Semi and non-parametric econometrics

Lee (2009): *Training, wages, and sample selection: Estimating sharp bounds on treatment effects*

Noémie Pinardon-Touati, Gauthier Schweitzer, Cyril Verluise

January 2018

1 Introduction

The sample selection problem arises when the outcome of interest is only observed for a non-randomly selected subpopulation. This may flaw causal analysis and is an ubiquitous phenomenon in many fields where treatment effect evaluations are conducted, such as labor, health, and educational economics.

In sample selection models (see the seminal work of Heckman, 1974, 1976, 1979), point identification of the parameters of interest commonly relies on tight functional form restrictions or on the availability of a valid instrument for selection. More recently, researchers have proposed bound estimators that require very few assumptions and do not rely on exclusion restrictions.[1] Lee (2009) contributes to this strand of the literature. He develops a trimming procedure for bounding average treatment effects in the presence of sample selection. It relies on the intuition developed by the principal stratification approach (Frangakis and Rubin, 2002), which consists in identifying causal effects within subgroups or stratum consisting of individuals with the same selection behavior. In contrast to existing methods, the procedure developed by Lee (2009) requires neither exclusion restrictions nor a bounded support for the outcome of interest.

He applies this method to the study of the effect of the Job Corps program on wages. Even with the aid of a randomized experiment, the impact of a training program on wages is difficult to study because of the sample selection problem: wage rates are only observed for those who are employed, and employment status itself may be affected by the training program. The key issue is that the subgroups defined by treated/control and employed/not employed are affected by post-randomization behaviors. Lee's estimator allows to obtain tight bounds on the coefficient of interest and shows a positive effect of the training on wages.

This document is organized as follows. Section 2 presents the method introduced by Lee (2009), highlights its key results and compares it with other bound estimators. Section 3 then reproduces the empirical investigation of Lee (2009) using simulated data.

2 Description of the procedure introduced by Lee (2009)

In this section, we present the procedure introduced by Lee (2009) for bounding average treatment effects in the presence of sample selection. We first present the methodology using an intuitive example, and then present the general procedure in detail. We insist on the role of each assumption for identification and on the intuition behind them. Finally, we compare this procedure with other existing methods.

[1] Partial identification of economic parameters in general goes back to Manski (1989).

2.1 Intuition

Consider the conventional setup, a la Heckman (1979), that models the wage determining process as:

$$\begin{aligned} Y^* &= \pi_1 + D\beta + U \\ Z^* &= \pi_2 + D\gamma + V \\ S &= 1_{[Z^* \geq 0]} \text{ and } Y = S \cdot Y^* \end{aligned} \tag{1}$$

where Y^* is the offered market wage, D is the indicator variable of receiving the treatment and Z^* is a latent variable representing the propensity to be employed. The wage is observed only for the employed. S is an indicator equal to 1 when the wage is observed. β is the (constant) treatment effect of interest. The treatment is assumed to be randomized, hence (U,V) and D are independent.

The observed means in the sample-selected treated and control group are equal to:

$$\mathbb{E}[Y|D=0, S=1] = \pi_1 + \mathbb{E}[U|D=0, V \geq -\pi_2], \tag{2}$$

$$\mathbb{E}[Y|D=1, S=1] = \pi_1 + \beta + \mathbb{E}[U|D=1, V \geq -\pi_2 - \gamma]. \tag{3}$$

This shows that when U and V are correlated, the difference in the means will generally be different from β. To understand the intuition behind Lee (2009), it is key to understand why this difference yields a biased estimate. Compared to individuals in the sample-selected treated group $(D=1, S=1)$, individuals in the sample-selected control group $(D=0, S=1)$ tend to have a higher V, that is to say a higher propensity to work conditional on D: we compare individuals with different selection behaviors. If V correlates (say, positively) with U, the sample-selected control group will have higher U, i.e. higher potential wages conditional on D. Therefore, the difference in means equals the coefficient of interest plus a selection bias due to the fact that we compare the wages of individuals who have systematically different potential wages given D.

Now, let us consider the following quantity

$$\mathbb{E}[Y|D=1, V \geq -\pi_2] \tag{4}$$

and note that the difference between (9) and (2) yields β. The preceding discussion should help understand why: now we compare individuals in the sample-selected treated and control group that share the same values of V. That is, we exclude from the sample-selected control groups the individuals that had the lowest V, i.e. the individuals that entered the labor market because of the treatment. In both groups, we thus end up with individuals who would have worked whether or not they received treatment. Therefore, the selection bias disappears and the difference in means across these two groups yields the coefficient of interest.

The general idea behind this approach is called principal stratification. It consists in identifying causal effects within subgroups or stratum consisting of individuals with the same selection behavior, i.e., being of the same "type" (here, the individuals with $V \geq -\pi_2$). This is useful because the selection problem does not arise within a particular stratum. We leave a more thorough discussion of the principal stratification approach to the next subsection.

Now that we have defined the causal effect of interest, we now have to find an estimation procedure for $\mathbb{E}[Y|D=1, V \geq -\pi_2]$, which is not observed. The bounding procedure proposed by Lee (2009) relies on the fact that:

$$\mathbb{E}[Y|D=1, S=1] = (1-p)\,\mathbb{E}[Y|D=1, V \geq -\pi_2] + p\,\mathbb{E}[Y|D=1, -\pi_2 - \gamma \leq V < -\pi_2] \tag{5}$$

where $p = \frac{\mathbb{P}[-\pi_2-\gamma \leq V < -\pi_2]}{\mathbb{P}[-\pi_2-\gamma \leq V]}$. We therefore have $\mathbb{E}[Y|D=1, V \geq -\pi_2] \leq \mathbb{E}[Y|D=1, S=1, Y \geq y_p]$, where y_p is the p^{th} quantile of the treatment group's observed Y distribution. This is true because among the selected population with $D=1, S=1$, no subpopulation with proportion $(1-p)$ can have a mean that is larger than the average of the largest $(1-p)$ values of Y. Besides, the trimming proportion p is equal to:

$$\frac{\mathbb{P}[S=1|D=1] - \mathbb{P}[S=1|D=0]}{\mathbb{P}[S=1|D=1]} = \frac{\mathbb{P}[V \geq -\pi_2 - \gamma] - \mathbb{P}[V \geq -\pi_2]}{\mathbb{P}[V \geq -\pi_2 - \gamma]} = \frac{\mathbb{P}[-\pi_2 - \gamma \leq V < -\pi_2]}{\mathbb{P}[-\pi_2 - \gamma \leq V]} = p$$

Intuitively, we do not know which individuals where induced to work by the treatment, but the worst-case scenario is that those individuals have the smallest p values of the distribution.

Note that the structure of the selection equation is key in this reasoning. The fact that γ is constant means that the effect of treatment on the probability of being observed goes in the same direction for all individuals. This ensures that the two populations are overlapping in terms of their characteristics V so that one subpopulation of the sample-selected treated group is comparable to the sample-selected control group. If γ is positive, the range of V in the sample-selected treated group is always larger than in the sample-selected control group. This would be the opposite if γ were negative. Besides, this ensures that p can be expressed in terms of observed probabilities.

2.2 General method

2.2.1 Definition of the estimand of interest

Let us consider a more general sample selection model:

$$\begin{aligned} S &= S_1 D + S_0(1-D) \\ Y^* &= Y_1^* D + Y_0^*(1-D) \\ Y &= S \cdot Y^* \end{aligned} \tag{6}$$

(Y_1^*, Y_0^*) are latent potential outcomes for the treated and control states, as in the classic Rubin causal model. (S_0, S_1) are "potential" sample selection indicators for the treated and control states. As before, S is an indicator variable that is equal to 1 if the individual is selected into the sample and D denotes treatment status. Assume that $(Y_1^*, Y_0^*, S_1, S_0, D)$ is i.i.d. across individuals. This model resembles the previous one but is much more general: it does not rely on a latent variable binary response model for the selection equation and the treatment effect is not assumed to be constant. As before, treatment assignment is assumed to be randomized so that we have:
A1. (Independence) : $(Y_1^*, Y_0^*, S_1, S_0) \perp D$.

As in the simple example above, in order to obtain a causal treatment effect, we want to compare outcomes within subpopulations whose individuals share the same potential values of the employment variable under both treatment arms. The principal stratification framework allows us to clarify this idea.

We define the basic principal stratification P_0 with respect to post-treatment variable S as the partition of individuals such that, within any set of P_0, all units have the same vector (S_1, S_0). In our case P_0 is given by:

$$\begin{aligned} 11 &= \{i : S_1 = S_0 = 1\} \\ 10 &= \{i : S_1 = 1, S_0 = 0\} \\ 01 &= \{i : S_1 = 0, S_0 = 1\} \\ 00 &= \{i : S_1 = S_0 = 0\} \end{aligned} \tag{7}$$

By definition, the stratum to which an individuals belongs is not affected by the treatment assignment. It can thus be viewed as a pretreatment covariate. Moreover, Assumption 1 guarantees to have the same distribution of potential outcomes in both treatment arms conditional on the stratum: $\{(Y_1^*, Y_0^*) \perp D\} \,|(S_1, S_0)$. Therefore, any effect defined conditional on a stratum is a well-defined causal effect.

The following correspondence between the observed values of D and S and the latent strata holds:

$$\begin{aligned} o(D=1,S=1) &= \{i : D=1, S_1=1\} \text{ i.e. i belongs to 11 or 10,} \\ o(D=1,S=0) &= \{i : D=1, S_1=0\} \text{ i.e. i belongs to 01 or 00,} \\ o(D=0,S=1) &= \{i : D=0, S_0=1\} \text{ i.e. i belongs to 11 or 01,} \\ o(D=0,S=0) &= \{i : D=0, S_0=0\} \text{ i.e. i belongs to 10 or 00.} \end{aligned} \tag{8}$$

In our setting, direct information on the causal effect can be found only in the 11 stratum of the always-respondents: since S represents non-response, only in this stratum can we observe both treated or control units. Therefore, the quantity of interest is:

$$\mathbb{E}[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1] = \mathbb{E}[Y_1^* | S_0 = 1, S_1 = 1] - \mathbb{E}[Y_0^* | S_0 = 1, S_1 = 1] \tag{9}$$

Note that this is the exact analog of our quantity of interest in the introductory subsection, as the subpopulation $V \geq -\pi_2$ corresponds to the individuals that would work irrespective of their treatment status, i.e. to the stratum 11.

Equation (10) shows that in each observed group we have a mixture of two strata. Therefore, it is not possible to point-identify the strata proportions, as well as the distribution of Y within the strata just by looking at the observed groups. To make progress toward estimating the quantity of interest, we further assume:
A2. (Monotonicity): $S_1 \geq S_0$ with probability 1.
Monotonicity rules out the existence of stratum 01. Therefore, we now have:

$$\begin{aligned} o(D=1,S=1) &= \{i : D=1, S_1=1\} \text{ subject i belongs to 11 or 10,} \\ o(D=1,S=0) &= \{i : D=1, S_1=0\} \text{ subject i belongs to 00,} \\ o(D=0,S=1) &= \{i : D=0, S_0=1\} \text{ subject i belongs to 11,} \\ o(D=0,S=0) &= \{i : D=0, S_0=0\} \text{ subject i belongs to 10 or 00.} \end{aligned} \tag{10}$$

As it will appear clearly below, monotonicity is critical for the results of Lee (2009). Monotonicity ensures that the sample-selected control group is composed only of individuals of strata 11. It therefore makes comparable a subset of the sample-selected treated group and the sample-selected control group. Besides, it allows to point-identify the strata proportions (proof below). Note that the assumption of a latent variable binary response model for the selection equation in the introductory model was implying monotonicity.

However, it is not possible to disentangle the distribution of Y between strata 11 and 10 in the sample-selected treated group: as in the simple example, we cannot identify which individuals were induced to work by the treatment. The next subsection deals with this issue.

2.2.2 Estimation strategy

With this definition of the treatment effect of interest in hand, we now turn to the estimation of the quantity of interest. Note that this is a separate problem, and several strategies can be envisioned

within the principal stratification framework, with a trade-off between the assumptions one is willing to make and the precision of the identification. Lee (2009) derives sharp bounds on the treatment effect under the two assumptions stated above. Alternative methods will be discussed in section 2.4.

Remember that the quantity of interest is:

$$\mathbb{E}[Y_1^*|S_0 = 1, S_1 = 1] - \mathbb{E}[Y_0^*|S_0 = 1, S_1 = 1] \tag{11}$$

As in the simple example above, the simple mean in the sample-selected control group yields the second term:

$$\begin{aligned}\mathbb{E}[Y|D = 0, S = 1] &= \mathbb{E}[Y_0^*|D = 0, S_0 = 1] \text{ using the definition of } Y \text{ and } S \text{ in (6)} \\ &= \mathbb{E}[Y_0^*|S_0 = 1] \text{ by A1} \\ &= \mathbb{E}[Y_0^*|S_0 = 1, S_1 = 1] \text{ by A2}\end{aligned}$$

Note that what ensures that the sample-selected control group is composed only of individuals of strata 11 is the monotonicity assumption.

As before, $\mathbb{E}[Y_1^*|S_0 = 1, S_1 = 1] = \mathbb{E}[Y|S_0 = 1, S_1 = 1, D = 1]$ is unobserved because the sample-selected treated group mixes strata 11 and 10. However, thanks to assumptions 1 and 2, this quantity can be bounded. The intuition is the same as that of the introductory example: the distribution of Y in the sample-selected treated group is a mix of the distribution of Y in strata 11 and 10 with known proportions, which implies constraints on the distribution of Y in strata 11.

We detail the steps of the proof more than what is done in the paper, because this provides useful intuitions on how the two assumptions allow to identify the bounds. Let us first show that the distribution of Y in the sample-selected treated group is a mix of the distribution of Y in strata 11 and 10 with known proportions. Let $F(y)$ be the c.d.f. of Y conditional on $D = 1, S = 1$.

$$\begin{aligned}F(y) = \mathbb{P}[Y \leq y|D = 1, S = 1] &= \mathbb{P}[Y_1^* \leq y|D = 1, S_1 = 1] \text{ using the definition of } Y \text{ and } S \text{ in (6)} \\ &= \mathbb{P}[Y_1^* \leq y|D = 1, S_1 = 1, S_0 = 1]\,\mathbb{P}[S_0 = 1|D = 1, S_1 = 1] \\ &+ \mathbb{P}[Y_1^* \leq y|D = 1, S_1 = 1, S_0 = 0]\,\mathbb{P}[S_0 = 0|D = 1, S_1 = 1] \\ &= (1 - p)N(y) + pM(y)\end{aligned}$$

where $M(y)$ denotes the c.d.f. of Y_1^*, conditional on $D = 1, S_0 = 0, S_1 = 1$, $N(y)$ denotes the c.d.f. of Y_1^*, conditional on $D = 1, S_0 = 1, S_1 = 1$ and

$$p = \mathbb{P}[S_0 = 0|D = 1, S_1 = 1] = \frac{\mathbb{P}[S_0 = 0, S_1 = 1|D = 1]}{\mathbb{P}[S_1 = 1|D = 1]} = \frac{\mathbb{P}[S_0 = 0, S_1 = 1]}{\mathbb{P}[S = 1|D = 1]} \text{ using A1 and the definition of } S,$$

that is to say that p is the proportion of stratum 10 in the sample-selected treated group.

Let us now show that $p = \frac{\mathbb{P}[S=1|D=1]-\mathbb{P}[S=1|D=0]}{\mathbb{P}[S=1|D=1]}$:

$$\frac{\mathbb{P}[S=1|D=1]-\mathbb{P}[S=1|D=0]}{\mathbb{P}[S=1|D=1]} = \frac{\mathbb{P}[S_1=1|D=1]-\mathbb{P}[S_0=1|D=0]}{\mathbb{P}[S=1|D=1]} \text{ using the definition of } S \text{ in (6)}$$

$$= \frac{\mathbb{P}[S_1=1]-\mathbb{P}[S_0=1]}{\mathbb{P}[S=1|D=1]} \text{ using A1}$$

$$= \frac{\mathbb{E}[S_1-S_0]}{\mathbb{P}[S=1|D=1]}$$

$$= \frac{0.\big(\mathbb{P}[S_0=1,S_1=1]+\mathbb{P}[S_0=0,S_1=0]\big)+1.\mathbb{P}[S_0=0,S_1=1]}{\mathbb{P}[S=1|D=1]} \text{ since by A2, } (S_1-S_0)\in\{0,1\}$$

$$= \frac{\mathbb{P}[S_0=0,S_1=1]}{\mathbb{P}[S=1|D=1]} = p$$

Thus $F(y) = pM(y) + (1-p)N(y)$ where p can be computed using know probabilities. Again, monotonicity is critical to ensure that we can pin down the proportion of "marginal individuals" (i.e. of stratum 10) in the sample-selected treated group. Intuitively, by assuming that strata 01 does not exist, we are left with two unknown proportions (of strata 11 and 10 in the population) and two pieces of sampling information (the proportions of respondents among treated and control units).

We then conclude using Lemma 1. Intuitively, $F(y) = pM(y) + (1-p)N(y)$ means that Y is a mixture of two random variables with a known mixing proportion p. Therefore, the expectation of the random variable with c.d.f. N cannot be greater than the expectation of the $(1-p)$ greatest values of the distribution of Y. Since N denotes the c.d.f. of $Y_1^*|D=1,S_0=1,S_1=1$, which by A1 is equivalent to the c.d.f. of $Y_1^*|S_0=1,S_1=1$, we have that:

$$\mathbb{E}[Y_1^*|S_0=1,S_1=1] \leq \mathbb{E}[Y|D=1,S=1,Y\geq y_p]$$

The lower bound on $\mathbb{E}[Y_1^*|S_0=1,S_1=1]$ can be found using the same methodology. This is how we get the results in Proposition 1a. of the paper.

Importantly, note that this procedure allows to identify the causal effect only on the always-observed group. In the introductory example, this did not matter since the treatment effect was assumed constant. With a potentially heterogeneous treatment effect, the quantity of interest corresponds to a local average treatment effect for stratum 11.

2.2.3 Including covariates

One can use these covariates to reduce the width of the bounds for the same estimand that has been discussed so far. A1 must now write $(Y_1^*, Y_0^*, S_1, S_0, X) \perp D$. The intuition is the following: conditional on X, we can repeat the exact same procedure as above and get bounds for the average treatment effect for the 11 stratum for each values of X. Taking the expectation of these bounds w.r.t. the distribution of X, we get bounds for the estimand that has been discussed so far. These bounds are necessarily narrower since any treatment effect that is consistent with an observed distribution of (Y,S,D,X), must also be consistent with the data after throwing away information on X, and observing only (Y,S,D).

2.3 Estimation and inference

The estimates of the bounds are sample analogs to the parameters defined in Proposition 1a. They will be detailed in the application section of this paper. The consistency and asymptotic normal-

ity of the estimator follow from general results on GMM estimators. The asymptotic variance depends on the variance of the trimmed outcome variable but also on the trimming threshold, which is an estimated quantile. There is also an added term that accounts for the estimation of the proportion of the distribution that has to be trimmed.

2.4 Comparison with other procedures

This subsection compares the procedure introduced by Lee (2009) with other potential methods, with a focus on the strength of the assumptions. Indeed, causal inference requires some assumptions, but the credibility of (causal) inference decreases with the strength of the maintained assumptions (Manski, 2003). We review only methods that can credibly solve the sample selection problem. For instance, identification using the conditional independence assumption would require to find covariates X such that $(Y_1^*, Y_0^*) \perp (S_1, S_0, D)|X$ so that $\mathbb{E}[Y|S = 1, D = 0, X] - \mathbb{E}[Y|S = 1, D = 1, X] = \mathbb{E}[Y_1^* - Y_0^*|X]$, which runs against theories of labour supply that account for the participation decision.

2.4.1 Identification with the data alone

Without making any assumptions, one can always construct worst-case scenario bounds of the treatment effect (Horowitz et Manski, 2000). When the support of the outcome is bounded, the idea is to impute the missing data with either the largest or the smallest possible values to compute the largest and smallest possible treatment effects consistent with the data. However, this imputation procedure cannot be used when the support is unbounded. Even when the support is bounded, if it is very wide, so too will be the width of the treatment effect bounds.

2.4.2 IV and Heckman selection

Post-randomization actions are usually described as problems of endogenous selection and represented by means of selection models (Heckman, 1974). The standard specification of a selection model is (1), adding covariates:

$$\begin{aligned} Y^* &= D\beta + X\pi_1 + U \\ Z^* &= D\gamma + X\pi_2 + V \\ S &= 1_{[Z^* \geq 0]} \text{ and } Y = S \cdot Y^* \end{aligned} \tag{12}$$

where (U,V) and D are independent conditional on X. As in the introductory model, monotonicity holds by construction. The goal of inference is estimating β in the first equation of (12), that should be valid for the whole population in the absence of complications (i.e., as if the data came from random sampling). Because observations come from a nonrandom sampling procedure, it is necessary to include a selection equation, in order to "correct" the estimation of the causal effect. This differs from the principal stratification approach which focuses on information contained in specific subgroups of units, aiming at producing valid inference conditional on such subgroups, without an a priori extension of the results to the whole population.

The textbook Heckman setting is fully parametric : identification is achieved by assuming the joint normality of (U, V). However, this parametric approach has been criticized for relying on too restrictive assumptions and for being vulnerable to misspecification (Puhani 2000 ; Grasdal 2001). Various extensions of the model have been proposed, which include semi and nonparametric versions. However, in order to point identify treatment effects semiparametrically or nonparametrically, one has to rely on exclusion restrictions. Notably, β can be identified if some of the

exogenous variables determine sample selection but do not have their own direct impact on the outcome of interest; i.e., some of the elements of π_1 are zero, while corresponding elements of π_2 are nonzero. Therefore, in any case, identification comes at the cost of much stronger restrictions than Lee (2009).

2.4.3 Principal stratification and parametric assumptions to obtain point indentification

Within the principal stratification framework, several estimation methods can be envisioned, relying on different assumptions. Common assumptions are : (a) exclusion restriction (ER), that zero effect on the intermediate variable implies zero effect on the outcome (e.g. Angrist et al. (1996)); (b) normal outcome distributions within principal strata (e.g. Zhang et al. (2009) and Frumento et al. (2012))[2]; (c) additional covariates or secondary outcomes (Mattei and Mealli, 2011; Mattei et al., 2013; Mealli and Pacini, 2013; Yang and Small, 2016; Jiang et al., 2016). Only parametric distributional assumptions can lead to point identification. This is because the normality assumption allows the use of results on finite mixture distribution theory to disentangle the distributions of Y in the 11 and 10 strata in the sample-selected treated group.

The procedure outlined by Lee (2009) has the advantage of relying on mild assumptions. Whether adding other assumptions would yield more precise bounds is then an empirical question.

3 Application de Lee (2009) à des données simulées

Dans cette section, on se propose de répliquer la méthode développée par Lee (2009). Cette méthode consiste à élaguer les données afin de borner l'estimateur de l'effet moyen d'un traitement lorsque la sélection n'est pas aléatoire.

3.1 Le problème

Plus précisément, à la suite de Lee (2009), nous considérons l'effet du Job Corps Program. On cherche à connaître l'effet (ici, *Intent To Treat effect*) de ce programme sur le salaire. Bien que le droit de participer au programme ait été déterminé de manière aléatoire, la difficulté provient du fait que seuls les salaires des individus effectivement employés, voire strictement supérieurs à un salaire minimum, sont observés. Or, il est fort probable que le traitement ait également un effet sur la probabilité d'être employé, cela conduit à un effet de sélection qu'il est nécessaire de prendre en compte pour évaluer rigoureusement l'effet du programme sur les salaires.

3.2 Explicitation du processus de sélection et simulation de la distribution des salaires observés

3.2.1 Explicitation du processus de sélection

À la suite d' Heckman (1979), on considère que le salaire observé provient de l'expression de deux variables latentes: le salaire potentiel et la propension à être employé. Formellement, le processus

[2]Note that these parametric assumptions are weaker than those required in the Heckman sample selection model, since only the distributions of Y within strata must be parameterized.

de détermination du salaire est décrit par le système (12), reproduit ici:

$$\begin{aligned} Y^* &= D\beta + X\pi_1 + U \\ Z^* &= D\gamma + X\pi_2 + V \\ S &= 1_{[Z^* \geq 0]} \text{ et } Y = S \cdot Y^* \end{aligned} \tag{13}$$

où Y^* et Z^* sont respectivement le logarithme du salaire potentiel et la propension à être employé. Le logarithme du salaire potentiel (Y^*) est représenté comme une combinaison de l'effet du programme sur le salaire (β), des caractéristiques de l'individu susceptibles d'affecter le salaire proposé (X) et d'un terme d'erreur (U). La propension à être employé est quant à elle modélisée comme la combinaison de l'effet causal du traitement (γ), des caractéristiques de l'individu qui affectent son employabilité (X) et d'un terme d'erreur (V). Suivant le cadre présenté par Lee (2009), nous considérons que les caractéristiques de l'individu affectant l'employabilité et le salaire sont les mêmes. Y^* représente le logarithme du salaire et non le salaire lui-même, afin de s'assurer que les salaires observés sont bien tous positifs.

Comme cela a été noté précédemment, il est important de noter que les termes d'erreur U et V sont potentiellement corrélés, sans quoi nous ferions l'hypothèse que la propension à être employé et le salaire potentiel sont indépendants. Une telle hypothèse serait déraisonnable, notamment au regard de concepts centraux de la théorie de l'offre de travail. Nous accordons donc une attention particulière à cet aspect au cours de notre simulation. Précisément, nous simulons U et V comme un couple de normales bivariées de covariance non nulle.

```
In [122]: # Choix de la taille de l'échantillon
          N = 10000

          # Simulation des erreurs
          sigma_mat = matrix(c(0.25, 0.5, 0.5, 1), nrow=2, ncol=2, byrow = TRUE)
          # matrice de var-cov des erreurs
          set.seed(123)
          UV = rmvnorm(n = N, mean = c(0,0), sigma = sigma_mat)
```

Dans le cadre de notre exercice de simulation, on assigne aléatoirement les individus aux groupes test et contrôle et on fixe les valeurs de β, γ, π_1 et π_2. Cela nous permet de calculer Y^*, Z^*. On en déduit Y, le logarithme du salaire observé. Il correspond à l'intersection des logarithmes des salaires potentiels (Y^*) et des emplois effectifs ($Z^* \geq 0$).

NB: Par souci de simplicité, X est de dimension N x 2, où N est la taille de l'échantillon. X est donc simplement la combinaison d'une constante et d'une unique variable explicative (comparable à un *propensity score*, c'est à dire un résumé univarié de l'ensemble des variables pré-traitement, tel que défini par Rosenbaum et Rubin (1983)). Cette variable indépendante est simulée suivant une loi normale centrée-réduite, suivant le cadre présenté par Mealli et Pacini (2008).

```
In [123]: # Assignation des individus a un groupe de traitement (proportion p)
          # et de controle (1-p)
          p = 0.5 # proportion de l'echantillon traitee
          D = rbinom(n = N, size = 1, prob = p) ## assignation des individus

          # Détermination des "vrais" paramètres
          beta = 0.2 # effet du traitement sur les salaires
```

```
gamma = 0.3 # effet du traitement sur l'employabilité
pi = matrix (c(9.8,0.35,0.3,1), nrow = 2, ncol = 2) # constantes des 2 équations

# X
X = matrix(data = 1, nrow = N, ncol = 2) # C1 = constante 1
X[,2] = rnorm(n = N, mean = 0, sd = 1)  # C2 = propensity score

# Calcul des variables latentes et du log du salaire observé
Y_star = D*beta + X%*%pi[,1] + UV[,1]
Z_star = D*gamma + X%*%pi[,2] + UV[,2]
Y = 0 + (Z_star>=0)*Y_star

# Variable indicatrice du salaire observé (1 si salaire observé, 0 sinon)
S = vector(length = N)
S = 0 + (Y>0)*1
```

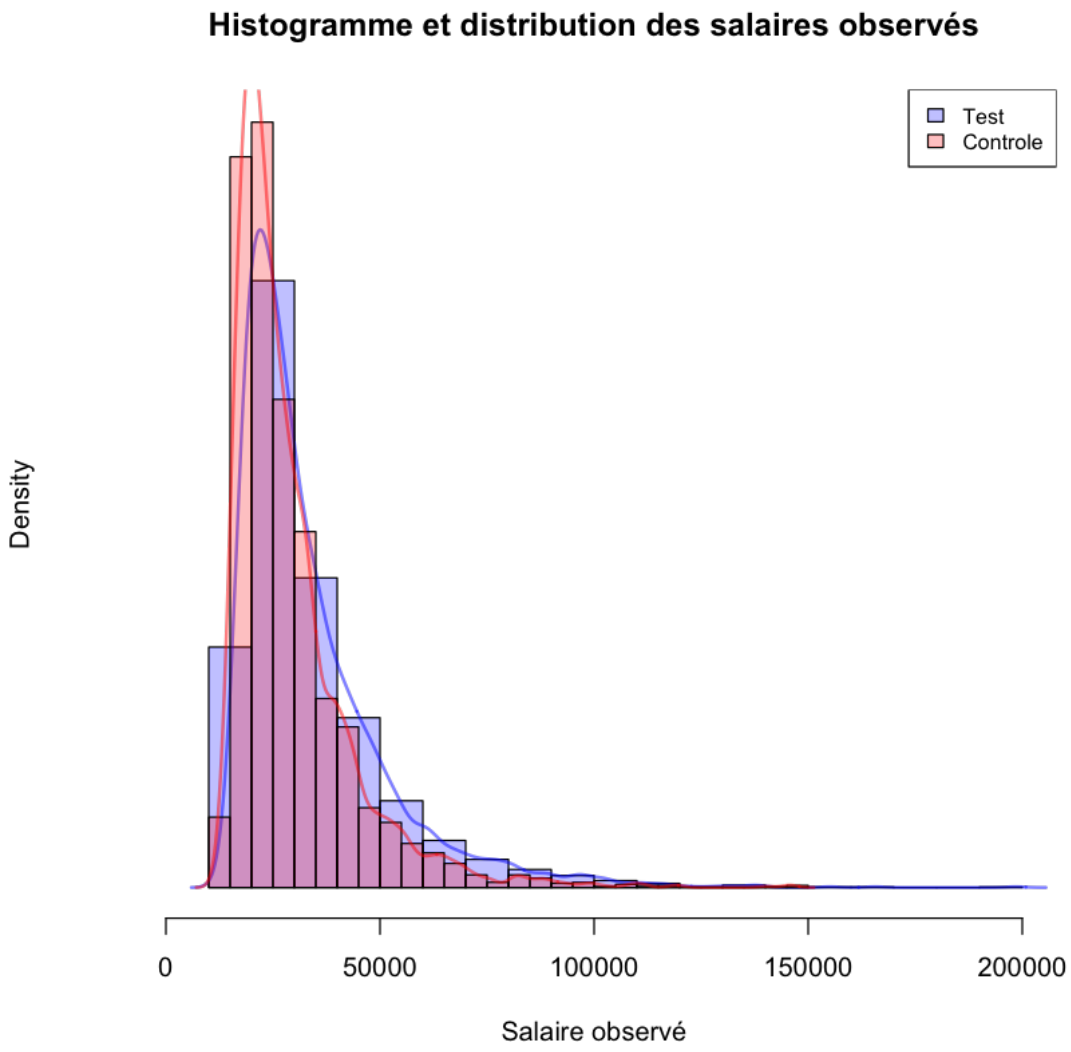
3.2.2 Distribution des salaires observés

Pour bien comprendre la situation, on trace l'histogramme des salaires observés et on calcule le taux de chômage pour les groupes test et contrôle. On observe que les salaires du groupe test sont en moyenne plus élevés et que le taux de chômage y est plus faible ce qui est cohérent avec nos attentes.

In [124]:
```
# Représentation de la distribution des salaires observés
hist(exp(Y[(D==1)&(S==1)]),breaks=20, freq = F, col=rgb(0,0,1,1/4),
     xlim = c(0,200000),xlab="Salaire observé",
     ylim=c(0,5e-5),yaxt='n',
     main="Histogramme et distribution des salaires observés")
lines(density(exp(Y[(D==1)&(S==1)])),lwd = 2,col = rgb(0,0,1,1/2))
hist(exp(Y[(D==0)&(S==1)]), freq = F, breaks=20, col=rgb(1,0,0,1/4),
     add=T)
lines(density(exp(Y[(D==0)&(S==1)])),lwd = 2,col = rgb(1,0,0,1/2))
legend('topright', legend=c("Test", "Controle"),
       fill=c(rgb(0,0,1,1/4), rgb(1,0,0,1/4)), cex=0.8)

# Calcul du taux de chômage
# Dans le groupe test et controle (resp)
u_test=sum((Y == 0)&(D == 1))/length(D ==1)
u_cont=sum((Y == 0)&(D == 0))/length(D ==0)
print(paste("Le taux de chômage dans le groupe test est de",
            u_test*100,"%"))
print(paste("Le taux de chômage dans le groupe contrôle est de",
            u_cont*100,"%"))
```

```
[1] "Le taux de chômage dans le groupe test est de 16.72 %"
[1] "Le taux de chômage dans le groupe contrôle est de 20.8 %"
```



A ce stade, l'approche de Heckman (1979) atteint ses limites. En effet, si elle permet une définition claire de l'effet de sélection théorique, elle suppose de faire une hypothèse d'exclusion pour identifier cet effet empiriquement. La pratique la plus courante est de considérer que la sélection provient de variables exogènes observables. Nous ne disposons pas de telles variables.

3.3 Idenfication de l'effet de sélection et élagage des données

3.3.1 Introduction à l'identification de l'effet de sélection par une approche du type *worst case scenario* a la Horowitz et Manski (2000)

Pour dépasser cette limite nous nous inspirons d'une seconde approche développée dans la littérature. Cette approche consiste à construire des *scenari* dits *worst case* afin d'encadrer l'effet du traitement. Dans ce cadre, l'idée développée par Horowitz et Manski (2000) est d'imputer les valeurs manquantes aux bornes du domaine des observations. De fait, en imputant successive-

ment toutes les données manquantes à la borne inférieure puis à la borne supérieure on obtient bien une borne inférieure puis une borne supérieure de l'effet du traitement.

Les auteurs définissent ainsi la borne supérieure comme suit:

$$P(Z^* \geq 0|D=1)E(Y|D=1) + P(Z^* < 0|D=1)Y^{UB} - (P(Z^* \geq 0|D=0)E(Y|D=0) + P(Z^* < 0|D=0)Y^{LB})$$

NB: La borne inférieure est obtenue par symétrie.

```
In [125]: ub_HM=( (1-u_test)*mean(Y[(S==1)&(D==1)]) + u_test*max(Y[S==1]) ) -
              ( (1-u_cont)*mean(Y[(S==1) & (D==0)]) + u_cont*min(Y[S==1]) )
          lb_HM=( (1-u_test)*mean(Y[(S==1)&(D==1)]) + u_test*min(Y[S==1]) ) -
          ( (1-u_cont)*mean(Y[(S==1) & (D==0)]) + u_cont*max(Y[S==1]) )

          print(paste("L'effet mesuré se situe dans l'intervalle [",
                      round(lb_HM,digits = 2),";",round(ub_HM, digits = 2),"]"))
```

```
[1] "L'effet mesuré se situe dans l'intervalle [ -0.43 ; 0.61 ]"
```

L'intervalle ainsi obtenu contient bien la vraie valeur de β. Toutefois, l'intervalle donné par les bornes ainsi définies est très large et donc très peu précis. Il contient notamment des valeurs positives et négatives.

Cette approche est de fait limitée par l'usage qu'elle fait des bornes du domaine d'observation de la variable dépendante. S'il est vrai que la distribution des salaires est bornée, ces bornes sont en fait un signal très pauvre. D'une part, la borne basse révèle pour partie des minimas décidés arbitrairement ou du moins sur la base de critères non économiques (ex: salaire minimum). D'autre part, la borne haute est très sensible à la présence d'*outliers* (ex: un milliardaire). L'information portée par les bornes est donc très pauvre et donne un encadrement très lâche de l'effet du programme.

3.3.2 Elagage des données a la Lee (2008)

La stratégie développée par Lee (2009) permet une meilleure prise en compte de l'information portée par la distribution observée des salaires. Cette approche a déjà fait l'objet d'une présentation théorique. Nous nous contentons ici d'en présenter les principaux aspects empiriques.

La construction du *worst case scenario* consiste ici à élaguer la queue (inférieure ou supérieure selon la borne voulue) de la distribution des salaires observés au sein du groupe test. L'objectif est de rendre les deux groupes "comparables". Intuitivement, cette approche comporte l'avantage de réduire la sensibilité aux valeurs extrêmes ou arbitraires et conduit à se concentrer sur le coeur de la distribution, là où se situe l'information qui nous intéresse.

La première étape consiste donc à calculer la proportion à élaguer. Elle est définie comme suit:

$$\frac{P(Z^* \geq 0|D=1) - P(Z^* \geq 0|D=0)}{P(Z^* \geq 0|D=1)}$$

```
In [126]: # On estime d'abord la proportion de trimming
          p = (sum((S==1)&(D == 1))/length(D ==1) -
          sum((S==1)&(D == 0))/length(D ==0))/(sum((S==1)&(D == 1))/length(D == 1))
```

La seconde étape revient à élaguer la distribution des salaires observés au sein du groupe test: par le haut pour obtenir une borne inférieure et par le bas pour obtenir une borne supérieure. Cela nous permet alors de calculer les bornes par une différence de moyennes.

Plus précisément, on a :

$$\Delta^{\hat{LB}} = \frac{\sum Y.S.D.1_{[Y \leq \hat{y_1 - p}]}}{\sum S.D.1_{[Y \leq \hat{y_1 - p}]}} - \frac{\sum Y.S.(1-D)}{\sum S.(1-D)}$$

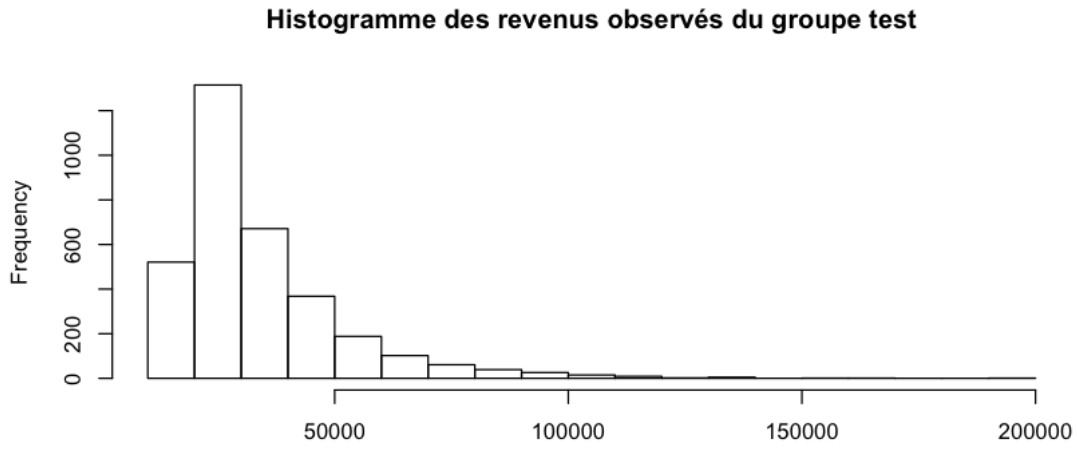
$$\Delta^{\hat{UB}} = \frac{\sum Y.S.D.1_{[Y \leq \hat{yp}]}}{\sum S.D.1_{[Y \leq \hat{yp}]}} - \frac{\sum Y.S.(1-D)}{\sum S.(1-D)}$$

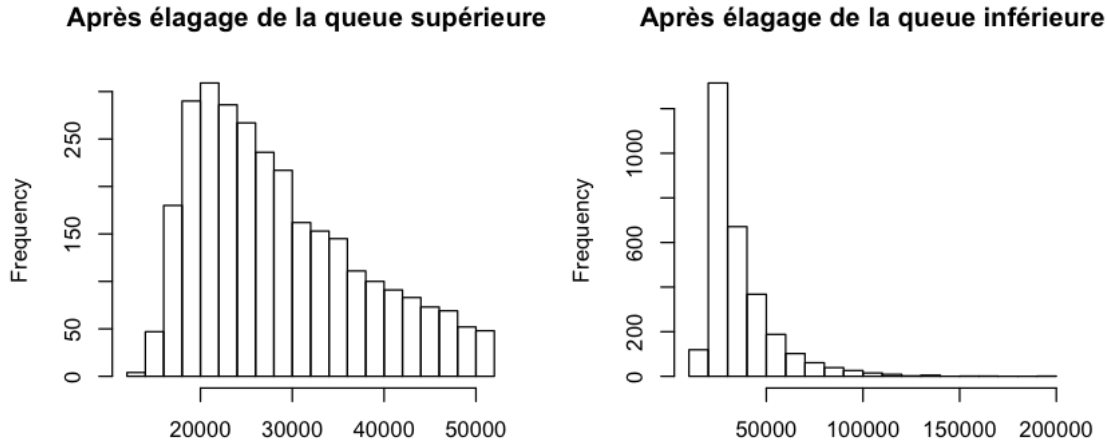
In [127]:
```
# On estime les quantiles associes
y_p = as.numeric(quantile(x = Y[(S==1)&(D == 1)], probs = c(p)))
y_1p = as.numeric(quantile(x = Y[(S==1)&(D == 1)], probs = c(1-p)))

# Représentation de l'élagage
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
hist(exp(Y[(S==1)&(D==1)]),breaks=20,xlab="",
     main="Histogramme des revenus observés du groupe test")
hist(exp(Y[(S==1)&(D==1)&(Y<=y_1p)]),breaks=20, xlab="",
     main="Après élagage de la queue supérieure ")
hist(exp(Y[(S==1)&(D==1)&(Y>=y_p)]),breaks=20, xlab="",
     main="Après élagage de la queue inférieure")

# Calcul des bounds
lb = mean(Y[(S==1)&(D==1)&(Y<=y_1p)]) - mean(Y[(S==1)&(D==0)])
ub = mean(Y[(S==1)&(D==1)&(Y>=y_p)]) - mean(Y[(S==1)&(D==0)])
print(paste("L'effet mesuré se situe dans l'intervalle [",
            round(lb,digits = 2),";",round(ub, digits = 2),"]"))
#(ub - lb) /var(Y[Y>0])
```

```
[1] "L'effet mesuré se situe dans l'intervalle [ 0.03 ; 0.22 ]"
```





On constate que la vraie valeur de β est bien dans l'intervalle obtenu. Plus intéressant, l'intervalle est beaucoup plus réduit que dans le cas de l'approche a la Horowitz et Manski (2000). Ces résultats sont très encourageants, il s'agit désormais d'en mesurer la validité et d'explorer la possibilité de les améliorer.

3.4 Validité et amélioration des résultats

3.4.1 Calcul de la variance

Afin d'étudier la validité des résultats donnés par l'approche de Lee (2009), on s'intéresse à la variance des bornes qu'on vient d'obtenir et on construit un intervalle de confiance à 95% pour l'effet du traitement.

Pour cela, on a recours aux formules de Lee (2009):

$$V^{LB} = \frac{V[Y|D=1,S=1,Y\leq y_{1-p_0}]}{E[SD](1-p_0)} + \frac{(y_{1-p_0}-\mu^{LB})^2 p_0}{E[SD](1-p_0)} + \Big(\frac{y_{1-p_0}-\mu^{LB}}{1-p_0}\Big)^2 V^P$$

$$V^{UB} = \frac{V[Y|D=1,S=1,Y\leq y_{p_0}]}{E[SD](1-p_0)} + \frac{(y_{p_0}-\mu^{UB})^2 p_0}{E[SD](p_0)} + \Big(\frac{y_{p_0}-\mu^{UB}}{p_0}\Big)^2 V^P$$

$$V^p = (1-p_0)^2\Big(\frac{1-\frac{\alpha_0}{1-p_0}}{E[D]\frac{\alpha_0}{1-p_0}} + \frac{1-\alpha_0}{(1-E[D])\alpha_0}\Big)$$

$$IC_{95\%} = \big[-1,96\frac{\hat{\sigma_{LB}}}{\sqrt{n}}; +1,96\frac{\hat{\sigma_{UB}}}{\sqrt{n}}\big]$$

In [128]:
```
# Calcul des mu et des variances incluses dans les formules de variance des bornes
mu_lb = mean(Y[(S==1)&(D==1)&(Y<=y_1p)])
mu_ub = mean(Y[(S==1)&(D==1)&(Y>=y_p)])
v1_temp = sum((Y[(S==1)&(D==1)&(Y<=y_1p)]-mu_lb)^2)/
(length(Y[(S==1)&(D==1)&(Y<=y_1p)])-1)
v2_temp = sum((Y[(S==1)&(D==1)&(Y>=y_p)]-mu_lb)^2)/
(length(Y[(S==1)&(D==1)&(Y>=y_p)])-1)

# Calcul des variances des bornes
v_lb = 1/(mean(S*D)*(1-p))*(v1_temp+(y_1p-mu_lb)^2*p)+(y_1p-mu_lb)^2*
((1-mean(S[D==0])-p*(1-mean(D)))/(mean(D)*mean(S[D==0])*(1-mean(D))))
v_ub = 1/(mean(S*D)*(1-p))*(v2_temp+(y_p-mu_ub)^2*p)+(y_p-mu_ub)^2*
((1-mean(S[D==0])-p*(1-mean(D)))/(mean(D)*mean(S[D==0])*(1-mean(D))))

# Calcul de l'IC
ic_lb = lb -1.96*(sqrt(v_lb/N))
ic_ub = ub +1.96*(sqrt(v_ub/N))
rm(y_p, y_1p, v1_temp, v2_temp, mu_lb, mu_ub)

print(paste("L'effet mesuré se situe avec une probabilité de 95% dans l'intervalle [
            round(ic_lb,digits = 2),";",round(ic_ub, digits = 2),"]"))
```

```
[1] "L'effet mesuré se situe avec une probabilité de 95% dans l'intervalle [ 0 ; 0.24 ]"
```

À ce stade plusieurs éléments méritent d'être notés :

- la vraie valeur de β se situe bien dans l'intervalle de confiance à 95%,
- cet intervalle est beaucoup moins ample que l'intervalle donné par une approche à la Horowitz et Manski (2011),
- nous ne parvenons pas à exclure un effet nul du traitement.

3.4.2 Amélioration des résultats

Si les deux premiers éléments sont très encourageants, le troisième reste préoccupant. Par conséquent, on s'intéresse à la possibilité de gagner en précision d'estimation. pour cela, on cherche à utiliser l'information portée par des variables explicatives corrélées à Y.

On développe donc une approche similaire à celle de Lee (2009). Pour cela, les étapes sont les suivantes. On commence par régresser le salaire observé sur les variables explicatives disponibles (en l'occurence X), pour l'échantillon de ceux dont on observe le salaire. A partir des coefficients obtenus, on prédit un salaire pour l'ensemble des individus. Sur la base de cette prédiction, on forme 5 groupes mutuellement exclusifs de salaires prédits, correspondant aux différents quintiles. A noter que chacun de ces groupes comprend aussi bien des individus de l'ensemble de "test" que des individus de l'ensemble de "contrôle".

```
In [129]: # Définition de 5 groupes
          # On commence par regresser le salaire sur X sans constante
          # (deja dans X) pour ceux dont on observe le salaire
          data = data.frame(Y[S==1],X[S==1,])
          colnames(data) = c("Y","X1", "X2")
          fit <- lm(Y~X1 + X2 -1, data = data)
          #summary(fit)

          # Calcul du salaire prédit
          Y_predicted = X %*% fit$coefficients

          # Calcul des quintiles empiriques
          quantiles = as.numeric(quantile(x = Y_predicted,
                                          probs = c(1/5,2/5,3/5,4/5)))
          group = vector(mode = "numeric", length = N)
          group = (Y_predicted < quantiles[1])*1
          for (i in 1:3) {
            group = group + (i+1)*
              ((Y_predicted>=quantiles[i])&(Y_predicted<quantiles[i+1]))
          }
          group = group + (Y_predicted > quantiles[4])*5
```

Au sein de chacun de ces groupes, on peut donc répliquer une méthode similaire à celle qu'on vient de développer sur l'ensemble des observations (élagage, calcul des bornes, calcul des variances et calcul d'un intervalle de confiance).

```
In [130]: # Définition de 5 vecteurs pour les variables d'interet
          p_cov = vector(mode = "numeric", length = 5)
          y_p_cov = vector(mode = "numeric", length = 5)
          y_1p_cov = vector(mode = "numeric", length = 5)
          lb_cov = vector(mode = "numeric", length = 5)
          ub_cov = vector(mode = "numeric", length = 5)

          for (i in 1:5){
            # Estimation de la proportion de trimming
            p_cov[i] = (sum((S==1)&(D==1)&(group==i))/length((D ==1)&(group==i)) -
```

```
  sum((S==1)&(D==0)&(group==i))/length((D ==0)&(group==i)))/
  (sum((S==1)&(D==1)&(group==i))/length((D==1)&(group==i)))

  # Estimation des quantiles associes
  y_p_cov[i] = as.numeric(quantile(x = Y[(S==1)&(D==1)&(group==i)],
  probs =c(p_cov[i])))
  y_1p_cov[i] = as.numeric(quantile(x = Y[(S==1)&(D==1)&(group==i)],
  probs =c(1-p_cov[i])))

  # Calcul des bounds
  lb_cov[i] = mean(Y[(S==1)&(D==1)&(group==i)&(Y<=y_1p_cov[i])]) -
  mean(Y[(S==1)&(D==0)&(group==i)])
  ub_cov[i] = mean(Y[(S==1)&(D==1)&(group==i)&(Y>=y_p_cov[i])]) -
  mean(Y[(S==1)&(D==0)&(group==i)])
}


# Calcul des variances dans chaque groupe
# Définition de 2 vecteurs pour les variables d'interet
v_lb_cov = vector(mode = "numeric", length = 5)
v_ub_cov = vector(mode = "numeric", length = 5)

for (i in 1:5){
  y_1p = y_1p_cov[i]
  y_p = y_p_cov[i]
  p = p_cov[i]
  # Calcul des mu et variances incluses dans les formules de variance des bornes
  mu_lb = mean(Y[(S==1)&(D==1)&(Y<=y_1p)&(group==i)])
  mu_ub = mean(Y[(S==1)&(D==1)&(Y>=y_p)&(group==i)])
  v1_temp = sum((Y[(S==1)&(D==1)&(Y<=y_1p)&(group==i)]-mu_lb)^2)/
  (length(Y[(S==1)&(D==1)&(Y<=y_1p)&(group==i)])-1)
  v2_temp = sum((Y[(S==1)&(D==1)&(Y>=y_p)&(group==i)]-mu_lb)^2)/
  (length(Y[(S==1)&(D==1)&(Y>=y_p)&(group==i)])-1)

  # Calcule des variances des bornes
  v_lb_cov[i] = 1/(mean(S[group==i]*D[group==i])*(1-p))*(v1_temp+(y_1p-mu_lb)^2*p)+
  (y_1p-mu_lb)^2*((1-mean(S[(D==0)&(group==1)])-p*(1-mean(D[group==i])))/
  (mean(D[group==i]*mean(S[(D==0)&(group==i)])*(1-mean(D[group==i]))))
  v_ub_cov[i] =1/(mean(S[group==i]*D[group==i])*(1-p))*(v2_temp+(y_p-mu_ub)^2*p)+
  (y_p-mu_ub)^2*((1-mean(S[(D==0)&(group==1)])-p*(1-mean(D[group==i])))/
  (mean(D[group==i])*mean(S[(D==0)&(group==i)])*(1-mean(D[group==i]))))
}
```

À partir de là, on peut calculer les bornes, les variances et l'intervalle de confiance de l'effet moyen agrégé du traitement en prenant en compte la pondération de chacun des groupes, suivant la méthode décrite dans Lee (2009)

In [131]:
```
# Calcul des coeffcients de pondération
weight = (1-p_cov)/sum((1-p_cov))
```

```r
# Calcul de la moyenne pondérée des ub et lw
ub_final = as.numeric(ub_cov %*% weight)
lb_final = as.numeric(lb_cov %*% weight)

# Variance totales de bornes
v_lb_final = as.numeric(v_lb_cov %*% weight +((lb_cov - lb_final)^2)%*% weight)
v_ub_final = as.numeric(v_ub_cov %*% weight +((ub_cov - ub_final)^2)%*% weight)

# IC
ic_lb_final = lb_final -1.96*(sqrt(v_lb_final/N))
ic_ub_final = ub_final +1.96*(sqrt(v_ub_final/N))

print(paste("L'effet mesuré se situe avec une probabilité de 95% dans l'intervalle [",
            round(ic_lb_final,digits = 2),";",round(ic_ub_final, digits = 2),"]"))
rm  (quantiles, i, Y_predicted,weight,v1_temp, v2_temp, mu_lb, mu_ub)
```

```
[1] "L'effet mesuré se situe avec une probabilité de 95% dans l'intervalle [ 0.01 ; 0.24 ]"
```

On constate que la méthode de Lee (2009) avec prise en compte des variables explicatives corrélées au salaire observé permet de gagner en précision et ainsi de s'assurer que la probabilité que l'effet du traitement soit non nul dépasse 95%.

3.5 Résultats

Pour rappel, la vraie valeur de l'effet du traitement sur le salaire est de 0,2. Une simple différence de moyenne conduit à un effet estimé de 0,14 mais ne saurait être considérée comme une estimation rigoureuse. Elle ne prend pas en compte l'effet de sélection lié au fait que le salaire n'est observé que pour les agents ayant un emploi et que l'employabilité est elle-même affectée par le traitement. Après avoir exploré plusieurs méthodes prenant en compte cet effet, il apparaît que la méthode de Lee (2009) avec usage de variables corrélées permet à la fois de prendre en compte cet effet de sélection et donne des résultats relativement précis (voir ci-dessous, Tableau récapitulatif des principaux résultats). Dans notre cas, cette méthode permet d'obtenir un intervalle de confiance à 95% près de 5 fois moins ample que l'approche de Manski et Horowitz (2000) et permet d'affirmer avec une probabilité supérieure à 95% que l'effet du traitement est non nul.

Pour conclure, on remarque qu'en dépit de ses bonnes peformances, la méthode de Lee (2009) semble être affectée par la forme de la distribution étudiée, en particulier de son asymétrie (*skewness*). On remarque ainsi que les bornes issues de la méthode de Manski et Horowitz (2000) sont symétriques autour de la vraie valeur. Ce n'est pas le cas des bornes obtenues par la méthode de Lee (2009). Alors que la borne haute est très proche de la vraie valeur, la borne basse en est très éloignée, ce qui pourrait refléter la *positive skewness* de la distribution des salaires. Il serait intéressant de se pencher sur l'existence de moyens permettants de corriger cet effet indésirable, ce qui, le cas échéant rendrait cette méthode d'autant plus attractive.
Il est également possible que cette asymétrie soit dûe aux modalités de génération des données, qui font que les individus qui ont rejoint le marché du travail du fait du traitement sont effectivement ceux qui ont les salaires les plus bas de la distribution du groupe de traitement. De ce fait, la réalité correspondrait assez fidèlement aux hypothèses qui sous-tendent le calcul de la borne supérieure.

Tableau récapitulatif des principaux résultats

Approche	$[LB;UB]$	$IC_{95\%}$
Manski et Horowitz (2000)	[-0,43;0,61]	NC
Lee (2009) sans covariates	[0,03;0,22]	[0;0,24]
Lee (2009) avec covariates	[0,04;0,21]	[0,01;0,24]

References

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444–472.

Frangakis, C. E., Rubin, D. B. (2002). Principal stratification in causal inference. Biometrics, 58, 21–29.

Frumento, P., Mealli, F., Pacini, B., and Rubin, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. J. Am. Statist. Ass., 107:450–466.

Grasdal, A. (2001), The performance of sample selection estimators to control for attrition bias. Health Econ., 10: 385–398.

Heckman, J., 1974. Shadow prices, market wages, and labor supply. Econometrica 42, 679–694.

Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error", Econometrica, 47, 153–161.

Heckman, J., 1990. Varieties of selection bias. American Economic Review 80, 313–318.

Horowitz, J. L. and Manski, C. F. (2000), "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data", Journal of the American Statistical Association, 95, 77–84.

Horowitz, J. L. and Manski, C. F. (2000), "Rejoinder: Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data", Journal of the American Statistical Association, 95, 87.

Jiang, Z., Ding, P., and Geng, Z. (2016). Principal causal effect identification and surrogate endpoint evaluation by multiple trials. J. R. Statist. Soc. B, in press.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. The Review of Economic Studies, 76, 1071–1102.

Manski, C. (1989), "Anatomy of the Selection Problem," Journal of Human Resources, 24, 343–360.

Mattei, A., Li, F., and Mealli, F. (2013). Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. Ann. Appl. Stat., 7:2336–2360.

Mattei, A. and Mealli, F. (2007). Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. Biometrics, 63:437–446.

Mattei, A. and Mealli, F. (2011). Augmented designs to assess principal strata direct effects. J. R. Statist. Soc. B, 73:729–752.

Mealli, F. and Pacini, B. (2008). Comparing principal stratification and selection models in parametric causal inference with nonignorable missingness. Computational Statistics and Data Analysis 53, 507-516.

Mealli, F. and Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. J. Am. Statist. Ass., 108:1120–1131.

Puhani, P. (2000), The Heckman Correction for Sample Selection and Its Critique. Journal of

Economic Surveys, 14: 53–68.
Rosenbaum, P., Rubin, D.B., (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41-55.
Yang, F. and Small, D. S. (2016). Using post-quality of life measurement information in censoring by death problems. J. R. Statist. Soc. B, 78:299–318
Zhang, J. L., Rubin, D. B., and Mealli, F. (2009). Likelihood-based analysis of causal effects via principal stratification: new approach to evaluating job-training programs. J. Am. Statist. Ass., 104:166–176.