicates
  n<- samplesize
  p<-length(Betas)-1    #number of variables without intercept
  B<-Betas
  Bout<-B[-(xout)]

  #------- create storage for loop

  Bhat   <- matrix(NA,R,length(B))
  Bhat.red <- matrix(NA,R,length(Bout))

  APE <- matrix(NA,R,length(B))
  APEred <- matrix(NA,R,length(Bout))

  deltaSE   <- matrix(NA,R,length(B))
  deltaSEred <- matrix(NA,R,length(Bout))

  ASE <- matrix(NA,R,length(B))
  ASEred <- matrix(NA,R,length(Bout))

  b.L <- matrix(NA,R,length(Bout))
  #-------- begin loop ---------

  for(t in 1:R){


    #-------  generate data  -----

    M<-replicate(p,rnorm(n,mean=0,sd=1))  #matrix with explanatory variables
    l.err<-rlogis(n)              #Logit error
    X <- cbind(1,M)            #data matrix, column with zeros & M
    Xout <- X[,-(xout)]

    #---- store variance-----
    v <- apply(X,2,var)
    b.L[t,] <- Bout*(sqrt(3.29)/(sqrt(3.29 + (B[xout]^2)*v[xout])))
    #------- configure  latent variable  -----

    Ylog <- X %*% B+l.err  #true latent variable

    #------- transform the latent variable into [0,1] vector -----

    LogLat <- rep(NA,n)
    for (j in 1:n){
      if(Ylog[j] > 0){ LogLat[j] = 1
      }else{              #condition, latent variable
        LogLat[j] = 0}
      }

    #--- use the [0,1] vector as a response variable in the logistic regression -
```

```
YHATlog <- glm(LogLat~X[,-1], family=binomial(link="logit"))
YHATlogOut <- glm(LogLat~Xout[,-1], family=binomial(link="logit"))

#-------- use logit-coefficients and obtain APE:s

logPDF <- dlogis(predict(YHATlog,type = "link"))
cof <- coef(YHATlog)
APEall <- mean(logPDF)*cof  # Wooldridge's average partial (marginal)  effect(s)

#-----store into storage-----
Bhat[t,] <-  cof
APE[t,] <- APEall
#--------------

logPDFout <- dlogis(predict(YHATlogOut,type = "link"))
cofOut <- coef(YHATlogOut)
APEout <- mean(logPDFout)*cofOut

#-----store into storage-----
Bhat.red[t,] <-  cofOut
APEred[t,] <- APEout
#----------------

#------ obtain Cramer's ASE
prob <- predict(YHATlog,type = "response")
probOut <- predict(YHATlogOut,type = "response")

ASE[t,]   <- mean(prob*(1-prob))*cof  #vector
ASEred[t,]   <- mean(probOut*(1-probOut))*cofOut
#------------------------------

#-------- estimate standard errors using
#---------   delta method, full model
z<-cof
pz <- length(z)

dAMEdB<-matrix(NA,pz,pz)   #this loop creates matrix dAMEdB where col=b,row = AME
for(i in 1:pz){
  dAMEdBi<-rep(NA,pz)
  for(j in 1:pz){
    c1<- z[i] # = cof1[i]   #value of coefficient (intercept = B_1,B_2,...B_n)
    m1<-predict(YHATlog,type = "link")
    H1 <- (-(2*c1*X[,j]*exp(2*m1))/((1+exp(m1))^3)+(c1*X[,j]*exp(m1))/((1+exp(m1))^2))
    dAMEdBi[j] <- mean(H1)
  }
  dAMEdB[i,]<-dAMEdBi
}

dAMEdBij <- matrix(NA,pz,pz)
for(i in 1:pz){
  m2<-predict(YHATlog,type = "link")
  H2 <- exp(m2)/((1+exp(m2))^2)
  dAMEdBij[i,i] <- dAMEdB[i,i] + mean(H2)
  dAMEdBij[-i,i] <- dAMEdB[-i,i]
}

deltaSEall <- sqrt(diag(dAMEdBij %*% vcov(YHATlog) %*% t(dAMEdBij)))

#-----store into storage-----
deltaSE[t,] <- deltaSEall
#-------------------------

#-------- estimate deltaSE, reduced model
```

```
  zOut<-cofOut
  pOut <- length(zOut)

  #----- standard errors, curtailed model

  dAMEdBOut<-matrix(NA,pOut,pOut) #this loop creates matrix dAMEdB where col=b,row = AME
  for(i in 1:pOut){
    dAMEdBiOut<-rep(NA,pOut)
    for(j in 1:pOut){
      c1Out<- zOut[i] # = cof1[i]   #value of coefficient (intercept = B_1,B_2,...B_n)
      m1Out<-predict(YHATlogOut,type = "link")
      H1Out <- (-(2*c1Out*Xout[,j]*exp(2*m1Out))/((1+exp(m1Out))^3)
                        + (c1Out*Xout[,j]*exp(m1Out))/((1+exp(m1Out))^2))
      dAMEdBiOut[j] <- mean(H1Out)
    }
    dAMEdBOut[i,]<-dAMEdBiOut
  }

  dAMEdBijOut <- matrix(NA,pOut,pOut)
  for(i in 1:pOut){
    m2Out<-predict(YHATlogOut,type = "link")
    H2Out <- exp(m2Out)/((1+exp(m2Out))^2)

    dAMEdBijOut[i,i] <- dAMEdBOut[i,i] + mean(H2Out)
    dAMEdBijOut[-i,i] <- dAMEdBOut[-i,i]
  }

  deltaSEout <- sqrt(diag(dAMEdBijOut %*% vcov(YHATlogOut) %*% t(dAMEdBijOut)))

  #----store into storage----
  deltaSEred[t,] <- deltaSEout
  #-------------------------
}

#-------- end loop --------
ASEtrue <- 0.25*B
ASEtrue.red <- 0.25*Bout


RSSase <- matrix(NA,dim(ASE)[1],dim(ASE)[2])
for(r in 1:dim(ASE)[2]){
  RSSase[,r] <- (ASE[,r] - ASEtrue[r])^2 }

RSSase.red <-matrix(NA,dim(ASEred)[1],dim(ASEred)[2])
for(rr in 1:dim(ASEred)[2]){
  RSSase.red[,rr] <- (ASEred[,rr] - ASEtrue.red[rr])^2}

MSEase <- apply(RSSase,2,sum)/R
MSEase.red <- apply(RSSase.red,2,sum)/R

#-------- define final result --------

mean.cof <- apply(Bhat,2,mean)
mean.cofOut <- apply(Bhat.red,2,mean)

mean.APE <- apply(APE,2,mean)
mean.APEred <- apply(APEred,2,mean)

mean.deltaSE <- apply(deltaSE,2,mean)
mean.deltaSEred <- apply(deltaSEred,2,mean)

mean.ASE <- apply(ASE,2,mean)
mean.ASEred <- apply(ASEred,2,mean)
```

```
  B.ch <-  mean.cofOut - mean.cof[-(xout)]
  APE.ch <- mean.APEred - mean.APE[-(xout)]
  deltaSE.ch <- mean.deltaSEred - mean.deltaSE[-(xout)]
  ASE.ch <- mean.ASEred - mean.ASE[-(xout)]
  MSE.ch <- MSEase.red - MSEase[-(xout)]

  beta.L <- apply(b.L,2,mean)


  result1 <-matrix(cbind(t(mean.cof),t(mean.APE),t(mean.deltaSE),t(mean.ASE),t(MSEase)),ncol=5)
  result2 <-matrix(cbind(t(mean.cofOut),t(mean.APEred), t(mean.deltaSEred),t(mean.ASEred),t(MSEase.red))
  result3 <-matrix(cbind(t(B.ch),t(APE.ch),t(deltaSE.ch),t(ASE.ch),t(MSE.ch)),ncol = 5)
  result4 <- matrix(beta.L[-1],ncol = 1)

  colnames(result1)<-c("mean.coef","mean.APE","mean.SE","mean.ASE","MSEase")
  colnames(result2)<-c("mean.coef.red","mean.APEred","mean.SEred","mean.ASEred","MSEase.red")
  colnames(result3)<-c("coef.dif","APE.dif","deltaSE.dif","ASE.dif","MSE.dif")
  colnames(result4) <- c("beta.L")
  return(list(result1,result2,result3,result4))

  #---------- end function ----------

}
```

Function used to generate the second set of simulations

The "mySimPlot" function generates the second set of simulations. The Figures 3 and 4 in section 4 are constructed using this function, including Figures 5 - 12 in the Appendix.

```
mySimPlot <- function(samplesize, replicates,K){

  true.b.2 <- K

  test.ASE <- matrix(NA,length(K),3)
  test.ASEred <- matrix(NA,length(K),2)

  test.MSEase <- matrix(NA,length(K),3)
  test.MSEase.red <- matrix(NA,length(K),2)

  test.Pr <- rep(NA,length(K))
  test.PrOut <- rep(NA,length(K))

  test.densty <- rep(NA,length(K))
  test.denstyR <- rep(NA,length(K))

  test.Bhat <- matrix(NA,length(K),3)
  test.Bhat.out <- matrix(NA,length(K),2)

  for(u in 1:length(K)){

    Betas <- c(0,1,K[u])
    xout <- 3

    #--------- re-define parameters ------
    R <- replicates
    n<- samplesize
    p<-length(Betas)-1   #number of variables without intercept
    B<-Betas
    Bout<-B[-(xout)]

    #--------- create storage for loop
```

```
ASE <- matrix(NA,R,length(B))
ASEred <- matrix(NA,R,length(Bout))

Bhat <- matrix(NA,R,length(B))
Bhat.out <- matrix(NA,R,length(Bout))

Pr <-  rep(NA,R)
PrOut <-  rep(NA,R)

m.densty <- rep(NA,R)
m.denstyR <- rep(NA,R)
#--------- begin loop ----------

for(t in 1:R){

  #--------  generate data  ------

  M<-replicate(p,rnorm(n,mean=0,sd=1))  #matrix with explanatory variables
  l.err<-rlogis(n)              #Logit error
  X <- cbind(1,M)           #data matrix, column with zeros & M
  Xout <- X[,-(xout)]

  #-------- configure  latent variable  ------

  Ylog <- X %*% B+l.err  #true latent variable

  #-------- transform the latent variable into [0,1] vector ------

  LogLat <- rep(NA,n)
  for (j in 1:n){
    if(Ylog[j] > 0){ LogLat[j] = 1
    }else{                  #condition, latent variable
      LogLat[j] = 0}
  }

  #-------- use the [0,1] vector as a response variable in the logistic regression ------

  YHATlog <- glm(LogLat~X[,-1], family=binomial(link="logit"), x = TRUE)
  YHATlogOut <- glm(LogLat~Xout[,-1], family=binomial(link="logit"),x = TRUE)

  cof <- coef(YHATlog)
  cofOut <- coef(YHATlogOut)

  #------ obtain Cramer's ASE
  prob <- predict(YHATlog,type = "response")  #gives probability, logistic CDF
  probOut <- predict(YHATlogOut,type = "response")

  m.densty[t] <- mean(dlogis(predict(YHATlog,type = "link")))
  m.denstyR[t] <- mean(dlogis(predict(YHATlogOut,type = "link")))

  ASE[t,]  <- mean(prob*(1-prob))*cof  #vector
  ASEred[t,]  <- mean(probOut*(1-probOut))*cofOut
  Pr[t] <- mean(prob)
  PrOut[t] <- mean(probOut)
  Bhat[t,] <- cof
  Bhat.out[t,] <- cofOut


  #--------- end 1 st loop ----------
  ASEtrue <- 0.25*B              #calculate MSE
  ASEtrue.red <- 0.25*Bout

  RSSase <- matrix(NA,dim(ASE)[1],dim(ASE)[2])
```

```
      for(i in 1:dim(ASE)[2]){
        RSSase[,i] <- (ASE[,i] - ASEtrue[i])^2 }

      RSSase.red <- matrix(NA,dim(ASEred)[1],dim(ASEred)[2])
      for(i in 1:dim(ASEred)[2]){
        RSSase.red[,i] <- (ASEred[,i] - ASEtrue.red[i])^2 }


      MSEase <- apply(RSSase,2,sum)/R
      MSEase.red <- apply(RSSase.red,2,sum)/R


      #---------- define result from the 1st loop----------

      mean.ASE <- apply(ASE,2,mean)
      mean.ASEred <- apply(ASEred,2,mean)

      mean.Bhat <- apply(Bhat,2,mean)
      mean.Bhat.out <- apply(Bhat.out,2,mean)

      #---------- store from te 2nd loop ----------

      test.ASE[u,]   <- mean.ASE
      test.ASEred[u,] <- mean.ASEred

      test.MSEase[u,] <- MSEase
      test.MSEase.red[u,] <- MSEase.red

      test.Pr[u] <- mean(Pr)
      test.PrOut[u] <- mean(PrOut)

      test.densty[u] <-mean(m.densty)
      test.denstyR[u] <- mean(m.denstyR)

      test.Bhat[u,] <- mean.Bhat
      test.Bhat.out[u,] <- mean.Bhat.out

    }

  }
  ape <- cbind(test.ASE[,2],test.ASEred[,2])
  colnames(ape) <- c("APE","APEred")


  MSEape <- cbind(test.MSEase[,2],test.MSEase.red[,2])
  colnames(MSEape) <- c("MSEape","MSEape.red")


  L.cdf <- cbind(test.Pr,test.PrOut)
  colnames(L.cdf) <- c("L.cdf","L.cdf.red")


  L.pdf <- cbind(test.densty,test.denstyR)
  colnames(L.pdf) <- c("L.pdf","L.pdf.red")


  beta.1 <- cbind(test.Bhat[,2],test.Bhat.out[,2])
  colnames(beta.1) <- c("beta.1","beta.1r")

  return(list(ape,MSEape,L.cdf,L.pdf,beta.1))
}
```

Note that the parameter K in the function is vector of the true beta.2 coefficients in the latent variable, that are to be omitted.

The following code generates figures:

```
#plot  beta.1 full/reduced model,   as function of omitted true beta.2
plot(K,a[[5]][,1],type = ”l”,col=”red”,lty=2,lwd = 2, ylab = expression(beta),
     xlab =  expression(beta[true]),cex.lab = 1.2,ylim = c(0.45,1.2),xlim = c(0.4,3.2))
lines(K,a[[5]][,2],type = ”l”,col=”blue”,lty=1,lwd=2)
legend( ”bottomleft”, c(expression(beta),expression(beta[red])),
        col = c(”red”,”blue”),lty = c(2, 1), lwd = 2,cex = 1.2, bty = ”n”)
#################################################################################################


#####  APE full/reduced model, as function of omitted true beta.2

plot(K,a[[1]][,1],type = ”l”,col=”red”,lty=2,lwd = 2,ylab = ”APE”,xlab =  expression(beta[true]),
     xlim = c(0.4,3.2),cex.lab = 1.2,ylim = c(0.1,0.22))
lines(K,a[[1]][,2],type = ”l”,col=”blue”,lty=2,lwd=2)
legend( ”topright”, c(expression(APE),expression(APE[red])),col = c(”red”,”blue”),lty = c(2, 1),
        lwd = 2,cex = 1.2, bty = ”n”)

#################################################################################################

#plot MSE full/reduced model,  as function of omitted true beta.2
plot(K,a[[2]][,1],type = ”l”,col=”red”, ylab = ”MSE”, xlab = expression(beta[true]),lty=2,lwd=2,
     ylim = c(0,0.08), xlim = c(0.4,3.2))
lines(K,a[[2]][,2],type = ”l”,col=”blue”,lty=1,lwd=2)
legend( ”topright”, c(expression(MSE), expression(MSE[red])), col = c(”red”,”blue”),
        lty = c(2, 1), lwd = 2,cex = 1.2, bty = ”n”)

#################################################################################################

# L.pdf full/reduced model, as function of omitted true beta.2
plot(K,a[[4]][,1],type = ”l”,col=”red”,lty=2,lwd=2, ylab = ”L.pdf”, xlab = expression(beta[true]),
     ylim = c(0.1,0.25),xlim = c(0.4,3.2))
lines(K,a[[4]][,2],type = ”l”,col=”blue”,lty=1,lwd=2)
legend( ”bottomleft”, c(expression(L.pdf), expression(L.pdf[red])), col = c(”red”,”blue”),
        lty = c(2, 1), lwd = 2,cex = 1.2, bty = ”n”)

#################################################################################################
```

Standalone function for estimating AME

```
ame <- function(o,dig = 3,bootSE=F,nboot){

  # ”o” is a glm object
  # ”dig” is number of digits in the output (default = 3)

  logPDF1 <- dlogis(predict(o,type = ”link”))
  cof1 <- coef(o)
  APE1 <- round(mean(logPDF1)*cof1,dig)

  z<-cof1
  p <- length(z)
  X <- o$x   # glm object ”o” should return design matrix (set ”x=TRUE”)

  #  deltaSE
  dAMEdB<-matrix(NA,p,p)  #this loop creates matrix dAMEdB where col = b, row = AME
  for(i in 1:p){
    dAMEdBi<-rep(NA,p)
     for(j in 1:p){
       c1<- z[i] # = cof1[i]   #value of coefficient
       m1<-predict(o,type = ”link”)
     H1 <- (-(2*c1*X[,j]*exp(2*m1))/((1+exp(m1))^3)+(c1*X[,j]*exp(m1))/((1+exp(m1))^2))
      dAMEdBi[j] <- mean(H1)
    }
    dAMEdB[i,]<-dAMEdBi
  }

  dAMEdBij <- matrix(NA,p,p)
  for(i in 1:p){
    m2<-predict(o,type = "link")
    H2 <- exp(m2)/((1+exp(m2))^2)

    dAMEdBij[i,i] <- dAMEdB[i,i] + mean(H2)
    dAMEdBij[-i,i] <- dAMEdB[-i,i]
  }

  deltaSE <- round(sqrt(diag(dAMEdBij %*% vcov(o) %*% t(dAMEdBij))),dig)

  # ------- bootstrap standard errors

  if(bootSE == T){

  R <- nboot
  bSE <- matrix(NA,R,length(cof1))
  for(i in 1:R){
    X1<-apply(X,2,sample)
    h<-dim(X1)[1]
    Y1 <- rbinom(h,1,0.5)
    oB <- glm(Y1~X1[,-1], family=binomial(link="logit"))

    logPDFB <- dlogis(predict(oB,type = "link"))
    cofB <- coef(oB)
    APEB <- mean(logPDFB)*cofB  # Wooldridge's average partial (marginal) effect(s)
    bSE[i,]<- APEB
  }

  bootSE <- round(apply(bSE,2,sd),dig)
  bootSE

  res1 <- matrix(cbind(t(APE1),t(deltaSE),t(bootSE)), ncol=3)
  colnames(res1)<-c("AME","deltaSE", "bootSE")
  rownames(res1)<- names(coefficients(o))
  return(res1)
   }

  res2 <- matrix(cbind(t(APE1),t(deltaSE)), ncol=2)
  colnames(res2)<-c("AME","deltaSE")
  rownames(res2)<- names(coefficients(o))
  return(res2)
}
```