Testing for Asymmetric Information with Neural Networks

Preliminary*

Serguei Maliar Bernard Salanié

December 31, 2022

Abstract

The positive correlation test for asymmetric information developed by Chiappori and Salanié (2000) has been applied in many contexts. Most of the literature only allows for a constant correlation; and they use parametric specifications for the choice of coverage and the occurrence of claims. Estimating conditional covariances non-parametrically is hard; and testing restrictions on variable correlation coefficients is a highly multiple testing problem. We relax these limitations using deep learning methods. We use a neural network to classify insurees into joint choice of coverage and claim occurrence categories. Then we use the double-debiasing methods of Chernozhukov et al. (2018) to obtain a consistent and asymptotically normal estimator of the correlation function. Finally, we apply the intersection test of Chernozhukov, Lee, and Rosen (2013) to the hypothesis that this correlation function only takes non-negative values. We conclude with an examination of the power of our procedure against small negative correlations when the correlation coefficient is assumed to be constant.

*We thank Simon Lee for his advice. We are grateful to Marc Maliar for his help with writing the TensorFlow code for estimation and testing, and to Xiangru Li for excellent research assistance.

Introduction

Chiappori and Salanié (2000) developed a test for asymmetric information in insurance that has been applied many times since. The test relies on the prediction that under either adverse selection or moral hazard, insuree's choices of coverage should be positively correlated with ex-post measures of their risk, such as the occurrence and severity of claims. Chiappori, Jullien, Salanié, and Salanié (2006) proved that this *positive correlation property* obtains quite generally in models of competitive equilibrium in insurance markets. Chiappori and Salanié (2013) surveys this literature.

In its simplest form, the positive correlation property states that the conditional correlation of coverage and ex-post risk should be positive, for *all* values of the vector of covariates that are observed both by the insurer and the insuree. Since these "public" covariates have at least half-a-dozen relevant dimensions, each with several possible values, this is a tall order. Most papers in this subfield have followed Chiappori and Salanié (2000), which relied on two parametric and one non-parametric procedure. They reasoned that the residuals of two binary choice models for coverage (e.g. minimal versus comprehensive) and ex-post risk (e.g. whether a claim was filed) should be positively correlated, if all public covariates are controlled for. They tested the hypothesis that the correlation of the generalized residuals of two univariate probits was zero. They also estimated a bivariate probit and tested the same hypothesis.

These two procedures are heavily parametric. While Chiappori and Salanié (2000) included 55 regressors in their probit regressions, this was only a very restricted subset of the covariates they could have constructed by interacting their categorical variables. Their bivariate probit also only allowed for a constant correlation. Since these tests depend both on the specification of the regression function and on the Gaussian homoskedastic distribution of the errors, Chiappori and Salanié (2000) also proposed a non-parametric test. They chose m binary covariates and constructed 2^m cells based on their values. In each cell, they computed the χ^2 test of independence of coverage and risk. Finally, they tested the null hypothesis that these 2^m statistics were drawn independently from a $\chi^2(1)$ distribution.

These methods have been extended in several directions. Kim, Kim, Im, and Hardin (2009) showed how the probit for coverage can be replaced by an ordered multinomial choice model when more than two types of contracts are available. Following ideas in Dionne, Gouriéroux, and Vanasse (2001, 2006), Su and Spindler (2013) and Spindler (2014) used a nonparametric test of conditional independence.

Ideally, one would want to do two things: estimate flexibly the conditional correlation of risk and coverage for any given values of the covariates, and test that that it is positive for all values in a given subset (e.g. for all male, 43-year old drivers who use a 7-year old car). To

focus the discussion, suppose that coverage y and ex-post risk z are binary variable, and let X denote all publicly observed covariates. Then the most basic form of the positive correlation property states that for all values of x, the covariance of y and z conditional on $X = x$ is non-negative. We will denote the underlying correlation coefficient as

$$\rho(x) = \frac{\text{cov}(y, z|X = x)}{\sqrt{V(y|X = x)V(z|X = x)}}.$$

Let us denote $p_{jk}(x)$ the probability that $y = j$ and $z = k$ given $X = x$, for $j, k = 0, 1$, and $p_{1\cdot}(x) = p_{10}(x) + p_{11}(x)$ (resp. $p_{\cdot 1}(x) = p_{01}(x) + p_{11}(x)$) the probability that $y = 1$ (resp. that $z = 1$) conditional on $X = x$. Then

$$\rho(x) = \frac{p_{11}(x) - p_{1\cdot}(x)p_{\cdot 1}(x)}{\sqrt{p_{1\cdot}(x)(1 - p_{1\cdot}(x))p_{\cdot 1}(x)(1 - p_{\cdot 1}(x))}}.$$

Estimating this correlation conditionally on given values of the covariates $X = x$ requires estimating the three above probabilities flexibly. This is not a simple task as many interactions between the covariates can have explanatory power. It is notoriously hard, for instance, to model the risk $p_{\cdot 1}(x)$ parsimoniously. Testing that ρ is non-negative over a subset of covariates is even harder: interesting subsets typically are very large, leading to a multiple testing problem where and the distributions of the estimated correlations for different x are definitely not independent.

Our purpose here is to apply machine learning to the estimation of the conditional probabilities $p_{jk}(x)$ and to rely on recent developments on double debiasing (Chernozhukov et al. (2018)) and on intersection tests (Chernozhukov, Lee, and Rosen (2013)) to test the positive correlation property. We also explore the power of our test against alternatives of the form $\rho(x) \equiv r\bar{h}o < 0$.

For simplicity, we focus on the basic binary-binary model above, which has been the focus of the largest part of the literature[1]. We use machine learning as just a type of model selection which yields good predictors, in the form of the estimated conditional probabilities $p_{jk}(x)$. There is a large (and expanding) catalog of machine learning on the market. We decided to use deep learning in order to explore how it can be used on a relatively small dataset whose size is typical of many microeconometric applications. More precisely, we will use feedforward neural networks with several hidden layers. We plan to explore other machine learning methods and in particular gradient boosting in future work.

Our method has three steps. First (Section 2) we fit a neural network to classify insurees into the four alternatives $y, z = 0, 1$. Then we use the double-debiasing methods of Chernozhukov

[1]As explained in Chiappori, Jullien, Salanié, and Salanié (2006) and Chiappori and Salanié (2013), the precise statement of the property also needs to be adapted in more general models.

et al. (2018) to fit a parametric model $\rho(x) = f(x, \beta)$ to the fitted probabilities of the alternatives; this gives us a consistent and asymptotically normal estimator of β (Section 3). Finally, Section 4 runs the intersection test of Chernozhukov, Lee, and Rosen (2013). We conclude with an examination of the power of our procedure against small negative correlations when the correlation coefficient is assumed to be constant under the alternative (Section 5).

1 The Data

Chiappori and Salanié (2000) obtained the data for their paper from the French federation of insurance companies, which ran a survey of automobile insurance in 1989. They selected a subsample that only includes "young drivers"—where "young" refers not to age, but to insurees who obtained their driver's licence within the past three years. We focus on an even narrower subsample, of insurees whose driving license is one year old at most. Since these individuals have no previous driving history, this removes concerns about the impact of experience rating on driving behavior; it also reduces the unobserved heterogeneity in the sample.

Our selection leaves us with a sample of 6,330 observations. Each observation includes information on the car (brand, model, age, power, ...), the client's demographics (age, profession, residence,...), the type of contract, and the claim record. Like Chiappori and Salanié (2000), we code the latter two variables as binary. In automobile insurance, the main distinction between contracts is whether they only include third-party (liability) coverage—which is compulsory in France—or they also cover damages that the insuree caused to his/her car. We will call the latter "comprehensive coverage" and we will neglect variations within this class, such as the amount of the deductible. Since the difference between third-party and comprehensive coverage only matters when the insuree is at fault, we define our claim variable accordingly. This results in the two binary variables $y, z = 0, 1$:

- $y = 1$ if the insuree opted for comprehensive coverage
- $z = 1$ if the insuree filed at least one at-fault claim.

We also define four indicator variables as $y_{jk} = 1$ if ($y = j$ and $z = k$) for $j, k = 0, 1$.

Table 1 shows how the 6,330 observations fall in the four corresponding alternatives.

The data on each insuree and car is quite rich. We use the same set of covariates x as Chiappori and Salanié (2000); they are created from the nine variables that insurers told us are the most important. We have (potentially) 36.000 categories of insurees: nine age categories, ten professions, four types of use, ten regions, five rural-to-urban codes, and gender; and 72 car categories: six categories for the performance of the car, and twelve for its age. Combining

Event	y	z	Alternatives	Number of observations
Third-party, no accident	0	0	$y_{00} = 1$	3,695
Third-party, accident	0	1	$y_{01} = 1$	301
Comprehensive, no accident	1	0	$y_{10} = 1$	2,202
Comprehensive, accident	1	1	$y_{11} = 1$	132
Total				6,330

Table 1: Observed Classification

them would yield more than 2.5 million dummy variables, a number that dwarfs the sample size. Even if we had a much larger dataset, there are many more variables in the data. This is a clear-cut "$p \gg n$" case, which calls for model selection.

The data records contracts that covered all or part of the year. Only about 40% of insurees were insured throughout the year, and roughly 15% were covered for less than two months. Like Chiappori and Salanié (2000), we use sampling weights w that represent the number of days that the insuree was insured during the year. We will indicate it in the notation with w subscripts when needed.

2 The Deep Learning Model

We first fit the bivariate probabilities $p_{jk} = \Pr(y = j, z = k|X = x)$ for $j, k = 0, 1$ with a neural network. The values of the nine covariates x enter the input layer. The neural network has D hidden layer, each with W neurons; the last hidden layer feeds into an output layer that consists two neurons, one for each of the alternatives $y, z = 0, 1$. It generates scores $s_{jk}(x)$ for $j, k = 0, 1$. We use the "softmax" function to turn them into 4-way probabilities:

$$p_{jk}(x) \equiv \frac{e^{s_{jk}(x)}}{e^{s_{00}(x)} + e^{s_{01}(x)} + e^{s_{10}(x)} + e^{s_{11}(x)}}. \tag{1}$$

We train the neural network with the Adam optimizer (Kingma and Ba, 2014) and the cross-entropy loss function, with our sampling weights: for a sample of observations $(x_i, y_i, z_i, w_i)_{i\in I}$, the loss is

$$L = \sum_{i\in I} w_i \sum_{j,k=0,1} \mathbb{1}(y_i = j, z_j = k) \log p_{jk}(x_i).$$

The neural network is implemented in the Python package Keras (Chollet, 2021) with the TensorFlow backend (Abadi et al., 2016). The neural network is trained for 100 epochs with a batch size of 32; we use m-fold validation.

The class of neural networks we consider has $p \equiv n_X W + (D-1)W^2 + 3W$ parameters, where

n_X is the number of covariates in the input layer. Adding up the number of categories (minus one) of our nine covariates, plus the constant, gives $n_X = 54$. This gives $p = W(57+(D-1)W)$. To illustrate this, a very modest neural network with $D = 2$ hidden layers of $W = 16$ neurons each has $p = 1,168$ parameters. With such a large number of parameters, overfitting is an obvious concern. While m-fold validation helps mitigate it, experience shows that it is not enough by itself. The neural network has many hyperparameters: its depth D and width W, the number of epochs of training, the number of folds n, the size of the mini-batches... These must also be selected while guarding against overfitting. A first approach is to optimize in hyperparameter space: we create a list of models and we select the one that gives the best fit on a test sample. This "hyperparameter tuning" can be very costly in large neural networks, which take days to train. An alternative approach, that has proved popular, specifies a sufficiently complex model upfront and uses dropout regularization (Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov, 2014) to attain the highest possible accuracy on the validation samples.

Dropout regularization consists of randomly dropping out some of the neurons in the hidden layers during training. The idea is that the network will learn to compensate for the missing neurons, and that the resulting model will be more robust to overfitting. The fraction d of neurons that are dropped out is called the dropout rate.

Since our sample size is relatively small, we will only fit smallish neural networks. This allows us to combine hyperparameter tuning and dropout regularization. More precisely, we estimated neural networks that varied in:

- the number D of hidden layers: we tried 1, 2 and 3 hidden layers;
- the number W of neurons in each layer: we experimented with values from 2 to 64;
- the dropout rate d: we tried values from 0 (no dropout) to 0.6.

We also experimented changing the activation function; as this had very little effect, we settled on the sigmoid and we won't report the results for other activation functions. For the same reason, we opted for 5-fold validation ($m = 5$). We found that using more than two hidden layers does not improve the fit, but results in overfitting. Furthermore, using more than 10 neurons per layer also leads to overfitting in the absence of dropout regularization. Dropout deals well with overfitting and allows to use a more powerful network with a larger number of neurons. The optimal dropout rate increases with the number of neurons in each hidden layer, as a higher dropout rate is needed to deal with the overfitting. Our final choice is a neural network with $D = 2$ hidden layers, each with $W = 32$ neurons, and a dropout rate of $d = 0.2$. We will use this model for all subsequent results in the paper.

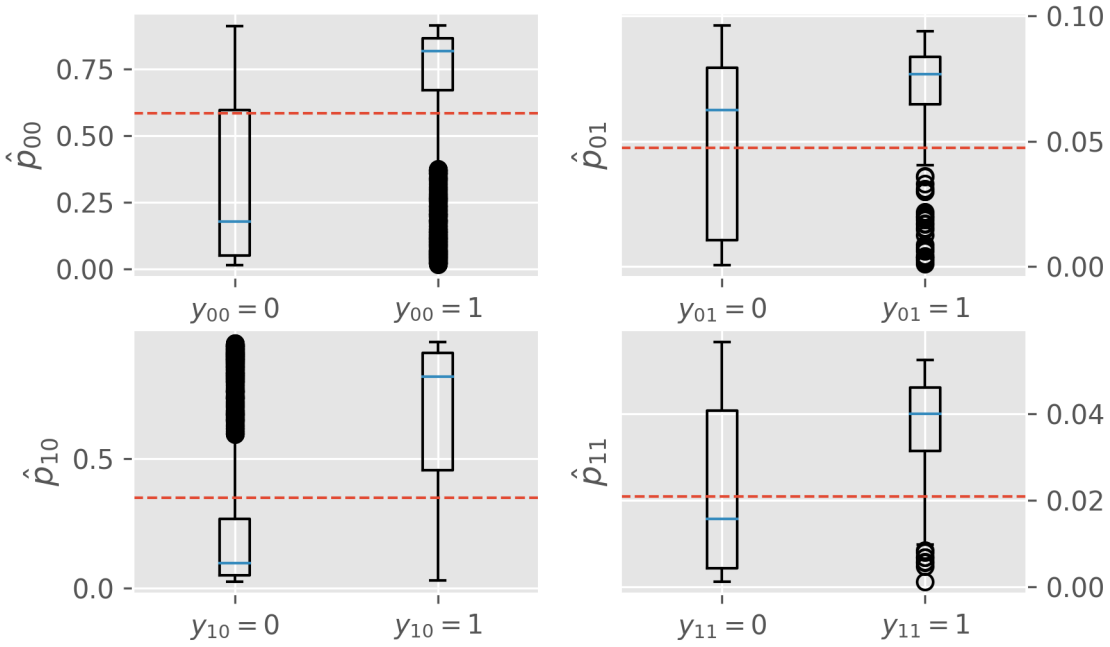

Figure 1: Fitting the y_{jk} variables

Figure 1 plots the estimated probabilities $\hat{p}_{jk}$, both when the corresponding y_{jk} is 0 and when it is 1. The dashed red line represents the mean of y_{jk} in the sample. A perfect fit would have $\hat{p}_{jk} = y_{jk}$ within each panel. It is clear that the left column performs better than the right column in this respect. It is not that surprising as there are more than 15 times as many observations with $z = 0$ as with $z = 1$: the neural network puts a large weight on fitting these observations. Figure **??** shows that the model can predict the purchase of coverage considerably better than the accident occurence. This is a common finding with insurance data. It is more surprising that the neural network overestimates the probability of a claim to the extent showed by the right panel. Again, this seems to be a side-effect of its efforts to fit the choice of contract (the variable y in the left panel).

In a linear model, we could use partial R^2's to evaluate the contribution of various covariates in explaining then left-hand side variable. In a neural network, a natural alternative is to train the model again while omitting one covariate and to measure the additional loss. This is a very partial indication, as the neural network interacts different groups of variables in potentially complex ways. Still, it is a reasonable starting point.

Our application has eight groups of variables; accordingly, we ran eight neural networks that each omits one of these groups. Figure 3 plots the results. A larger positive number denotes

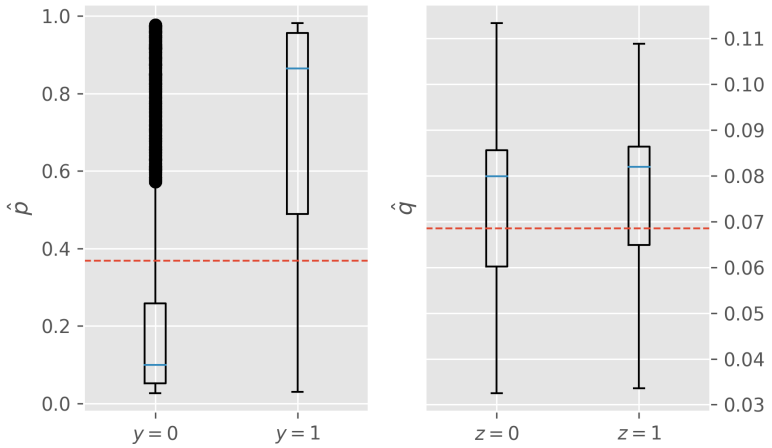

Figure 2: Fitting the y and z variables

that this group of variables contributes more to the quality of the fit, which we measure both for each of the forur subsamples $y_{jk} = 1$ and for the full sample.

Only three groups of variables contribute (taken by themselves) to the fit: the rural/urban zone, the gender of the insuree, and most importantly the age of the car. The latter simply shows that drivers of older cars are less likely to buy comprehensive coverage (hence the large contribution of the car age dummies to explaining the variable y).

3 Estimation of ρ

A naive estimate of the conditional correlation coefficient would be the ratio of the weighted conditional covariance to the product of the weighted conditional standard errors:

$$\hat{\rho}_n(x) = \frac{\widehat{\mathrm{cov}_w}(y, z|X = x)}{\hat{\sigma}_w(y|X = x)\hat{\sigma}_w(z|X = x)}.$$

Given our estimates $\hat{p}_{jk}$ of the probabilities of the four alternatives, we can substitute

$$\hat{\rho}_n(x) = \frac{E\left(w(y - \hat{p}(x))(z - \hat{q}(x))|X = x\right)}{\sqrt{E\left(w(y - \hat{p}(x))^2|X = x\right) E\left(w(z - \hat{q}(x))^2|X = x\right)}}.$$

where $\hat{p}(x) \equiv \hat{p}_{10}(x) + \hat{p}_{11}(x)$ estimates $\Pr(y = 1|X = x)$ and $\hat{q}(x) \equiv \hat{p}_{01}(x) + \hat{p}_{11}(x)$ estimates $\Pr(z = 1|X = x)$.

In these equations, the neural network estimates $\hat{p}_{jk}(x)$ act as nuisance parameters. Their

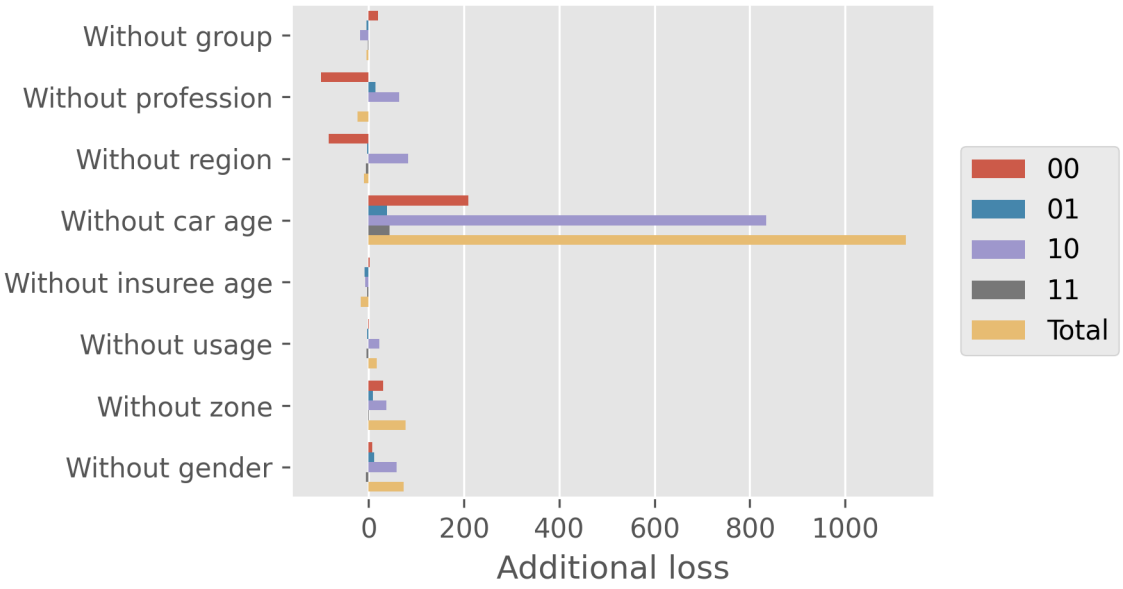

Figure 3: Omitting groups of variables

relatively slow rate of convergence invalidates standard inference procedures on the parameters of the correlation coefficient ρ, as well as tests of the positive correlation property. To remedy this, we will use the double-debiasing method of Chernozhukov et al. (2018), which extends the idea of Neyman orthogonalization to a much broader range of models and estimation methods.

3.1 Double debiasing

Suppose for simplicity that we seek to fit a parametric model for the correlation ρ:

$$\rho_0(x) = f(x, \beta_0)$$

for a known function f and unknown parameters $\beta_0 \in \mathbb{R}^q$. The intuition of Neyman orthogonalization can be described simply. Suppose that the nuisance parameters $\eta = (p_{00}, p_{01}, p_{10}, p_{11})$ and the parameters of interest β are estimated from a set of equations $\hat{M}(\hat{\beta}, \hat{\eta}) = 0$, where $\hat{M}$ converges to M and $M(\beta_0, \eta_0) = 0$. Given an appropriate set of assumptions, the standard Taylor expansion around the true values (β_0, η_0) gives

$$\nabla_\beta \hat{M}(\beta_0, \eta_0)(\hat{\beta} - \beta_0) \simeq -\hat{M}(\beta_0, \eta_0) - \nabla_\eta \hat{M}(\beta_0, \eta_0)(\hat{\eta} - \eta_0).$$

The presence of the second term on the right-hand side is what makes the η nuisance parameters: the estimation error insurance $\hat{\eta}$ contaminates the asymptotic distribution of $\hat{\beta}$. To get rid of the nuisance, we need to change the estimating equation to one for which the gradient of M with respect to η is zero at the true values (β_0, η_0). This can be done by projecting the estimating equation on the orthogonal subspace to the gradient $\nabla_\eta \hat{M}(\beta_0, \eta_0)$.

In our application, the nuisance parameters are (after dropping the first alternative $y = z = 0$) the probabilities $\eta_0 = (p_{01}, p_{10}, p_{11})$, which are functions of the covariates x; and the parameters of interest $\beta_0 \in \mathbb{R}^q$ define $\rho_0(x) = f(x, \beta_0)$. The moment conditions that define η_0 and β_0 can be written as

$$E\left(M(y, z, w, \rho(x), \eta(x) \mid X = x\right) = 0$$

where $M = (m_{01}, m_{10}, m_{11}, m_\beta)$ and

$$\begin{aligned}
m_{01} &= w\left((1-y)z - p_{01}(x)\right) \\
m_{10} &= w\left(y(1-z) - p_{10}(x)\right) \\
m_{11} &= w\left(yz - p_{11}(x)\right) \\
m_\beta &= w\left((y - p(x))(z - q(x)) - f(x, \beta_0)R(x)\right)
\end{aligned}$$

where $p = p_{10} + p_{11}$, $q = p_{01} + p_{11}$, and $R(x)^2$ is the product of the weighted variances of y and z conditional on x:

$$R(x)^2 \equiv \frac{E\left(w(y - p(x))^2 | X = x\right)}{E(w|X = x)} \frac{E\left(w(z - q(x))^2 | X = x\right)}{E(w|X = x)} = p(x)(1 - p(x))q(x)(1 - q(x)).$$

Note that since we used weights to define the p_{jk} and therefore p and q, the sampling weights do not explicitly appear in this formula[2].

Given neural network predictors $\hat{\eta}(x)$, the double debiasing procedure of Chernozhukov et al. (2018) modifies the moment condition $\hat{E}\left(m_\beta(x, \hat{\beta}, \hat{\eta}) | X = x\right) = 0$ into the following:

$$\sum_{i=1}^{n} \psi\left(y_i, z_i, f(x_i, \hat{\beta}), \hat{\eta}(x_i)\right) = 0,$$

where ψ is a well-chosen vector of q linear combinations of the four moment conditions in M:

$$\psi = \mu_{01} m_{01} + \mu_{10} m_{10} + \mu_{11} m_{11} + \mu_\beta m_\beta.$$

More precisely, the $(4, q)$ matrix $\mu \equiv (\mu'_{01}, \mu'_{10}, \mu'_{11}, \mu'_\beta)'$ is given by

$$\mu(x) = \Omega^{-1}(S - P_{\Gamma(x)} S)$$

where S is any $(4, q)$ moment selection matrix; Ω is any $(4, 4)$ positive definite matrix weighting matrix; $\Gamma(x)$ is the $(4, 3)$ matrix $\nabla_\eta E(M|X = x)$; and $P_{\Gamma(x)}$ is the projection on $\Gamma(x)$ for the metric given by Ω:

$$P_\Gamma = \Gamma\left(\Gamma'\Omega^{-1}\Gamma\right)^{-1}\Gamma'\Omega^{-1}$$

so that μ projects the selected moment condition on the space that is orthogonal to $\nabla_\eta M$.

[2]As an example,

$$\begin{aligned} E(w(y - p(x))^2 | X = x) &= E(wy^2 | X = x) - 2p(x)E(wy|X = x) + p(x)^2 E(w|X = x) \\ &= E(wy|X = x)(1 - 2p(x)) + p(x)^2 E(w|X = x) \\ &= E(w|X = x)p(x)(1 - p(x)). \end{aligned}$$

3.1.1 Choosing S and Ω

The matrix Γ is clearly a function of x. The matrices S and Ω may be chosen to be constant, or they can also be functions of x. In fact, the semiparametrically efficient choice has

$$S^*(x) = \nabla_\beta E(M|X=x) \text{ and } \Omega^*(x) = E(MM'|X=x),$$

computed at the true values (β_0, η_0); both matrices S^* and Ω^* depend on the values of the covariates.

In the definitions of $\Gamma(x)$ and of the semiparametrically efficient moment selection matrix $S^*(x)$, the gradients are taken at the true value for this particular value of x. These two matrices have a very simple form. The first three rows of $S^*(x)$ are zero since β only appears in m_β. Its fourth row is

$$\frac{\partial E(m_\beta|X=x)}{\partial \beta} = -R(x)E(w|X=x)\frac{\partial f}{\partial \beta'}(x,\beta).$$

Since $S^*(x)$ has such a simple form, we always use it in our inference procedure.

The semiparametrically efficient $\Omega^*(x)$ is more complex. Still, it is easy to evaluate numerically. Taking Ω to be the four-dimensional identity matrix I_4 does not achieve the semiparametric efficiency bound, but it still gives a consistent estimator, and the calculations are more transparent. We will use both choices $\Omega = I_4$ and $\Omega = \Omega^*$ in our simulations.

3.1.2 The Debiased Moment Conditions

For illustrative purposes, we give here the formula for the doubly debiased moment conditions when Ω is taken to be the identity matrix I_4. Let us denote $f_\beta(x,\beta)$ the gradient of $\rho(x)$ in β, and define $T_{jk}(x) = R(x)R_{jk}(x)$ for $j,k = 0,1$. Define the natural estimators of the naive moment conditions:

$$\begin{aligned}
\hat{m}_\beta(x,\beta) &= w\left((y-\hat{p}(x))(z-\hat{q}(x)) - f(x,\beta)\hat{R}(x)\right)\\
\hat{m}_{01}(x) &= w((1-y)z - \hat{p}_{01}(x))\\
\hat{m}_{10}(x) &= w(y(1-z) - \hat{p}_{10}(x))\\
\hat{m}_{11}(x) &= w(yz - \hat{p}_{11}(x));
\end{aligned}$$

Using plug-in estimates for $\hat{R}$ and the $\hat{T}_{jk}$, we show in the Appendix that our estimator of β solves

$$\sum_{i=1}^{n} \psi(x_i, \hat{\beta}) = 0$$

where

$$\psi(x,\beta) = \Big(f(x,\beta_0)(\hat{T}_{01}(x)\hat{m}_{01}(x) + \hat{T}_{10}\hat{m}_{10}(x) + \hat{T}_{11}\hat{m}_{11}(x)) - \hat{R}(x)\hat{m}_\beta(x,\beta) \Big) \hat{V}(x), \qquad (2)$$

where the q-vector $\hat{V}(x)$ is a consistent estimator of

$$V(x) = \frac{R^2(x)E(w|X=x)}{R^2(x) + f^2(x,\beta_0)(T_{01}^2(x) + T_{10}^2(x) + T_{11}^2(x))}.$$

In practice, we use

$$\hat{V}(x) = \frac{\hat{R}^2(x)\hat{E}(w|X=x)}{\hat{R}^2(x) + f^2(x,\hat{\beta})(\hat{T}_{01}^2(x) + \hat{T}_{10}^2(x) + \hat{T}_{11}^2(x))}.$$

Under the hypothesis that the function $\rho(x) = f(x,\beta)$ is well-specified, the estimates of its parameters $\hat{\beta}$ are $\sqrt{n}$-consistent and asymptotically normal.

To compute their variance-covariance, we use the sandwich estimator, as recommended by Chernozhukov et al. (2018). We first evaluate the matrices

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i,\hat{\beta})\psi(x_i,\hat{\beta})'$$

and

$$\hat{J} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial\psi}{\partial\beta}(x_i,\hat{\beta}).$$

Then we estimate the variance-covariance matrix of $\hat{\beta}$ with $\hat{J}^{-1}\hat{I}\hat{J}^{-1}/n$. Finally, we use the delta method to obtain the variance of $\hat{\rho}(x)$ at any point x:

$$\hat{\sigma}_\rho^2(x) = V\hat{\rho}(x) = \frac{\partial f}{\partial\beta}(x,\hat{\beta})\; V\hat{\beta}\; \frac{\partial f}{\partial\beta}(x,\hat{\beta})'. \qquad (3)$$

3.2 Estimation Results

Figure 4 plots the density of our estimated $\hat{\rho}_i$ over the sample. To give an idea of the sampling variation, we also plot the density of $\hat{\rho}_i \pm 1.96\sigma(\hat{\rho}_i)$, where $\sigma(\rho)$ estimates $E(\hat{\sigma}_\rho|\hat{\rho}_i = \rho)$. The density of the estimated $\hat{\rho}_i$ is negatively skewed; more than 60% of the estimates are negative. The correlations are weak, however: 99% of the mass is in the $[-0.2, 0.2]$ interval. The correlations are estimated fairly precisely, with standard errors of the order of 0.1.

While men (young men especially) are notoriously riskier prospects than women, no such difference emerges in their correlation coefficients. In fact, the densities of $\hat{\rho}_i$ are remarkably

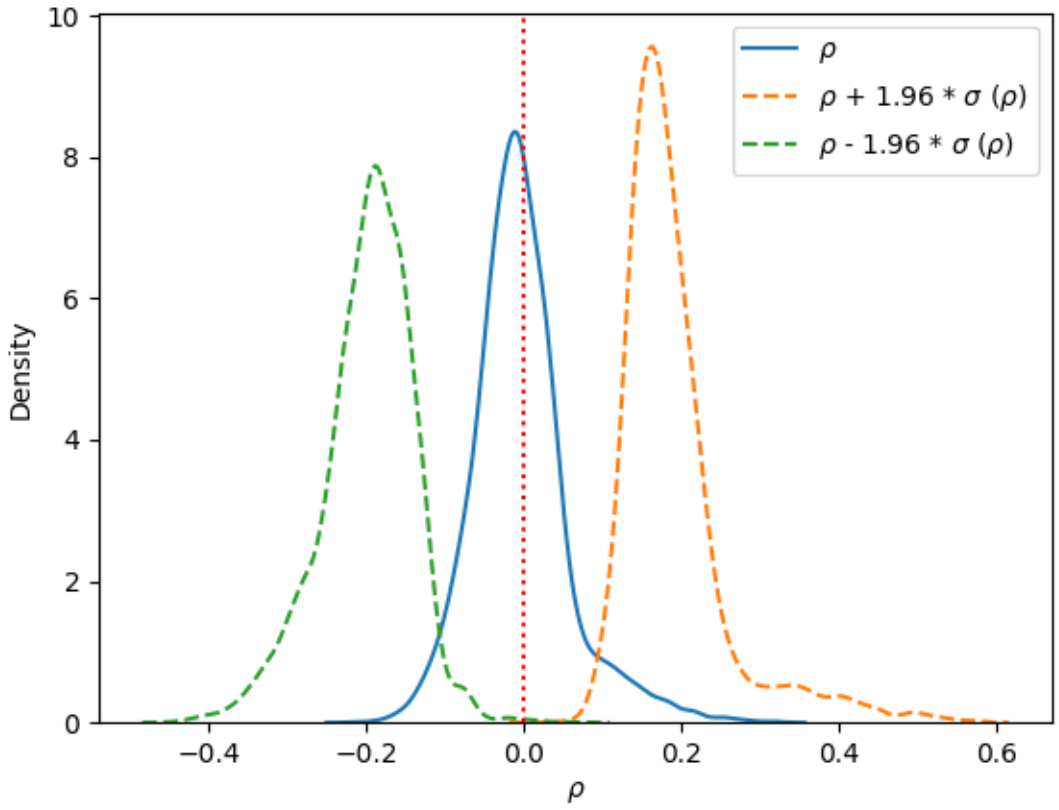

Figure 4: Density of the Estimated $\rho(x)$

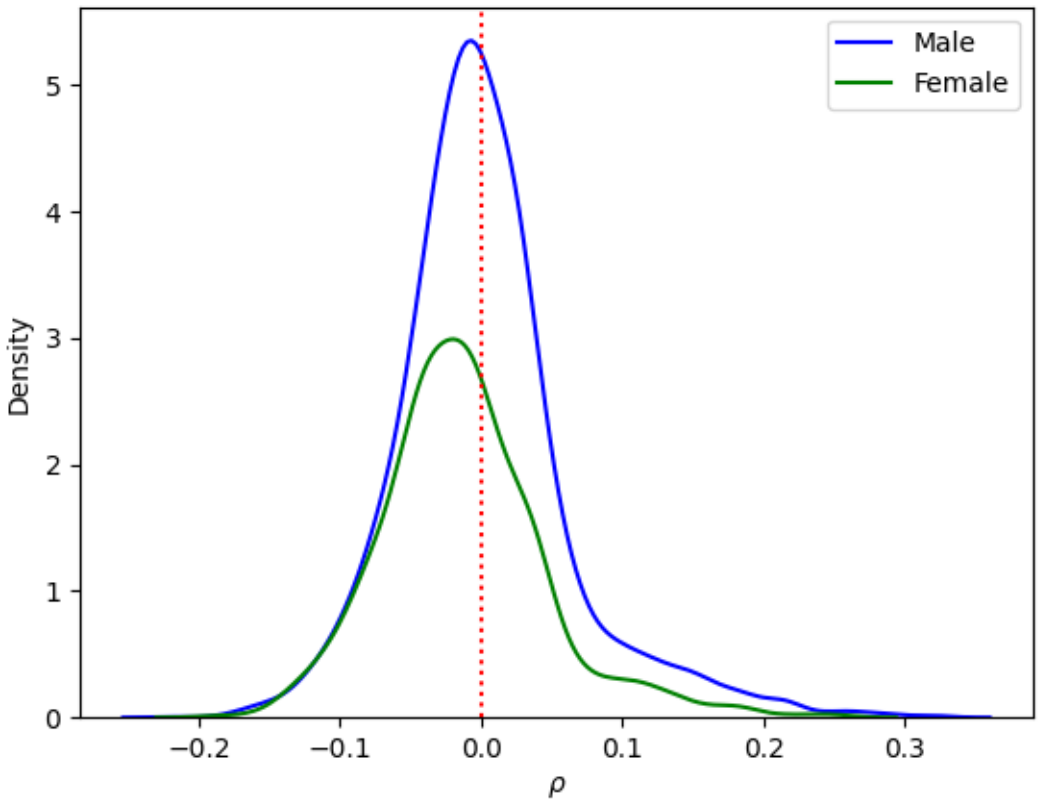

Figure 5: Density of the Estimated $\rho(x)$ by Gender

similar across genders[3], as seen in Figure 5.

There is more heterogeneity across ages. Figure 6 suggests that insurees in their twenties have a more negative correlation than older insurees. Still, the amplitude of these variations is quite small.

4 Testing the Positive Correlation Property

Our ultimate goal is to test the hypothesis that the correlation of y and z is non-negative conditional on *any* value of the covariates x. For any given domain A, the hypothesis

$$\inf_{x \in A} \rho_0(x) \geq 0$$

[3]The difference in mass reflects the fact that (in 1988 France) it is more often the man who holds the insurance contract in a couple.

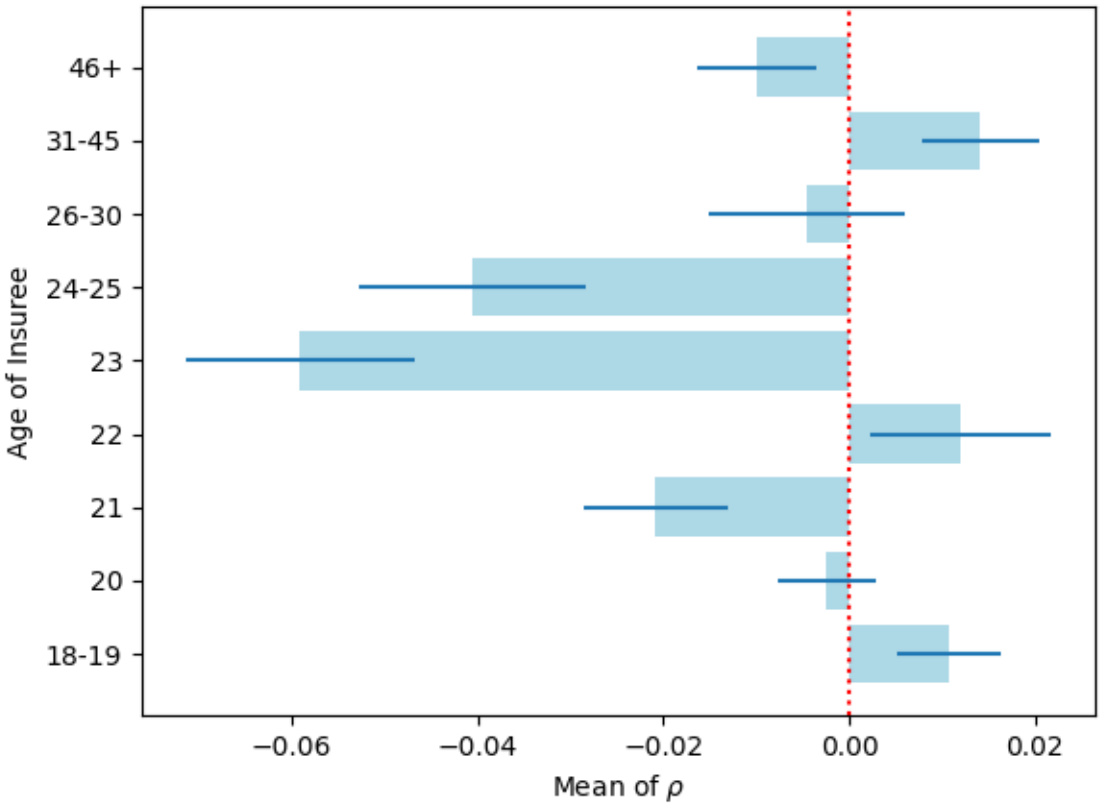

Figure 6: Mean of the Estimated $\rho(x)$ by Insuree Age

is an *intersection hypothesis*. Given our parametric specification for $\rho(x)$, a natural estimate for the left-hand side is

$$\hat{\rho}_m = \inf_{x \in A} f(x, \hat{\beta}),$$

where $\hat{\beta}$ is the doubly debiased estimator we obtained in Section 3. One look at Figure 4 shows that $\hat{\rho}_m$ is negative; in our sample, it equals -0.23. This would suggest a clear rejection of the positive correlation property. The error in estimating β makes $\hat{\rho}_m$ a very bad choice of test statistic, however: the infimum is pulled downwards by sampling variation.

To avoid overrejecting the positive correlation property, we need to choose our test statistic more carefully. Following Chernozhukov, Lee, and Rosen (2013), we will reject the positive correlation property at level α if $\underline{\rho} < 0$, where

$$\underline{\rho} = \inf_{x \in \hat{A}} \left(\hat{\rho}(x) + k \hat{\sigma}_\rho(x) \right). \tag{4}$$

In this formula, $\hat{\sigma}_\rho(x)$ denotes the estimated standard error of $\hat{\rho}(x) \equiv f(x, \hat{\beta})$ from Section 3. The value of k and the subset $\hat{A} \subset A$ are selected as follows:

1. for each sample point $x_i \in A$, we compute the q-vector

$$h_i = (V\hat{\beta})^{1/2} \frac{\partial f}{\partial \beta}(x_i, \hat{\beta});$$

2. we draw a large number R of iid values η_r from $N(0, I_q)$;

3. for each $r = 1, \ldots, R$, we define

$$\bar{h}_r = \max_{x_i \in A} \frac{h_i' \eta_r}{\|h_i\|};$$

4. we let k_0 be the γ_n quantile of the R numbers $\bar{h}_r$, with $\gamma_n = 1 - 0.1/\log n$;

5. we define $\hat{A}$ to be the set of sample points x_i such that

$$\hat{\rho}(x_i) \leq \min_{x_j \in A} \left(\hat{\rho}(x_j) + k_0 \frac{\|h_j\|}{\sqrt{n}} \right) + 2k_0 \frac{\|h_i\|}{\sqrt{n}};$$

6. finally, we let k be the $(1-\alpha)$-quantile of the values

$$\hat{h}_r = \max_{x_i \in \hat{A}} \frac{h_i' \eta_r}{\|h_i\|}.$$

In our application k_0 is close to 4.5, and the set $\hat{A}$ only contains 12 observations. The values

of k and $\underline{\rho}$ are given in Table 2. The values of k are much larger than the critical values that a standard normal approximation would suggest. The corresponding test statistic $\underline{\rho}$ is positive at all usual levels, so that the positive correlation property is not rejected.

Level	k	$\underline{\rho}$
1%	4.56	0.127
2.5%	4.32	0.115
5%	4.13	0.105
10%	3.91	0.094

Table 2: The intersection test for $\rho \gg 0$

Out of curiosity, we also ran the intersection test for the hypothesis that $\rho(x)$ is negative for all x. This rejects the (hypothetical) *negative* correlation property at level α if $\overline{\rho} > 0$, where

$$\overline{\rho} = \max_{x \in \hat{A}'} \left(\hat{\rho}(x) - k' \hat{\sigma}_\rho(x) \right) \tag{5}$$

and $\hat{A}'$ and k' are selected by a procedure very similar to the one we explained above. k'_0 is smaller this time, ast 2.26, and $\hat{A}'$ has 3 elements only. Table 3 gives the results. This hypothesis is not rejected either. One could also combine these tests to evaluate a third hypothesis: that

Level	k'	$\overline{\rho}$
1%	2.51	-0.095
2.5%	2.17	-0.080
5%	1.84	-0.053
10%	1.48	-0.022

Table 3: The intersection test for $\rho \ll 0$

all values of $\rho(x)$ have the same sign. By the Bonferroni inequality, this combined hypothesis is rejected at level α if $\underline{\rho} > 0$ at level $\alpha/2$ and $\overline{\rho} < 0$ at level $\alpha/2$. Using Tables 2 and 3 shows that the "same sign" hypothesis is rejected at all levels above 2%.

The intuition for these mixed results is simple and can be grasped from Figure 4. With such a smallish sample size, the standard error on the estimates of $\rho(x)$ is of the order of 0.1. Our estimate of the smallest $\rho(x)$ is

$$\underline{\rho} = \min_{x \in \hat{A}} \left(\hat{\rho}(x) + k \hat{\sigma}_\rho(x) \right)$$

and that of the largest $\rho(x)$ is

$$\overline{\rho} = \max_{x \in \hat{A}'} \left(\hat{\rho}(x) - k' \hat{\sigma}_\rho(x) \right).$$

Since most of the variation of the estimated $\rho(x)$ is in $[-0.2, 0.2]$, even with relatively small k and k' the min and the max are in inverse order. This is in contrast with the parametric specifications used in the literature, which come with smaller standard errors. Here a larger sample size is needed to reach more convincing conclusions.

5 The power curve

While the sequence of procedures we have been using is designed to yield a test with the correct size, we wanted to evaluate its power to detect violations of the positive correlation property.

5.1 A Generic Bootstrap Procedure

There are many ways we could parameterize alternatives when $\rho(x)$ may vary with x. To evaluate the power of our testing procedure against an alternative given by a function $x \to \rho(x) = r(x) < 0$, we need to simulate distributions of (y, z) with a correlation coefficient $r(x)$. The most natural way to do it is to fix the marginal distributions of y and of z at their estimated values and to impose their correlation. For any value of the covariates x, we fix $p_{10}(x) + p_{11}(x) = \hat{p}(x)$ and $p_{01}(x) + p_{11}(x) = \hat{q}(x)$, where $\hat{p}$ and $\hat{q}$ are the predictions of the neural network of Section 2. Recall that the correlation coefficient is

$$\rho(x) = \text{corr}(y, z|x) = \frac{p_{11}(x) - p(x)q(x)}{R(x)}$$

with $R(x) = \sqrt{\hat{p}(x)(1-\hat{p}(x))\hat{q}(x)(1-\hat{q}(x))}$. To set it at a value $r(x)$ we use the following formulæ:

$$\begin{aligned}
p_{11}(x) &= \hat{p}(x)\hat{q}(x) + r(x)\hat{R}(x) \\
p_{01}(x) &= \hat{q}(x) - p_{11}(x) \\
p_{10}(x) &= \hat{p}(x) - p_{11}(x) \\
p_{00}(x) &= 1 - \hat{p}(x) - \hat{q}(x) + p_{11}(x).
\end{aligned} \tag{6}$$

This can only be done if all of these four numbers are between zero and one, however. This requirement corresponds to the Fréchet-Hoeffding inequalities:

$$\max(0, \hat{p}(x) + \hat{q}(x) - 1) \leq p_{11}(x) \leq \min(\hat{p}(x), \hat{q}(x)).$$

Simple calculations show that in our setting, they constrain $r(x)$ to be chosen within the interval

$$-\min\left(\hat{a}(x), \frac{1}{\hat{a}(x)}\right) \leq r(x) \leq \min\left(\hat{b}(x), \frac{1}{\hat{b}(x)}\right) \tag{7}$$

where $\hat{a}(x) = \sqrt{\frac{\hat{p}(x)\hat{q}(x)}{(1-\hat{p}(x))(1-\hat{q}(x))}}$ and $\hat{b}(x) = \sqrt{\frac{\hat{p}(x)(1-\hat{q}(x))}{\hat{q}(x)\hat{p}(x)}}$.

The upper bound of the interval is always positive and we can neglect it since we are interested in alternatives $r < 0$. Suppose we want to evaluate the power of our procedure at an alternative correlation function r that is above the lower bound of the interval for every observed value of x. We would

1. compute the $p_{jk}(x_i)$ for each observation i from the formulæ above
2. for a large number of simulation runs $l = 1, \ldots, L$:
   (a) generate samples (y_i^l, z_i^l) from random draws using the probabilities from step 1
   (b) then estimate new $\hat{p}_{jk}^l(x)$ using the neural network of Section 2
   (c) estimate $\hat{\rho}^l$ as per Section 3
   (d) and run the intersection test at 5% to get an estimate $\underline{\rho}^l$ of the minimum of the ρ^l correlation function;
3. then count the proportion of simulation runs $l = 1, \ldots, L$ for which $\underline{\rho}^l < 0$.

The proportion obtained in step 3 represents the power of the test that $\rho(x) \geq 0$ for all x against the alternative that $\rho(x) = r(x) < 0$ for all x.

5.2 The Case of a Constant Correlation

For simplicity, we focus for now on the maintained hypothesis that the correlation coefficient ρ does not vary with x; that is, the function $f(x, \beta)$ of Section 3 is simply $\rho = \beta$. There is of course no need for the intersection test when ρ is constant; we can simply test the hypothesis that $\rho \geq 0$ by running a one-sided Student test. We reject the positive correlation hypothesis at the 5% level if

$$\hat{t} \equiv \frac{\hat{\rho}}{\hat{\sigma}_\rho} < -1.64.$$

This test statistic is asymptotically pivotal as $\hat{\rho}$ has standard asymptotics. As such, it can be improved by boostrapping it. This can be done with a nonparametric bootstrap:

1. draw B samples from the (y, z, x) data with replacement

(a) on each such sample b, estimate ρ^b by the same method to get $\hat{\rho}^b$, $\hat{\sigma}^b_\rho$, and the Student statistic $\hat{t}^b = \hat{\rho}^b / \hat{\sigma}^b_\rho$

2. compute C, the 5% lower percentile of the $(\hat{t}^b)_{b=1}^B$

3. finally, reject the null hypothesis at 5% iff $\hat{t} < C$.

The drawback is that for each new sample generated in step 1, we need to fit the neural network anew. A less time-consuming alternative consists in replacing step 1 with

1'. keep the covariates x as in the observed sample and draw B samples of (y, z) from the probabilities $\hat{p}_{jk}(x)$ estimated in Section 2.

We call it the "parametric bootstrap", even though the probabilities are generated by a neural network.

We applied power calculations to the one-sided t-test with the three critical values: the asymptotic approximation; the nonparametric bootstrap; and the parametric bootstrap.

5.2.1 Testing the null hypothesis for the original data

We start by estimating a constant value of ρ on the original dataset. When we use the optimal Ω matrix, we obtain $\hat{\rho} = -0.007$ and $\hat{\sigma}_\rho = 0.025$; the corresponding Student statistic $\hat{t}$ is -0.28, well above the standard critical value of -1.64. Therefore the positive correlation property is not rejected when we use standard asymptotics[4]. To get a better handle on the properties of this testing procedure, we ran four versions of a small ($B = 100$) bootstrap, using both our parametric and our nonparametric bootstrap, with the optimal matrix $\Omega = \Omega^*$ and with the identity matrix $\Omega = I_4$. Table 4 shows that the standard test rejects about twice more often than it should. This makes our non-rejection of the positive correlation property even more robust.

Bootstrap experiment	Bootstrapped p-value
$\Omega = \Omega^*$, nonparametric bootstrap	0.12
$\Omega = I_4$, nonparametric bootstrap	0.09
$\Omega = \Omega^*$, parametric bootstrap	0.09
$\Omega = I_4$, parametric bootstrap	0.13

Table 4: Bootstrapping the standard test

[4]With $\Omega = I_4$ we get very similar results: $\hat{\rho} = -0.004, \hat{\sigma}_\rho = 0.015$, and a Student statistic -0.24.

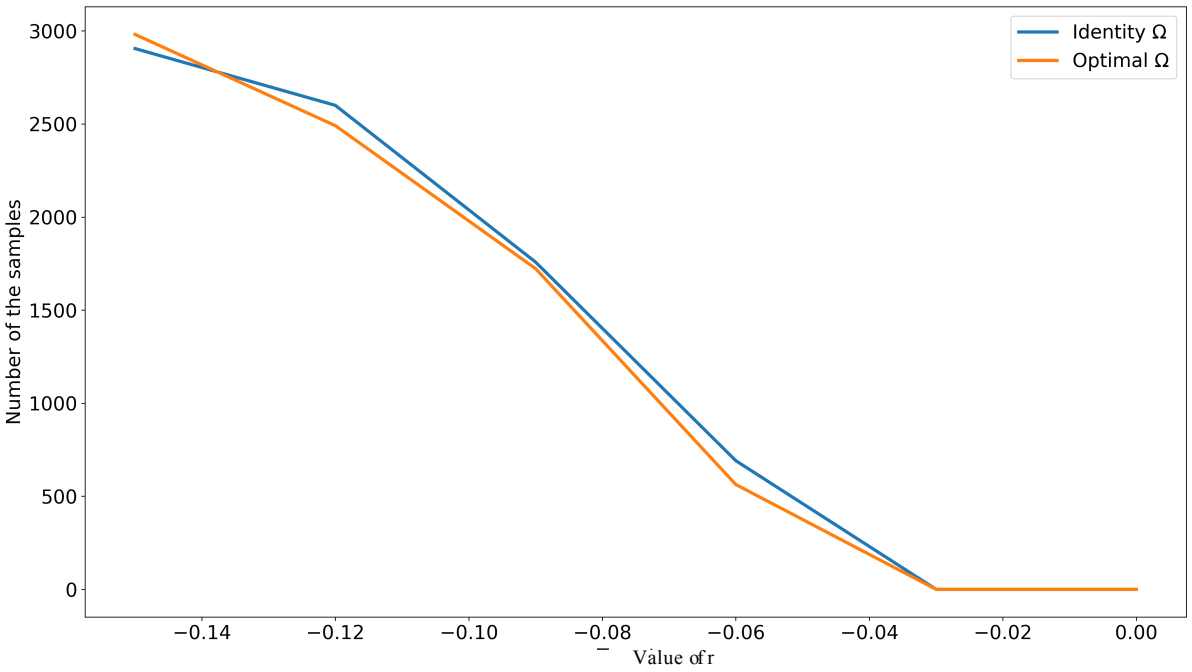

Figure 7: Number of samples that violate the r bound

5.2.2 Power of the test for constant ρ

To explore the power of the test, we implement a similar bootstrap experiment by generating artificial data in which the "true" coefficient ρ is set at a given level r. We use the formulæ in (6) to generate the probabilities. We only report the results with what we called the nonparamatric bootstrap here; the results with the parametric bootstrap are very similar. With a constant r, the lower bound (7) on the value of r becomes

$$-\min_x \min\left(\hat{a}(x), \frac{1}{\hat{a}(x)}\right) \leq r.$$

This turns out to be a rather demanding criterion. Figure 7 shows that as r becomes more negative, the number of samples for which $-\min\left(\hat{a}(x), \frac{1}{\hat{a}(x)}\right)$ is larger than r quickly becomes large.

This in itself is strong evidence that much of the observed data is incompatible with even a fairly small negative correlation coefficients. We will discard these samples when bootstrapping.

We again consider both the optimal matrix $\Omega = \Omega^*$ and the identity matrix $\Omega = I_4$. As the nonparametric boostrap procedure is time consuming, we only consider only six values for r: $r = 0$ (the boundary of the null hypothesis) and five values under the alternative, $r = -0.03, -0.06, -0.09, -0.12, -0.15$; we connect the resulting data points on our graphs for

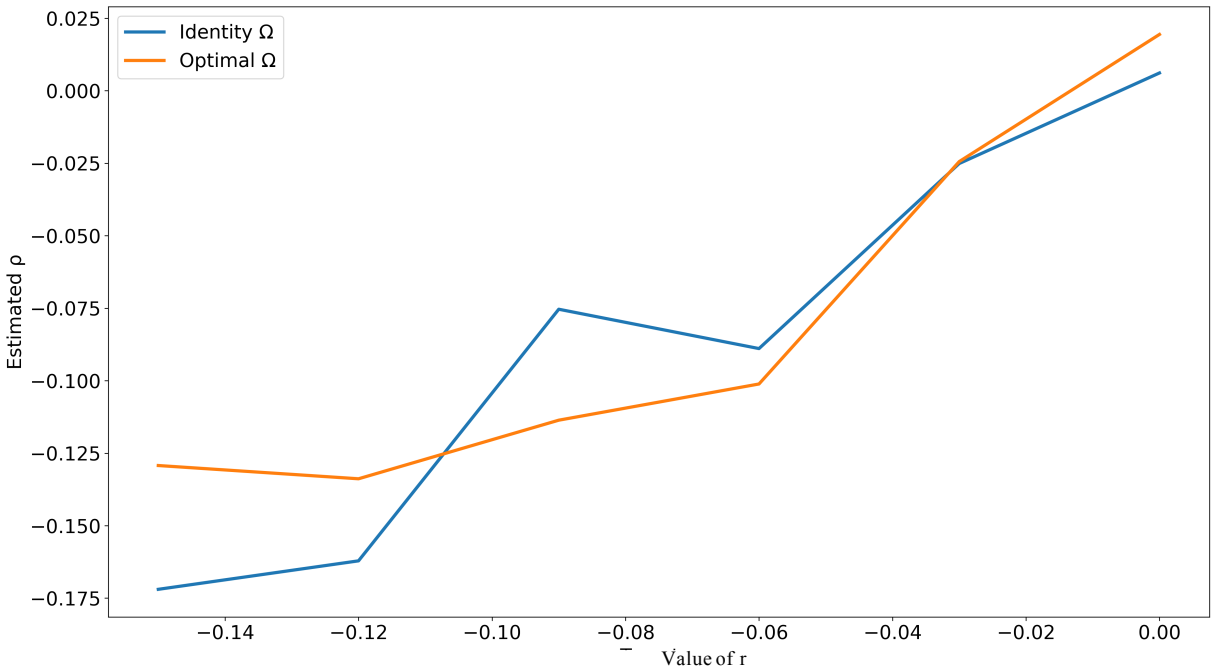

Figure 8: Estimated ρ as a function of r

visual convenience. Note that even when $r = -0.09$, we have to discard about one quarter of the sample; our results are probably more informative for $r = -0.03$ and $r = -0.06$ than for more negative values.

Figure 8 plots the estimated ρ as a function of r, along with a dashed line $\hat{\rho} = r$.Our procedure delivers estimates of ρ that are reasonably close to the true value r. Finally, Figure 9 plots the power of the Student test with the standard critical value of -1.64. Even with a small negative correlation $r = -0.06$, our test rejects the null hypothesis of positive correlation with close to probability one. The results are quite similar with the optimal matrix $\Omega = \Omega^*$ and with the identity matrix $\Omega = I_4$. In this case of constant correlation, the power of the test is very satisfactory.

Concluding Remarks

With our small sample of $6,330$ observations, deep learning cannot go very deep: a small neural network with two hidden layers and a small dropout rate world best. While it fits the choice of coverage well, it does not improve on parametric procedures for the claim occurrence. Double debiasing gives is a consistent and asymptotically normal estimate of a parameterized correlation function; given the small sample size, its pointwise standard error is relatively large

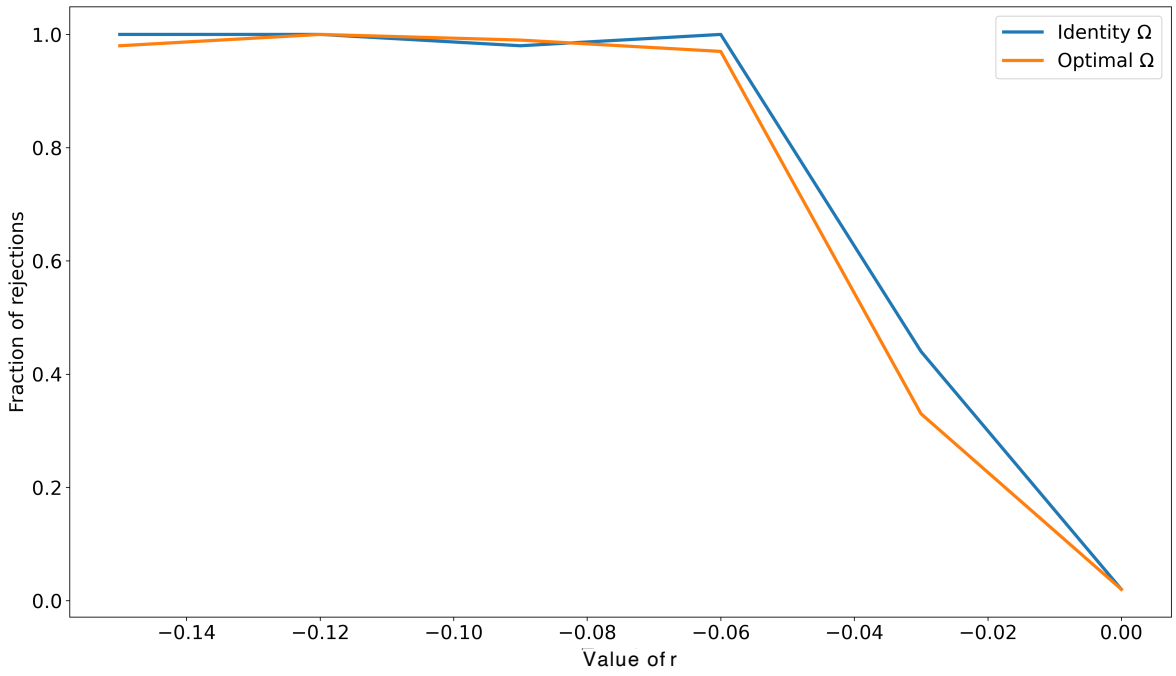

Figure 9: Power of the one-sided Student test as a function of r

and the intersection test for the positive correlation property gives mixed results.

Specializing the parameterized correlation function to be a constant, we obtain stronger results: the positive correlation property is not rejected, and the power of the test is high enough that it rules out any negative correlation, except for levels very close to zero. Still, the fairly large standard deviation in the estimated correlation function $\hat{\rho}(x)$ suggests that constant ρ is probably not the right model.

In further work, we plan to apply our three-step method to larger samples of insurees, and to explore the value of reducing the number of covariates to those that seem to have the largest effect.

References

Abadi, M., et al. (2016): "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," *CoRR*, abs/1603.04467.

Chernozhukov, V., et al. (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21(1), C1–C68.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.

Chiappori, P., B. Jullien, B. Salanié, and F. Salanié (2006): "Asymmetric information in insurance: general testable implications," *The RAND Journal of Economics*, 37(4), 783–798.

Chiappori, P., and B. Salanié (2000): "Testing for Asymmetric Information in Insurance Markets," *Journal of Political Economy*, 108(1), 56–78.

Chiappori, P., and B. Salanié (2013): "Asymmetric Information in Insurance Markets: Predictions and Tests," in *Handbook of Insurance*, ed. by G. Dionne, pp. 397–422. Springer New York.

Chollet, F. (2021): *Deep Learning with Python, Second Edition*. Manning.

Dionne, G., C. Gouriéroux, and C. Vanasse (2001): "Testing for evidence of adverse selection in the automobile insurance market: A comment," *Journal of Political Economy*, 109, 444–453.

——— (2006): "Informational content of household decisions with applications to insurance under asymmetric information," in *Competitive Failures in Insurance Markets: Theory and Policy Implications*, ed. by P. Chiappori, and C. Gollier, CESifo Seminar Series, pp. 159–184. MIT Press, Cambridge, MA:.

Kim, H., D. Kim, S. Im, and J. W. Hardin (2009): "Evidence of Asymmetrie Information in the Automobile Insurance Market: Dichotomous Versus Multinomial Measurement of Insurance Coverage," *Journal of Risk and Insurance*, 76, 343–366.

Kingma, D. P., and J. Ba (2014): "Adam: A Method for Stochastic Optimization," .

Spindler, M. (2014): "Econometric Methods for Testing for Asymmetric Information: A Comparison of Parametric and Nonparametric Methods with an Application to Hospital Daily Benefits," *The Geneva Risk and Insurance Review*, 39, 254–266.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014): "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, 15, 1929–1958.

Su, L., and M. Spindler (2013): "Nonparametric Testing for Asymmetric Information," *Journal of Business and Economic Statistics*, 31, 208–225.

Appendix

Let us derive (2). We drop the x from the notation, and we denote $f \equiv f(x, \beta_0)$ and $f_\beta = \frac{\partial f}{\partial \beta}(x, \beta_0)$.

With $\Omega = I_4$, the matrix μ' is simply $S^* - \Gamma \left(\Gamma'\Gamma\right)^{-1} \Gamma' S^*$. Simple calculations show that

$$\Gamma(x) = -E(w|X = x) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ f(x,\beta_0)R_{01}(x) & f(x,\beta_0)R_{10}(x) & f(x,\beta_0)R_{11}(x) \end{pmatrix} \tag{8}$$

where $R_{jk}(x)$ is the derivative of $R(x)$ with respect to $p_{jk}(x)$:

$$\begin{aligned} R_{01} &= \frac{\partial R}{\partial p_{01}} = \frac{p(1-p)(1-2q)}{2R} \\ R_{10} &= \frac{\partial R}{\partial p_{10}} = \frac{q(1-q)(1-2p)}{2R} \\ R_{11} &= \frac{\partial R}{\partial p_{11}} = R_{01} + R_{10} = \frac{(p+q-2pq)(1-p-q)}{2R}. \end{aligned}$$

We use (8) to compute

$$\Gamma'\Gamma = \left(E(w|x)\right)^2 \begin{pmatrix} 1 + f^2 R_{01}^2 & f^2 R_{01} R_{10} & f^2 R_{01} R_{11} \\ f^2 R_{01} R_{10} & 1 + f^2 R_{10}^2 & f^2 R_{10} R_{11} \\ f^2 R_{01} R_{11} & f^2 R_{10} R_{11} & 1 + f^2 R_{11}^2 \end{pmatrix}$$

and

$$\left(\Gamma'\Gamma\right)^{-1} = \frac{1}{D} \begin{pmatrix} 1 + f^2(R_{10}^2 + R_{11}^2) & -f^2 R_{01} R_{10} & -f^2 R_{01} R_{11} \\ -f^2 R_{01} R_{10} & 1 + f^2(R_{01}^2 + R_{11}^2) & -f^2 R_{10} R_{11} \\ -f^2 R_{01} R_{11} & -f^2 R_{10} R_{11} & 1 + f^2(R_{01}^2 + R_{10}^2) \end{pmatrix}$$

with

$$D = \left(1 + f^2(R_{01}^2 + R_{10}^2 + R_{11}^2)\right) \left(E(w|x)\right)^2.$$

Moreover, we know that the first three rows of S^* are zero and its fourth row is $-RE(w|x)f'_\beta$. This gives the $(3, q)$ matrix

$$\Gamma' S^* = f \ \left(E(w|x)\right)^2 \ R \times \begin{pmatrix} f'_\beta R_{01} \\ f'_\beta R_{10} \\ f'_\beta R_{11} \end{pmatrix}$$

It is easy to see that S^* belongs to the space spanned by the columns of Γ, so that $(\Gamma'\Gamma)^{-1}\Gamma' S^*$ is collinear with S^*. In fact,

$$(\Gamma'\Gamma)^{-1}\Gamma' S^* = C \times \begin{pmatrix} f'_\beta R_{01} \\ f'_\beta R_{10} \\ f'_\beta R_{11} \end{pmatrix}$$

with

$$C = \frac{fR}{1 + f^2(R_{01}^2 + R_{10}^2 + R_{11}^2)}.$$

Finally, we get

$$\mu' = S^* - C\Gamma \begin{pmatrix} f'_\beta R_{01} \\ f'_\beta R_{10} \\ f'_\beta R_{11} \end{pmatrix},$$

that is

$$\mu' = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -RE(w|x)f'_\beta \end{pmatrix} + CE(w|x) \begin{pmatrix} f'_\beta R_{01} \\ f'_\beta R_{10} \\ f'_\beta R_{11} \\ f f'_\beta (R_{01}^2 + R_{10}^2 + R_{11}^2) \end{pmatrix}$$

Rearranging terms,

$$\mu' = E(w|x) \begin{pmatrix} CR_{01} \\ CR_{10} \\ CR_{11} \\ (Cf(R_{01}^2 + R_{10}^2 + R_{11}^2) - R) \end{pmatrix} f'_\beta.$$

Since

$$Cf(R_{01}^2 + R_{10}^2 + R_{11}^2) - R = -\frac{R}{1 + f^2(R_{01}^2 + R_{10}^2 + R_{11}^2)} = -\frac{C}{f},$$

we can write

$$\mu' = CE(w|x) \begin{pmatrix} R_{01} \\ R_{10} \\ R_{11} \\ -1/f \end{pmatrix} f'_\beta$$

and we get

$$\begin{aligned}\psi &= CE(w|x)\left(R_{01}m_{01} + R_{10}m_{10} + R_{11}m_{11} - \frac{m_\beta}{f}\right)\ f'_\beta \\ &= \frac{RE(w|x)}{1 + f^2(R_{01}^2 + R_{10}^2 + R_{11}^2)}\left(f(R_{01}m_{01} + R_{10}m_{10} + R_{11}m_{11}) - m_\beta\right)\ f'_\beta.\end{aligned}$$

Finally, using $T_{jk} \equiv R \times R_{jk}$ and reintroducing the arguments:

$$\psi(x,\beta) = \left(f(x,\beta_0)\ (T_{01}(x)m_{01}(x) + T_{10}(x)m_{10}(x) + T_{11}m_{11}(x)) - R(x)m_\beta(x,\beta)\right)\ V(x)$$

with

$$V(x) \equiv \frac{R(x)^2 E(w|X=x)}{R(x)^2 + f^2(x,\beta_0)(T_{01}^2(x) + T_{10}^2(x) + T_{11}^2(x))}\ f'_\beta(x,\beta_0).$$

Note that since μ is defined at the true values, β_0 appears in this formula.