Testing for a functional form of mean regression in a fully parametric environment

Stanislav Anatolyev[*]

CERGE-EI, Czech Republic and New Economic School, Russia

July 2017

Abstract

We develop a test for a restricted functional form of a mean regression when a complex distributional model for all variables is estimated. The test statistic is an average squared deviation from the estimated hypothesized form of the form implied by the estimated parametric model, and is asymptotically distributed as a mixture of chi-squared distributions. The test is easy to implement using numerical derivatives, and it performs well in samples of typical size. We illustrate the test using data on labor market characteristics of U.S. young men.

Keywords: mean regression, functional form, specification test, wage regression

JEL codes: C12, C21

[*] Address: Stanislav Anatolyev, CERGE-EI, Politických vězňů 7, 11121 Prague 1, Czech Republic; e-mail stanislav.anatolyev@cerge-ei.cz. I am grateful to the Co-Editor and two anonymous referees for useful suggestions that significantly improved the presentation. I also thank Nikolay Kudrin for excellent research assistance.

1 Introduction

In a fully parametric setup when the distributional specification is available, one may be interested in whether the mean regression takes a particular restricted functional form. While the unrestricted regression may be inferred from the specified distribution and estimated from the data, it is likely to allow a rich variety of shapes.[1] In such a case, it is often interesting whether the shape of the mean regression reduces to some functional form implied by economic theory, tradition in the literature, or visual inspection; at the same time, it may be problematic to test directly for parametric restrictions embedded in the hypothesized shape. As an example, Figure 2 from our illustrative application based on a complex mixed continuous/discrete distribution presents two regressions of a wage variable on a variable representing education and on a variable representing age, derived from the estimator of the fully parametric (i.e. joint distributional) model. One of these regressions looks quite like linear to a naked eye, and is widely assumed to be linear in the literature, but is it truly linear? The other may seem to be cubic or quartic, but is it truly such? Do the observable deviations from a low-order polynomial owe merely to the sampling error, or do they evidence against these simple forms of the conditional mean?

In this paper, we develop a test for a parametric functional form of a mean regression when the full parametric model for all variables is estimated.[2] A natural test statistic is the average squared deviation of the regression function implied by the estimated parametric model from the hypothesized functional form. We derive the asymptotic distribution of the test statistics, which turns out to be a weighed sum of chi-squared distributions with one degree of freedom. Even though the test statistic is non-pivotizable (except possibly in some special cases when the distribution collapses to a single scaled chi-squared distribution with one degree of freedom), the test is easy to implement by employing estimates of the weights by using numerical derivatives of the true and hypothesized regression functions and the score function. We demonstrate good size and power properties of the test in finite samples using two simple stylized models – one is based on bivariate normality, and the other on a mixed continuous/discrete marginals linked by a copula. Finally, we illustrate the test using Card's (1995) data on wage, education and age of a few thousand U.S. young men. Despite the regressions may look seemingly linear to a naked eye, the test decidedly rejects linearity of regressions of log-wage on education, log-wage on age, and log-wage on both education and age, as well as of low-order (quartic) polynomial analogs

[1] Unless the distributional specification is very simple, as, for example, joint normalily, in which case the mean regression is necessarily linear.

[2] Another appropriate context is semiparametric where a *conditional* distribution is specified and estimated in the first place. However, this case is less practical, because specifying a conditional distribution typically entails specifying the conditional mean as a part of the modeling strategy.

of these.

There exists a variety of tests for a parametric form of a mean regression against nonparametric alternatives; see, for instance, Härdle and Mammen (1990) and Horowitz and Spokoiny (2001). This is also a valid approach to testing for a regression parametric specification. However, when the whole framework is parametric, more natural is to utilize it and perform testing within the parametric distributional model. In addition, from the technical standpoint, the nonparametric tests usually involve kernel estimation of the mean regression and bootstrapping of the test statistic, so their implementation is more involved than that of the test proposed here.

The paper is structured as follows. In Section 2 the setup is described, the assumptions are laid out, and properties of auxiliary estimates are derived. In Section 3 the test statistic and its asymptotic properties are presented, and implementation of the test is described. Section 4 contains two illustrative examples, accompanied by simulation evidence. In Section 5, we illustrate how the test works using labor market data. Finally, Section 6 concludes. All proofs and tedious derivations are relegated to the Appendix. Notes on notation: $\|\cdot\|$ denotes the L_2 norm of a matrix, by $\dim(\cdot)$ we denote dimensionality of a vector, by $\mathrm{rk}(\cdot)$ – its rank.

2 Setup and estimation

Suppose there is a parametric density[3] model $f(u,v|\theta)$, $\theta\in\Theta$ for a pair of possibly multidimensional random variables (u,v), u being scalar and v being possibly multidimensional, and let θ^0 be the true value of parameter θ. The implied mean regression for u given v is the value at θ^0 of the conditional expectation function

$$E\left(u|v,\theta\right)=\int_{-\infty}^{+\infty}uf(u|v,\theta)du,$$

where

$$f(u|v,\theta)=\frac{f(u,v|\theta)}{g(v|\theta)}$$

is the conditional density of u given v, and

$$g(v|\theta)=\int_{-\infty}^{+\infty}f(u,v|\theta)du$$

[3] We call simply by 'density' what may in fact be a mass in the case discrete variables are considered, or a density/mass in the case mixed continuous/discrete variables are. The integrals considered from now on are then redefined accordingly.

is the marginal density of v. The estimated implied regression is then

$$E\left(u|v,\hat{\theta}\right)=\frac{1}{g(v|\hat{\theta})}\int_{-\infty}^{+\infty}uf(u,v|\hat{\theta})du, \tag{1}$$

where $\hat{\theta}$ is the maximum likelihood estimator of θ^0:

$$\hat{\theta}=\arg\min_{\theta\in\Theta}\sum_{i=1}^{n}\ln f(u_i,v_i|\theta).$$

We would like to compare the implied regression (1) to the parametric functional form $\psi\left(v,\beta^0\right)$, where

$$\beta^0=\arg\min_{\beta\in\mathrm{B}}E\left[\left(u-\psi\left(v,\beta\right)\right)^2\right].$$

The estimator $\hat{\beta}$ of the (pseudo)true value of the parameter β^0 is based on least squares[4]:

$$\hat{\beta}=\arg\min_{\beta\in\mathrm{B}}\sum_{i=1}^{n}\left(u_i-\psi\left(v_i,\beta\right)\right)^2.$$

Denote $\vartheta=\left(\theta',\beta'\right)'$ and $\vartheta^0=\left(\theta^{0\prime},\beta^{0\prime}\right)'$. Because the test to be developed will need to use information on asymptotic correlatedness between $\hat{\theta}$ and $\hat{\beta}$, we frame the two estimation problems inside one joint optimization problem[5]

$$\hat{\vartheta}\equiv\left(\hat{\theta}',\hat{\beta}'\right)'=\arg\max_{\theta\in\Theta,\beta\in\mathrm{B}}\frac{1}{n}\sum_{i=1}^{n}\left\{Q\left(u_i,v_i|\vartheta\right)\equiv\ln f\left(u_i,v_i|\theta\right)-\frac{1}{2}\left(u_i-\psi\left(v_i,\beta\right)\right)^2\right\},$$

and the asymptotic variance estimate $\hat{V}_{\vartheta}$ for $\hat{\vartheta}$ can be obtained numerically from this optimization problem. The factor $\frac{1}{2}$ is added for convenience of computing the derivatives; its presence (or presence of any other positive factor) does not affect the estimator or its properties.

Let us have a closer look at the structure of $\hat{V}_{\vartheta}$. Because $\hat{\vartheta}$ is an extremum estimator, it has a sandwich form $H^{-1}\Omega H^{-1}$, where H is a Hessian matrix, and Ω is a variance matrix of first derivatives. Because of an additive structure of $Q\left(u,v|\vartheta\right)$ in θ and β, H has a block-diagonal form

$$H=\left[\begin{array}{cc}H_f & 0\\ 0 & -M_{\psi\psi}\end{array}\right],$$

[4] The use of other consistent criteria is also possible. The test can be modified in a straightforward way.

[5] Even in the most likely case when $\psi\left(v,\beta\right)$ is linear in β and the solution for $\hat{\beta}$ is known in a closed form, the ML estimator $\hat{\theta}_f$ is likely not, so one still has to solve a nonlinear optimization problem. The closed form of $\hat{\beta}$ can be conveniently used as an starting (and final) point for β during optimization.

where

$$H_f = E\left[\frac{\partial^2 \ln f(u,v|\theta^0)}{\partial\theta\partial\theta'}\right] = E\left[\frac{\partial^2 Q\left(u,v|\theta^0,\beta^0\right)}{\partial\theta\partial\theta'}\right]$$

and

$$M_{\psi\psi} = E\left[\frac{\partial\psi\left(v,\beta^0\right)}{\partial\beta}\frac{\partial\psi\left(v,\beta^0\right)}{\partial\beta'}\right] = -E\left[\frac{\partial^2 Q\left(u,v|\theta^0,\beta^0\right)}{\partial\beta\partial\beta'}\right].$$

Next,

$$\Omega = \left[\begin{array}{cc} -H_f & M_{uf\psi} \\ M'_{uf\psi} & M_{u^2\psi\psi} \end{array}\right],$$

where

$$\begin{aligned} M_{uf\psi} &= E\left[\left(u-\psi\left(v,\beta^0\right)\right)\frac{\partial \ln f(u,v|\theta^0)}{\partial\theta}\frac{\partial\psi\left(v,\beta^0\right)}{\partial\beta'}\right], \\ M_{u^2\psi\psi} &= E\left[\left(u-\psi\left(v,\beta^0\right)\right)^2\frac{\partial\psi\left(v,\beta^0\right)}{\partial\beta}\frac{\partial\psi\left(v,\beta^0\right)}{\partial\beta'}\right] \end{aligned}$$

and the northwest corner is occupied by H_f because of information matrix equality (recall that $\hat{\theta}$ is an ML estimate). Putting the pieces together,

$$\hat{V}_\vartheta = H^{-1}\Omega H^{-1} = \left[\begin{array}{cc} -H_f^{-1} & -H_f^{-1}M_{uf\psi}M_{\psi\psi}^{-1} \\ -M_{\psi\psi}^{-1}M'_{uf\psi}H_f^{-1} & M_{\psi\psi}^{-1}M_{u^2\psi\psi}M_{\psi\psi}^{-1} \end{array}\right].$$

Note that while H is necessarily of full rank, the matrix Ω may well be singular. In one of examples in Section 4, $\mathrm{rk}(\Omega) = 4$ while $\dim(\vartheta) = 5$. The matrices H_f, $M_{\psi\psi}$, $M_{u\psi\psi}$ and $M_{u^2\psi\psi}$ can be easily estimated by numerical derivatives and the parameter estimate $\hat{\vartheta}$ already obtained.

We make a number of assumptions that guarantee existence of the above moments and ensure joint consistency and asymptotic normality of ML and LS estimates $\hat{\theta}$ and $\hat{\beta}$.

Assumption 1 *The following about data generation holds:*
(a) the data $\{(u_i,v_i)\}_{i=1}^n$ is a random sample from a population with probability density $f(u,v|\theta^0)$ and finite $E\left[u^2\right]$;
(b) the parameter set Θ is a compact subset of $\mathbb{R}^{\dim(\theta)}$, and θ^0 is in the interior of Θ;
(c) for any $\theta\in\Theta$ such that $\theta\neq\theta^0$, it holds that $f(u,v|\theta)\neq f(u,v|\theta^0)$;
(d) $f(u,v|\theta)$ is continuous in θ on Θ and twice continuously differentiable in θ in a neighborhood $\mathfrak{N}_\theta$ of Θ;
(e) the following moments are finite: $E\left[\sup_{\theta\in\Theta}|f(u,v|\theta)|\right]$, $E\left[\sup_{\theta\in\mathfrak{N}_\theta}\left\|\partial^2\ln f(u,v|\theta)/\partial\theta\partial\theta'\right\|\right]$, and the following functions are integrable: $\sup_{\theta\in\mathfrak{N}_\theta}\|\partial f(u,v|\theta)/\partial\theta\|$, $\sup_{\theta\in\mathfrak{N}_\theta}\left\|\partial^2 f(u,v|\theta)/\partial\theta\partial\theta'\right\|$;

(f) the matrix H_f is non-singular.

Assumption 2 *The following about the hypothesized regression function holds:*
(a) the parameter set B *is a compact subset of* $\mathbb{R}^{\dim(\beta)}$*, and* β^0 *is in the interior of* B;
(b) for any $\beta \in \mathrm{B}$ *such that* $\beta \neq \beta^0$*, it holds that* $\psi(v,\beta) \neq \psi(v,\beta^0)$*;*
(c) $\psi(v,\beta)$ *is continuously differentiable in* β *on* B *and twice continuously differentiable in* β *in a neighborhood* $\mathfrak{N}_\beta$ *of* B;
(d) the following moments are finite: $E\left[\sup_{\beta\in\mathrm{B}}\psi(v,\beta)^2\right]$, $E\left[\sup_{\beta\in\mathrm{B}}\|\partial\psi(v,\beta)/\partial\beta\|^2\right]$,
$E\left[\sup_{\beta\in\mathfrak{N}_\beta}\|\partial\psi(v,\beta)/\partial\beta\cdot\partial\psi(v,\beta)/\partial\beta'\|\right]$, $E\left[\sup_{\beta\in\mathfrak{N}_\beta}\left\|\partial^2\psi(v,\beta)/\partial\beta\partial\beta'\right\|^2\right]$;
(e) the matrix $M_{\psi\psi}$ is non-singular.

Lemma 1: Suppose assumptions 1–2 hold. Then $\hat{\vartheta}\xrightarrow{p}\vartheta^0$.

For future use, define

$$\Delta = E\left[\left(\frac{\partial E(u|v,\theta^0)}{\partial\vartheta}-\frac{\partial\psi(v,\beta^0)}{\partial\vartheta}\right)\left(\frac{\partial E(u|v,\theta^0)}{\partial\vartheta}-\frac{\partial\psi(v,\beta^0)}{\partial\vartheta}\right)'\right].$$

A natural estimator of Δ is

$$\hat{\Delta}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial E(u|v_i,\hat{\theta})}{\partial\vartheta}-\frac{\partial\psi(v_i,\hat{\beta})}{\partial\vartheta}\right)\left(\frac{\partial E(u|v_i,\hat{\theta})}{\partial\vartheta}-\frac{\partial\psi(v_i,\hat{\beta})}{\partial\vartheta}\right)'.$$

For simplicity, we assume that this evaluation occurs without computational error.[6] We make additional technical assumptions that ensure finiteness of Δ and consistency of $\hat{\Delta}$.

Assumption 3 *The following moments exist and are finite:* $E\left[\sup_{\theta\in\Theta}\|\partial\ln f(u|v,\theta)/\partial\theta\|^2\right]$, $E\left[\sup_{\theta\in\mathfrak{N}_\theta}\|E\left[u\cdot\partial\ln f(u|v,\theta)/\partial\theta|v\right]\|^2\right]$.

Lemma 2: Suppose assumptions 1–3 hold. Then $\hat{\Delta}\xrightarrow{p}\Delta$.

[6] There are several sources of computational errors: software's round-off error, error in evaluation of integrals on a finite domain, error from neglecting tails of functions being integrated, and error in evaluation of derivatives. See Judd (1998) for information about orders of some of these approximation errors. For example, two-sided differences in evaluation of first derivatives lead to errors of order $O\left(h^2+h^{-1}\epsilon\right)$, where h is a step size and ϵ is an error in computation of the function being integrated (which may exceed the round-off error) (Judd, 1998, section 7.7); numerical integration on a bounded interval using the Gaussian–Chebychev quadrature causes errors of order $O\left(\left(2^{2m}(2m)!\right)^{-1}\right)$, where m is a number of quadrature nodes (Judd, 1998, section 7.2). We assume that the total computational error is sufficiently controllable so that it does not affect the test statistic to the precision used to compute it.

Note that because of convenient partitioning of ϑ into θ and β and dependence $E\left(u|v,\theta\right)$ only on θ and of $\psi\left(v,\beta\right)$ only on β, differentiation inside Δ also separates out, and one can rewrite

$$\frac{\partial E\left(u|v,\theta^0\right)}{\partial\vartheta}-\frac{\partial\psi\left(v,\beta^0\right)}{\partial\vartheta}=\left[\begin{array}{c}\dfrac{\partial E\left(u|v,\theta^0\right)}{\partial\theta}\\ -\dfrac{\partial\psi\left(v,\beta^0\right)}{\partial\beta}\end{array}\right]. \qquad (2)$$

While the bottom entry can be computed analytically, for the top entry one can use the machinery of numerical derivatives in a straightforward way.

3 Test and asymptotics

Suppose that $\psi\left(v,\beta\right)$ is specified so that it may be equal, almost surely, to $E\left(u|v,\theta\right)$ derived from a fully parametric model $f(u,v|\theta)$ for some combination of θ and β. The null hypothesis to be tested is

$$H_0: E\left(u|v,\theta^0\right)=\psi\left(v,\beta^0\right) \quad \text{a.s.}$$

Denote

$$\Lambda=H^{-1}\Omega H^{-1}\Delta.$$

The test of the null H_0 is based on the comparison at data points of regression values implied by the full parametric model and by the hypothesized regression function. The sample squared deviations statistic is

$$\hat{D}=\frac{1}{n}\sum_{i=1}^{n}\left(E(u|v_i,\hat{\theta})-\psi(v_i,\hat{\beta})\right)^2,$$

which is the sample analog to

$$D=E\left[\left(E\left(u|v,\theta^0\right)-\psi\left(v,\beta^0\right)\right)^2\right],$$

which is zero under H_0 and nonzero otherwise.

The following theorem provides the asymptotic distribution of $\hat{D}$ under the null, which turns out to be a weighted sum of chi-squared distributions.

Theorem 1: Suppose assumptions 1–2 hold and $\mathrm{rk}(\Lambda)\neq 0$. Then, under H_0,

$$n\hat{D}\overset{d}{\to}\mathcal{D},$$

where

$$\mathcal{D} \overset{d}{=} \sum_{j=1}^{\dim(\vartheta)} \lambda_j \zeta_j^2,$$

where $\{\lambda_j\}_{j=1}^{\dim(\vartheta)}$ are eigenvalues of Λ, and $\{\zeta_j^2\}_{j=1}^{\dim(\vartheta)} \sim IID\,\chi^2_{(1)}$.

To implement the test, one computes $\hat{D}$, constructs consistent estimates $\hat{H}$ and $\hat{\Delta}$ of H and Δ and finds eigenvalues $\{\hat{\lambda}_j\}_{j=1}^{\dim(\vartheta)}$ of $\hat{\Lambda} = \hat{H}^{-1}\hat{\Omega}\hat{H}^{-1}\hat{\Delta}$. Then one simulates the distribution of

$$\hat{\mathcal{D}} \overset{d}{=} \sum_{j=1}^{\dim(\vartheta)} \hat{\lambda}_j \zeta_j^2,$$

and reads off its relevant right quantile to use as critical values for $n\hat{D}$.[7]

Note that Λ may well be of reduced rank, and it may be of rank even lower than $\dim(\theta)$ and/or $\dim(\beta)$. In one of examples in Section 4, $\text{rk}(\Lambda) = 1$ while $\dim(\theta) = 3$ and $\dim(\beta) = 2$; in the second example, $\text{rk}(\Lambda) = 3$ while $\dim(\theta) = 5$ and $\dim(\beta) = 2$. This happens because typically there is a great deal of collinearity between the derivative of the true regression, the derivative of the hypothesized regression, and the score, at least under the null. This phenomenon does not, however, pose any difficulties in implementation in case $\text{rk}(\Lambda)$ is *a priori* unknown (which is typically the case) as the other $\dim(\vartheta) - \text{rk}(\Lambda)$ eigenvalues of Λ are zeros.

Consider now the situation when the null hypothesis does not hold. More precisely, the null does not hold if for no parameter value the functional form $\psi(v, \beta)$ coincides with the true regression almost surely. Note that in this case β_0 is interpreted as a pseudotrue value of β as the true value does not exist. The following theorem says that under any alternative, the test statistic diverges.

Theorem 2: Suppose assumptions 1–2 hold, and $\Pr\{E\left(u|v, \theta^0\right) \neq \psi(v, \beta)\} > 0$ for any $\beta \in \mathrm{B}$. Then

$$n\hat{D} \xrightarrow{p} +\infty.$$

Theorem 2 implies that the test is consistent against any deviations from the true specification, i.e. when the regression function does not equal, on a set of positive measure however small it is, to the hypothesized specification $\psi(v, \beta)$ for any $\beta \in \mathrm{B}$. The power of the test is

[7] As a practical matter, simulation of the null distribution can be implemented very easily given the collection of eigenvalues. For example, in GAUSS, the vector of simulated values can be computed using the statement `sumc(lambda.*rndn(d,S)^2);`. Here, the vector `lambda` contains the eigenvalues, `d` is the dimension of ϑ, and `S` is the number of simulations.

expected to be greater the larger is this set on which the two functions (evaluated at the true and pseudotrue parameter values, respectively) deviate from each other, and/or the larger are those deviations.

4 Illustrations and simulations

In this section we elaborate on two examples of data generating processes to illustrate the construction of the test and verify its finite sample performance.

The aim of our first experiment is to analyze the size of the test in a simplest setup, and, even more importantly, to see whether the use of numerical derivatives delivers good enough precision in controlling the size of the test. Here all variables are continuous, the regression function has a known form, and the matrices related to first and second derivatives are computable in a closed form. Namely, we use a jointly normal model for the two variables

$$f(u,v|\theta)=\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{(u-\mu_u)^2-2\rho\,(u-\mu_u)\,(v-\mu_v)+(v-\mu_v)^2}{2\,(1-\rho^2)}\right),$$

where $\theta=(\mu_u,\mu_v,\rho)'$. Due to joint normality, the regression function is linear: $E\,(u|v,\theta)=\mu_u+\rho\,(v-\mu_v)$. We use this fact to verify performance of the test in finite samples in terms of size properties, setting $\psi\,(v,\beta)=a+bv$, where $\beta=(a,b)'$. Notice that there is *a priori* no doubt that the tested regression functional form is true. The total dimensionality of the parameter vector is $\dim\,(\vartheta)=5$.

In Appendix A2 we derive that

$$\Lambda=2\rho^2\frac{1-\rho^2}{1+\rho^2}\begin{bmatrix}0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&\mu_v&0&-\mu_v\\0&0&-1&0&1\end{bmatrix}. \tag{3}$$

We rule out the cases $\rho=\pm1$ as these values sit on the boundary of the parameter set $[-1,1]$ for ρ. In the formulation of Theorem 1, we also rule out the case $\rho=0$ which leads to Λ being a zero matrix with $\mathrm{rk}(\Lambda)=0$. The test will not work properly when $\rho=0$.

Provided that $\rho\neq0$ and $\rho\neq\pm1$, the rank of Λ is unity no matter what the parameter values are, and only non-zero eigenvalue is $\lambda_\rho=2\rho^2\frac{1-\rho^2}{1+\rho^2}$. Note that even though $\dim\,(\vartheta)=5$, we have $\mathrm{rk}(H)=5$, $\mathrm{rk}(\Omega)=4$, $\mathrm{rk}(\Delta)=2$, yet $\mathrm{rk}(\Lambda)=1$. Because $\mathrm{rk}(\Lambda)=1$, the limiting distribution in fact simplifies to λ_ρ times a $\chi^2_{(1)}$ distribution. Thus, the critical values can be computed simply as $\hat{\lambda}_\rho$ times an appropriate quantile of the tabulated $\chi^2_{(1)}$ distribution, where $\hat{\lambda}_\rho$ is λ_ρ with the ML estimate $\hat{\rho}$ plugged in place of ρ.

This result will be used as an 'analytic' benchmark when one uses analytical derivatives. To that end, we set the limiting distribution as described in the previous paragraph. The other, 'numerical' value for Λ is obtained as $\hat{H}^{-1}\hat{\Omega}\hat{H}^{-1}\hat{\Delta}$, where $\hat{H}$, $\hat{\Omega}$ and $\hat{\Delta}$ are estimates of H, Ω and Δ using numerical derivatives.[8]

The pairs $\{(u_i, v_i)\}_{i=1}^n$ are drawn from the bivariate normal distribution with means $\mu_u^0 = \mu_v^0 = 1$, unit variances and correlation $\rho^0 = 0.5$. The following simulation results are based on 2000 simulations; the rejection rates are expressed in percentages.

rejection rates					
analytical			numerical		
10%	5%	1%	10%	5%	1%
$n = 100$					
10.5	5.6	1.0	15.0	8.9	3.0
$n = 500$					
10.1	5.1	1.2	7.9	4.0	1.0
$n = 2000$					
10.6	5.0	0.7	8.5	4.8	1.2

The size control is excellent even for small samples when analytical derivatives are used. When one computes numerical derivatives instead, there are expectedly some size distortions, which go away quickly as the sample size grows. For samples of a few thousand, the size control is of no concern, at least for low-dimensional setups.

In our second experiment, we will analyze the size and power of the test in a more realistic setup. Here the data are mixed continuous and discrete. The continuous u has a logistic distribution, the discrete v is drawn from a three-point distribution, and the dependence is induced by the Farlie–Gumbel–Morgenstern (FGM) copula. These choices are due to availability of the joint PDF/PMF and CDF/CMF in a closed form, simplicity of the form of mean regression,

[8] The derivatives are computed using two-sided differences with the step of $h\theta$ componentwise, where $h = 10^{-5}$. The integrals involved in evaluation of expectations are computed via Gauss–Chebychev quadrature with $m = 100$ quadrature nodes on $[-8, 8]$. Such precision is more than sufficient not to worry about the error ϵ of computation of the function being integrated; see the previous footnote.

simplicity of tuning the parameters so that the regression function is linear or non-linear, and, finally, conceptual similarity to our illustrative empirical application.

The continuous marginal has the density

$$f_u(u|\mu,\gamma) = \frac{\exp\left(-\gamma^{-1}\left(u-\mu\right)\right)}{\gamma\left[1+\exp\left(-\gamma^{-1}\left(u-\mu\right)\right)\right]^2}$$

and cumulative distribution function

$$F_u(u|\mu,\gamma) = \frac{1}{1+\exp\left(-\gamma^{-1}\left(u-\mu\right)\right)}.$$

We set the true value of μ to be zero in order to obtain symmetry. The three-point distribution of the discrete marginal is $v \in \{-1,0,+1\}$ with marginal PMF $g(v)$ represented by the corresponding collection of probabilities $q \in \{q_{-1}, 1-q_{-1}-q_{+1}, q_{+1}\}$ with CMF $G(v) = q_{-1}1_{\{v\leq -1\}} + \left(1-q_{-1}-q_{+1}\right)1_{\{v\leq 0\}} + q_{+1}1_{\{v\leq 1\}}$. The dependence is induced by the FGM copula

$$C(w_1,w_2) = w_1w_2(1+\rho(1-w_1)(1-w_2)),$$

where $\rho \in [-1,+1]$ and $\rho > 0$ implies positive, although moderate at most, dependence.

Let $\theta = (\mu,\gamma,q_{-1},q_{+1},\rho)'$. It is shown in Appendix A2 that the joint density/mass is

$$\begin{aligned} f(u,v|\theta) &= f_u(u|\mu,\gamma)q_{-1}^C\left(F_u(u|\mu,\gamma)\right)^{1_{\{v=-1\}}} q_0^C\left(F_u(u|\mu,\gamma)\right)^{1_{\{v=0\}}} q_{+1}^C\left(F_u(u|\mu,\gamma)\right)^{1_{\{v=+1\}}} \\ &= \gamma^{-1}\omega\left(u\right)\left(1-\omega\left(u\right)\right)\varphi\left(u\right)_{-1}^{1_{\{v=-1\}}}\left(1-\varphi_{-1}\left(u\right)-\varphi_{+1}\left(u\right)\right)^{1_{\{v=0\}}}\varphi\left(u\right)_{+1}^{1_{\{v=+1\}}}, \end{aligned}$$

where

$$\omega\left(u\right) = \frac{\exp\left(-\gamma^{-1}\left(u-\mu\right)\right)}{1+\exp\left(-\gamma^{-1}\left(u-\mu\right)\right)}$$

and

$$\begin{aligned} \varphi_{-1}\left(u\right) &= q_{-1}+\rho\left(1-2\left(1-\omega\left(u\right)\right)\right)q_{-1}\left(1-q_{-1}\right) \\ \varphi_{+1}\left(u\right) &= q_{+1}-\rho\left(1-2\left(1-\omega\left(u\right)\right)\right)q_{+1}\left(1-q_{+1}\right) \end{aligned}$$

If, in addition to $\mu^0 = 0$, we set $q_{-1}^0 = q_{+1}^0$, then, due to a symmetry around the origin, the regression function will be linear: $E\left(u|v\right) = \lambda v$, where λ depends on q_{-1}^0. If we set $q_{-1}^0 \neq q_{+1}^0$, the symmetry ceases to take place, and the regression function is no longer linear. We again set $\psi\left(v,\beta\right) = a + bv$, where $\beta = (a,b)'$, and study size properties when $q_{-1}^0 = q_{+1}^0$ and power properties when $q_{-1}^0 \neq q_{+1}^0$. The total dimensionality of the parameter vector is $\dim\left(\vartheta\right) = 7$.

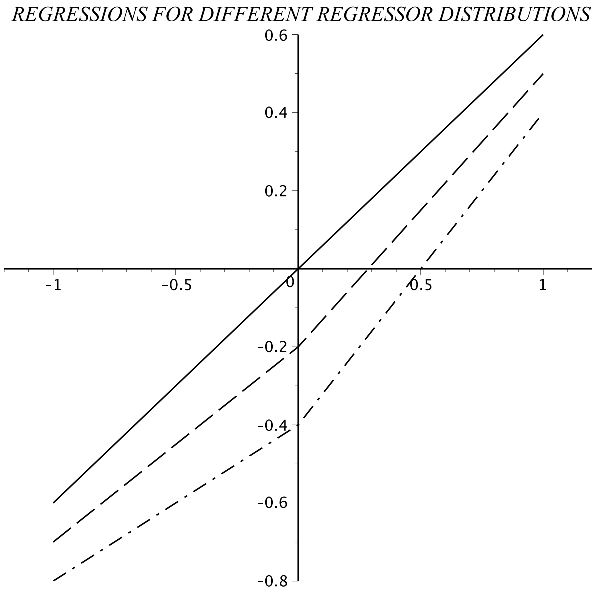

Figure 1: Regressions for the second experiment, linear and non-linear

The variables $\{u_i\}_{i=1}^n$ are drawn from the standard logistic distribution (i.e. with $\mu^0 = 0$ and $\gamma^0 = 1$). We set $\rho^0 = 1$ implying the correlation coefficient of $\frac{1}{3}$. Then, for a given i and given pair (q_{-1}^0, q_{+1}^0), we compute $\varphi_{-1}(u_i)$ and $\varphi_{+1}(u_i)$ and use these to generate the variables $\{v_i\}_{i=1}^n$ from the three-point distribution $\{-1, 0, +1\}$ with corresponding probabilities $\{\varphi_{-1}(u_i), 1 - \varphi_{-1}(u_i) - \varphi_{+1}(u_i), \varphi_{+1}(u_i)\}$. We set the pair (q_{-1}^0, q_{+1}^0) to three values, one of which implies a linear regression, while the other two imply non-linear ones (see Figure 1).

The following table contains simulation results for samples of small size $n = 100$, moderate size $n = 500$, and big size $n = 2000$. The results are based on 2000 simulations; the rejection rates are expressed in percentages.[9]

parameters		true regression at v			line on Figure 1	$n = 100$			$n = 500$			$n = 2000$		
q_{-1}^0	q_{+1}^0	-1	0	$+1$		10%	5%	1%	10%	5%	1%	10%	5%	1%
0.4	0.4	-0.6	0	0.6	solid	8.1	3.7	0.4	10.2	5.2	0.8	8.7	4.7	1.1
0.3	0.5	-0.7	-0.2	0.5	dashed	8.3	4.3	0.5	39.6	20.3	3.8	93.7	92.7	79.6
0.2	0.6	-0.8	-0.4	0.4	dash-dotted	10.9	4.8	0.7	93.4	85.3	34.1	98.6	97.8	94.9

Except for small samples, the size and power figures are favorable. The actual rejection rates

[9] See the previous computational footnote, except that the domain of integration is now $[-20, 20]$.

shown in the first line are quite close to nominal test sizes. The power figures are impressive, especially for large samples, and even though the true regression line does not deviate much from a linear form, the test detects it pretty often from a sample of a moderate size. With small samples, the null rejection rates fall short of nominal rates a bit, and the test has hard time detecting small deviations from the null. While a hundred observations are clearly not sufficient for the test to work properly, increasing the sample size severalfold straightens out the rejection rates and makes the properties of the test very attractive.

5 Illustrative application

In this section we illustrate the test using the labor market data from Card (1995). These data contain, in particular, wage, education and age of a sample of U.S. men of size $n = 3010$ taken in 1976. The main variable is logarithm of wages (`lwage76`), and regressors are education (`ed76`) and age (`age76`). We run bivariate and trivariate full parametric models for the pairs (`lwage76`,`ed76`), (`lwage76`,`age76`) and the triple (`lwage76`,`ed76`,`age76`), compute implied regressions of log wages on one or two regressors, and test them for linearity using the test developed in this paper.[10]

Because the regressand is a continuous variables while both regressors are discrete, we construct the joint distribution by using the copula machinery. The marginal density for the continuously distributed log wages is chosen to be the skew-normal distribution (Azzalini, 1985):

$$u = \mu + \sigma w,$$

where μ is a location parameter, σ is a scale parameter,

$$f_w\left(w|\gamma\right) = 2\phi\left(w\right)\Phi\left(\gamma w\right),$$

and[11] γ is a shape parameter that indexes the degree of skewness; the distribution reduces to the regular normal when $\gamma = 0$. In total, the skew-normal density $f_u\left(u|\theta_u\right)$ and its CDF $F_u(u|\theta_u)$ are characterized by three parameters in $\theta_u = \left(\mu, \sigma, \gamma\right)'$. Azzalini, Dal Cappello and Kotz (2003) argue that this distribution (among others) well approximates the real log income data. Below are the results of fitting the marginal skew-normal density to the variable `lwage76`.

[10] We do not make any attempt to interpret these regressions as any sort of causal relationships. A causal approach when `ed76` is involved requires an acknowledgement of its endogeneity and needs instrumental variables for consistent estimation; see Card (1995) and the rest of the returns to schooling literature.

[11] Note that μ and σ are *not* the mean and standard deviation of u.

marginal skewed normal: `lwage76`

parameter	μ	σ	γ
estimate	6.586	0.550	-1.100
standard error	0.034	0.021	0.162

The Kolmogorov–Smirnov statistic (the maximal difference between the empirical distribution function and estimated CDF) equals 0.0168, and, normalized by $\sqrt{n}$, equals 0.921, which is quite smaller than the critical value even at the 20% significance level (e.g., Massey, 1995).

The marginal distributions of the variables `ed76` and `age76` are categorical, with a number of categories being $k_1 = 18$ for the former and $k_2 = 11$ for the latter,[12] and with categorical probabilities $q_\ell = (q_j)_{j=1}^{k_\ell}$, $\ell = 1, 2$ subject to $\sum_{j=1}^{k_\ell} q_j = 1$. Let us denote the CMF of this distribution by $G_v(v|q) = \sum_{j=1}^{\lfloor v \rfloor} q_j$. The estimates are shown in the following tables.

marginal categorical: `ed76`

parameter	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}	q_{11}	q_{12}	q_{13}	q_{14}	q_{15}	q_{16}	q_{17}	q_{18}
estimate, $\times 10^2$	0.03	0.06	0.09	0.09	0.32	0.53	0.96	2.24	2.67	4.19	5.40	33.04	9.20	8.63	5.28	15.25	5.06	6.96
standard error, $\times 10^2$	0.03	0.04	0.05	0.05	0.10	0.13	0.18	0.27	0.29	0.37	0.41	0.85	0.52	0.51	0.41	0.65	0.40	0.46

marginal categorical: `age76`

parameter	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}	q_{11}
estimate	0.131	0.122	0.128	0.114	0.105	0.078	0.066	0.055	0.071	0.063	0.068
standard error	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.004	0.005	0.004	0.005

Because the two/three components are both discrete and continuous, we extend the method of Anatolyev and Gospodinov (2010) of constructing a joint distribution of mixed marginals to the case of multiple values in the discrete marginal's support[13] using copula machinery. We

[12] While the minimal values of `ed76` is 1, that of `age76` is 24, therefore we simply subtract 23 from `age76` upfront for convenience.

[13] In Anatolyev and Gospodinov (2010), the discrete marginal is Bernoulli.

employ the Gaussian copula because it is simple and convenient, easily interpretable, and allows natural extension to higher dimensions with a reasonable increase in the degree of parameterization. When there is only one discrete regressor, the Gaussian copula has only one correlation parameter ϱ. It is derived in Appendix A3 that the joint density is

$$f(u,v|\theta) = f_u\left(u|\theta_u\right) f^C(u,v|\theta),$$

where

$$f^C(u,v|\theta) = \Phi\left(\frac{\Phi^{-1}(G(v|q)) - \varrho\Phi^{-1}(F_u(u|\theta_u))}{\sqrt{1-\varrho^2}}\right) - \Phi\left(\frac{\Phi^{-1}(G(v-1|q)) - \varrho\Phi^{-1}(F_u(u|\theta_u))}{\sqrt{1-\varrho^2}}\right)$$

is ‘distorted’ categorical probability, and $\theta = (\theta_u', \varrho, q')'$ collects all 21 or 14 parameters.

Maximization of the joint (log) likelihood yields estimates of parameters of the marginals very close to figures reported above but with lower standard errors, and the estimates of the copula as in the following tables:

copula: `ed76`	
parameter	ϱ
estimate	0.327
standard error	0.017

copula: `age76`	
parameter	ϱ
estimate	0.290
standard error	0.017

One can see that the estimates of bivariate degrees of dependence are highly statistically significant and moderately large in value.

Figure 2 shows the estimated mean regressions. In the case of `ed76`, it may appear that the true functional form is linear, which is what the corresponding literature tends to focus on. In the case of `age76`, linearity does not seem to hold, but a low-order polynomial like a cubic form may be appropriate. To verify whether these conjectures hold, we first perform the test for a linear mean regression:

$$\psi\left(v,\beta\right) = a + bv.$$

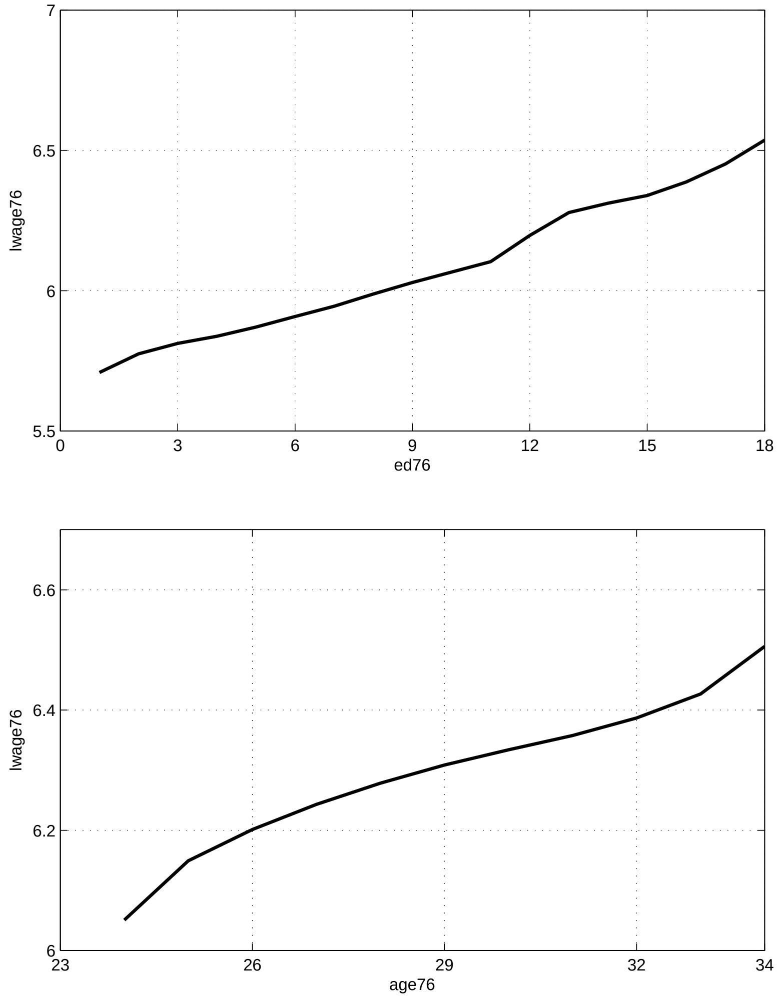

Figure 2: Estimated mean regression with regressor `ed76` (top panel) or `age76` (bottom panel)

The test results are in the following tables.

regression on: `ed76`		
test statistic		1.05
	10%	1.49
critical values, $\times 10^{-5}$	5%	1.79
	1%	2.47

regression on: `age76`		
test statistic		2.00
	10%	1.56
critical values, $\times 10^{-5}$	5%	1.99
	1%	3.00

The hypothesis of a linear regression form is decidedly rejected for both regressors at any conventional significance level; in fact, the exceedance is huge. We conclude that the form of the actual mean regression differs from what is usually assumed in regressions of wages on its determinants.

Labor econometricians often add in their linear regressions a square of a variable related to duration (e.g., work experience[14]); Murphy and Welch (1990) show that even fourth powers may be needed. Therefore, we have also run the test with low-order polynomial hypothesized regression forms: $\psi_2(v,\beta) = a + bv + cv^2$ and $\psi_4(v,\beta) = a + bv + cv^2 + dv^3 + fv^4$. These functional forms are also rejected at any conventional significance level.

When there are two discrete regressors, the Gaussian copula has a 3 × 3 correlation matrix

$$R = \begin{bmatrix} 1 & \varrho_0 & \varrho_1 \\ \varrho_0 & 1 & \varrho_2 \\ \varrho_1 & \varrho_2 & 1 \end{bmatrix}$$

with 3 distinct parameters ϱ_0, ϱ_1, ϱ_2. It is derived in Appendix A3 that the joint density is

$$f(u, v_1, v_2|\theta) = f_u(u|\theta_u) f^C(u, v_1, v_2|\theta),$$

where

$$\begin{aligned} f^C(u_1, v_1, v_2) &= \Phi_2(\varphi_1(v_1), \varphi_2(v_2)|\varphi_u(u)) - \Phi_2(\varphi_1(v_1-1), \varphi_2(v_2)|\varphi_u(u)) \\ &\quad -\Phi_2(\varphi_1(v_1), \varphi_2(v_2-1)|\varphi_u(u)) + \Phi_2(\varphi_1(v_1-1), \varphi_2(v_2-1)|\varphi(u)) \end{aligned}$$

for $v_1, v_2 \in \{0,1\}$ are 'distorted' bivariate categorical probabilities, where

$$\begin{aligned} \varphi_\ell(v) &= \Phi^{-1}(G_\ell(v)), \quad \ell = 1, 2, \\ \varphi_u(u) &= \Phi^{-1}(F_u(u)), \end{aligned}$$

and $\theta = (\theta_u', \varrho_0, \varrho_1, \varrho_2, q_1', q_2')'$ collects all 33 parameters.

Maximization of the joint (log) likelihood yields estimates of parameters of the marginals very close to figures reported above but with lower standard errors, and the estimates of the copula as in the following tables:

[14] More precisely, the potential experience is defined as age minus education less 6.

copula: ed76 and age76			
parameter	ϱ_0	ϱ_1	ϱ_2
estimate	0.187	0.328	0.290
standard error	0.019	0.017	0.017

One can see that the estimates of bivariate degrees of dependence ϱ_1 and ϱ_2 are very close to those from bivariate models with similar standard errors. The degree of dependence between the two regressors ϱ_0 is estimated to be quite modest but significantly different from zero.

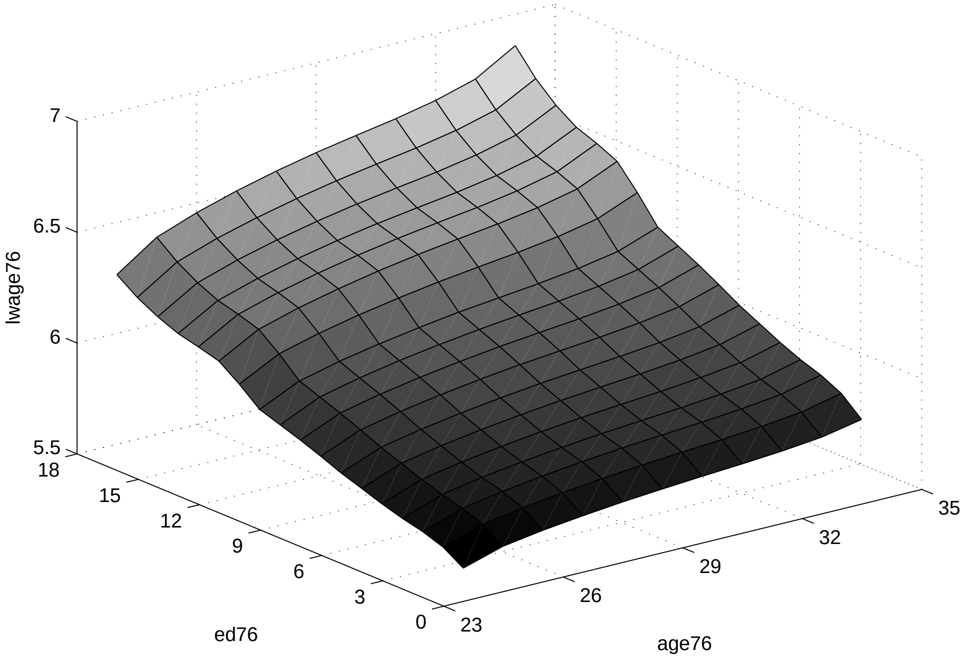

Figure 3: Estimated mean regression with regressors ed76 and age76

Figure 3 shows the surface of the estimated mean regression which is arguably close to a plane. We perform the test for a linear mean regression:

$$\psi(v_1, v_2, \beta) = a + b_1 v_1 + b_2 v_2.$$

The test results are:

regression on: `ed76` and `age76`		
test statistic		3.03
critical values, $\times 10^{-5}$	10%	2.63
	5%	3.09
	1%	4.11

The hypothesis of a linear regression form is decidedly rejected for both regressors at any conventional significance level. We also repeat this exercise for the form quadratic in both regressors $\psi_{22}(v_1, v_2, \beta) = a + b_1 v_1 + b_2 v_2 + c_1 v_1^2 + c_{12} v_1 v_2 + c_2 v_2^2$, as well as, motivated by the study of Murphy and Welch (1990), for the form linear in education and quartic in age, $\psi_{14}(v_1, v_2, \beta) = a + b_1 v_1 + b_2 v_2 + c_2 v_2^2 + c_{12} v_1 v_2 + d_2 v_2^3 + d_{12} v_1 v_2^2 + f_2 v_2^4 + f_{12} v_1 v_2^3$, as well as the same form with age v_2 replaced by potential experience that equals $v_2 + 17 - v_1$.[15]

These functional forms are also decidedly rejected at any conventional significance level. Evidently, the observable “bumps” in the curves/surface in Figures 2 and 3 are not due to a sampling error only, but rather are built-in attributes of the shapes of regressions. The overall results imply that the true mean regressions are not likely to reduce to low-order polynomials in the conditioning variables but rather take more complex functional forms, which is contradictory to popular empirical practices.[16]

6 Conclusion

We have developed a test for a restricted functional form of a mean regression function when a parametric distribution for all variables is specified and estimated. The test is based on mean-square comparison of the estimated regression implied by the joint density and estimated hypothesized functional form. The test statistic is asymptotically mixed chi-squared distributed, with the coefficients computable from the true and hypothesized regression functions and the score function. The size and power properties are favorable for sample sizes usually employed. A possible direction of future research may be extension of the test to causal regressions estimated by instrumental variables.

[15] See footnotes 12 and 14.

[16] While the regressions we have considered here are not causal (see footnote 10), the rejections obtained indirectly indicate probable misspecification of similar causal relationships used in the returns to schooling literature.

A Appendix: proofs

Proof of Lemma 1. Consistency and asymptotic normality of $\hat{\vartheta}$ follow from Newey and McFadden (1994, theorems 2.5, 2.6, 3.3 and 3.4) using Assumptions 1 and 2. □

Proof of Lemma 2. Note that

$$E\left[\left\|\frac{\partial E(u|v,\theta^0)}{\partial\vartheta}\right\|^2\right]=E\left[\left\|\int_{-\infty}^{+\infty}u\frac{\partial f(u|v,\theta^0)}{\partial\theta}du\right\|^2\right]=E\left[\left\|E\left[u\frac{\partial\ln f(u|v,\theta^0)}{\partial\theta}|v\right]\right\|^2\right]<\infty,$$

which follows from Assumption 3. Next,

$$\begin{aligned}E\left[\left\|\frac{\partial E(u|v,\theta^0)}{\partial\vartheta}\frac{\partial\psi\left(v,\beta^0\right)}{\partial\vartheta'}\right\|\right] &\leq E\left[\left\|\frac{\partial E(u|v,\theta^0)}{\partial\vartheta}\right\|\left\|\frac{\partial\psi\left(v,\beta^0\right)}{\partial\vartheta'}\right\|\right]\\ &\leq E\left[\left\|\frac{\partial E(u|v,\theta^0)}{\partial\vartheta}\right\|^2\right]^{1/2}E\left[\left\|\frac{\partial\psi\left(v,\beta^0\right)}{\partial\vartheta'}\right\|^2\right]^{1/2}<\infty,\end{aligned}$$

which follows from the previous and Assumption 2(f). Finally, $M_{\psi\psi}$ is finite by Assumption 2(f). This shows finiteness of Δ.

Now,

$$\begin{aligned}&E\left[\sup_{\theta\in\mathfrak{N}_\theta,\beta\in\mathfrak{N}_\beta}\left\|\left(\frac{\partial E(u|v_i,\theta)}{\partial\vartheta}-\frac{\partial\psi(v_i,\beta)}{\partial\vartheta}\right)\left(\frac{\partial E(u|v_i,\theta)}{\partial\vartheta}-\frac{\partial\psi(v_i,\beta)}{\partial\vartheta}\right)'\right\|\right]\\ \leq\ &E\left[\sup_{\theta\in\mathfrak{N}_\theta,\beta\in\mathfrak{N}_\beta}\left(\left\|\frac{\partial E(u|v_i,\theta)}{\partial\vartheta}\right\|+\left\|\frac{\partial\psi(v_i,\beta)}{\partial\vartheta}\right\|\right)^2\right]\\ \leq\ &2E\left[\sup_{\theta\in\mathfrak{N}_\theta}\left\|E\left[u\frac{\partial\ln f(u|v,\theta)}{\partial\theta}|v\right]\right\|^2\right]+2E\left[\sup_{\beta\in\mathfrak{N}_\beta}\left\|\frac{\partial\psi(v_i,\beta)}{\partial\beta}\right\|^2\right]<\infty\end{aligned}$$

by Assumptions 2(d) and 3. Then, by Lemma 4.3 of Newey and McFadden (1994), $\hat{\Delta}\xrightarrow{p}\Delta$. □

Proof of Theorem 1. Take a second-order stochastic expansion of $n\hat{D}$ around the true parameter value ϑ^0:

$$n\hat{D}=\sum_{i=1}^{n}\left(E\left(u|v_i,\theta^0\right)-\psi\left(v_i,\beta^0\right)\right)^2+\sqrt{n}\left.\frac{\partial\hat{D}}{\partial\vartheta'}\right|_{\vartheta^0}\hat{\zeta}_\vartheta+\hat{\zeta}_\vartheta'\left.\frac{1}{2}\frac{\partial^2\hat{D}}{\partial\vartheta\partial\vartheta'}\right|_{\vartheta^0}\hat{\zeta}_\vartheta+O_P\left(\frac{1}{\sqrt{n}}\right),$$

where

$$\hat{\zeta}_\vartheta=\sqrt{n}\left(\hat{\vartheta}-\vartheta^0\right)\xrightarrow{p}\zeta_\vartheta\overset{d}{=}\mathcal{N}\left(0,V_\vartheta\right),$$

and $V_\vartheta=H^{-1}\Omega H^{-1}$ is the asymptotic distribution of $\hat{\vartheta}$. Under H_0, the leading term is zero.

Next, under H_0,

$$\begin{aligned} \left.\frac{\partial \hat{D}}{\partial \vartheta'}\right|_{\vartheta^0} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \vartheta}\left(E\left(u|v,\theta^0\right)-\psi\left(v_i,\beta^0\right)\right)^2 \\ &= 2\frac{1}{n}\sum_{i=1}^{n}\left(E\left(u|v,\theta^0\right)-\psi\left(v_i,\beta^0\right)\right)\frac{\partial\left(E\left(u|v,\theta^0\right)-\psi\left(v_i,\beta^0\right)\right)}{\partial \vartheta} \\ &= 0. \end{aligned}$$

Finally,

$$\begin{aligned} \frac{1}{2}\left.\frac{\partial^2 \hat{D}}{\partial \vartheta \partial \vartheta'}\right|_{\vartheta^0} &= \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial \vartheta \partial \vartheta'}\left(E\left(u|v_i,\theta^0\right)-\psi\left(v_i,\beta^0\right)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \vartheta'}\left[\left(E\left(u|v_i,\theta^0\right)-\psi\left(v_i,\beta^0\right)\right)\left(\frac{\partial E\left(u|v_i,\theta^0\right)}{\partial \vartheta}-\frac{\partial \psi\left(v_i,\beta^0\right)}{\partial \vartheta}\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial E\left(u|v_i,\theta^0\right)}{\partial \vartheta}-\frac{\partial \psi\left(v_i,\beta^0\right)}{\partial \vartheta}\right)\left(\frac{\partial E\left(u|v_i,\theta^0\right)}{\partial \vartheta}-\frac{\partial \psi\left(v_i,\beta^0\right)}{\partial \vartheta}\right)' \\ &\quad +\frac{1}{n}\sum_{i=1}^{n}\left(E\left(u|v_i,\theta^0\right)-\psi\left(v_i,\beta^0\right)\right)\frac{\partial}{\partial \vartheta'}\left(\frac{\partial E\left(u|v_i,\theta^0\right)}{\partial \vartheta}-\frac{\partial \psi\left(v_i,\beta^0\right)}{\partial \vartheta}\right) \\ &\underset{H_0}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial E\left(u|v_i,\theta^0\right)}{\partial \vartheta}-\frac{\partial \psi\left(v_i,\beta^0\right)}{\partial \vartheta}\right)\left(\frac{\partial E\left(u|v_i,\theta^0\right)}{\partial \vartheta}-\frac{\partial \psi\left(v_i,\beta^0\right)}{\partial \vartheta}\right)' \\ &= \Delta + o_P\left(1\right) \end{aligned}$$

by the law of large numbers (see the proof of Lemma 2) and because

$$\begin{aligned} & E\left[\left\|\left(\frac{\partial E\left(u|v,\theta^0\right)}{\partial \vartheta}-\frac{\partial \psi\left(v,\beta^0\right)}{\partial \vartheta}\right)\left(\frac{\partial E\left(u|v,\theta^0\right)}{\partial \vartheta}-\frac{\partial \psi\left(v,\beta^0\right)}{\partial \vartheta}\right)'\right\|\right] \\ &\qquad \leq E\left[\left\|\frac{\partial E\left(u|v,\theta^0\right)}{\partial \vartheta}-\frac{\partial \psi\left(v,\beta^0\right)}{\partial \vartheta}\right\|^2\right] \\ &\qquad \leq 2E\left[\left\|\frac{\partial E\left(u|v,\theta^0\right)}{\partial \vartheta}\right\|^2+\left\|\frac{\partial \psi\left(v,\beta^0\right)}{\partial \vartheta}\right\|^2\right] < \infty. \end{aligned}$$

Summarizing, we have that under H_0,

$$n\hat{D} = \zeta_\vartheta' \Delta \zeta_\vartheta + o_P(1).$$

Now, using Lemma 3.2 from Vuong (1989), we get that

$$n\hat{D} \xrightarrow{p} \sum_{j=1}^{\dim(\vartheta)} \lambda_j \zeta_j^2,$$

where $\{\lambda_j\}_{j=1}^{\dim(\vartheta)}$ are eigenvalues of $V_\vartheta \Delta = \Lambda$, and $\{\zeta_j^2\}_{j=1}^{\dim(\vartheta)} \sim IID\chi^2_{(1)}$. □

Proof of Theorem 2. It follows from the proof of Theorem 1 that

$$n\hat{D} = \sum_{i=1}^{n} \left(E\left(u|v_i, \theta^0\right) - \psi\left(v_i, \beta^0\right)\right)^2 + O_P\left(\sqrt{n}\right).$$

Because $E\left(u|v, \theta^0\right) \neq \psi\left(v, \beta^0\right)$ almost surely, we have that $n\hat{D}$ tends to $+\infty$ as $n \to \infty$ as it is positive by construction. □

B Appendix: details on simulation experiments

Consider the setup of the first experiment. Because $E\left(u|v\right) = \mu_u + \rho\left(v - \mu_v\right)$ and $\psi\left(v, \beta\right) = a + bv$, we compute that

$$\frac{\partial E\left(u|v\right)}{\partial \vartheta} - \frac{\partial \psi\left(v, \beta\right)}{\partial \vartheta} = \begin{bmatrix} 1 \\ -\rho \\ v - \mu_v \\ -1 \\ -v \end{bmatrix}.$$

Note that there are only two non-collinear elements. Hence,

$$\Delta = \begin{bmatrix} 1 & -\rho & 0 & -1 & -\mu_v \\ -\rho & \rho^2 & 0 & \rho & \rho\mu_v \\ 0 & 0 & 1 & 0 & -1 \\ -1 & \rho & 0 & 1 & \mu_v \\ -\mu_v & \rho\mu_v & -1 & \mu_v & 1 + \mu_v^2 \end{bmatrix},$$

which, expectedly, has a rank of 2.

The logdensity is

$$\ln f(u, v|\theta) = -\ln 2\pi - \frac{1}{2}\ln\left(1 - \rho^2\right) - \frac{\left(u - \mu_u\right)^2 - 2\rho\left(u - \mu_u\right)\left(v - \mu_v\right) + \left(v - \mu_v\right)^2}{2\left(1 - \rho^2\right)},$$

and its derivatives are

$$\frac{\partial \ln f(u,v|\theta)}{\partial \theta}=\begin{bmatrix} \frac{1}{1-\rho^2}\left((u-\mu_u)-\rho\left(v-\mu_v\right)\right) \\ \frac{1}{1-\rho^2}\left((v-\mu_v)-\rho\left(u-\mu_u\right)\right) \\ \frac{\rho}{1-\rho^2}-\frac{\rho}{\left(1-\rho^2\right)^2}\left(\left(u-\mu_u\right)^2+\left(v-\mu_v\right)^2\right)+\frac{1+\rho^2}{\left(1-\rho^2\right)^2}\left(u-\mu_u\right)\left(v-\mu_v\right) \end{bmatrix}.$$

Then

$$E\left[\frac{\partial^2 \ln f(u,v|\theta)}{\partial\theta\partial\theta'}\right]=E\begin{bmatrix} -\frac{1}{1-\rho^2} & \frac{\rho}{1-\rho^2} & 0 \\ \frac{\rho}{1-\rho^2} & -\frac{1}{1-\rho^2} & 0 \\ 0 & 0 & -\frac{1+\rho^2}{\left(1-\rho^2\right)^2} \end{bmatrix}.$$

The derivatives of the hypothesized regression function are

$$\frac{\partial\psi\left(v,\beta\right)}{\partial\beta}=\begin{pmatrix}-1\\-v\end{pmatrix},$$

and hence

$$E\left[\frac{\partial\psi\left(v,\beta\right)}{\partial\beta}\frac{\partial\psi\left(v,\beta\right)}{\partial\beta'}\right]=\begin{bmatrix} 1 & \mu_v \\ \mu_v & 1+\mu_v^2 \end{bmatrix}.$$

So, the (minus) inverted Hessian is

$$-H^{-1}=\begin{bmatrix} 1 & \rho & 0 & 0 & 0 \\ \rho & 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{\left(1-\rho^2\right)^2}{1+\rho^2} & 0 & 0 \\ 0 & 0 & 0 & 1+\mu_v^2 & -\mu_v \\ 0 & 0 & 0 & -\mu_v & 1 \end{bmatrix}.$$

Next we compute

$$\begin{aligned}
E\left[\frac{\partial \ln f(u,v|\theta)}{\partial \theta}\frac{\partial \ln f(u,v|\theta)}{\partial \theta'}\right] &= \begin{bmatrix} \frac{1}{1-\rho^2} & -\frac{\rho}{1-\rho^2} & 0 \\ -\frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} & 0 \\ 0 & 0 & \frac{1+\rho^2}{(1-\rho^2)^2} \end{bmatrix}, \\
E\left[(u-\psi(v,\beta))^2 \frac{\partial \psi(v,\beta)}{\partial \beta}\frac{\partial \psi(v,\beta)}{\partial \beta'}\right] &= \left(1-\rho^2\right)\begin{bmatrix} 1 & \mu_v \\ \mu_v & 1+\mu_v^2 \end{bmatrix}, \\
E\left[(u-\psi(v,\beta)) \frac{\partial \ln f(u,v|\theta)}{\partial \theta}\frac{\partial \psi(v,\beta)}{\partial \beta'}\right] &= \begin{bmatrix} 1 & \mu_v \\ -\rho & -\rho\mu_v \\ 0 & 1 \end{bmatrix}.
\end{aligned}$$

Hence, the matrix of expected cross-products of the elements of the score vector is

$$\Omega = \begin{bmatrix} \frac{1}{1-\rho^2} & -\frac{\rho}{1-\rho^2} & 0 & 1 & \mu_v \\ -\frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} & 0 & -\rho & -\rho\mu_v \\ 0 & 0 & \frac{1+\rho^2}{(1-\rho^2)^2} & 0 & 1 \\ 1 & -\rho & 0 & 1-\rho^2 & \left(1-\rho^2\right)\mu_v \\ \mu_v & -\rho\mu_v & 1 & \left(1-\rho^2\right)\mu_v & \left(1-\rho^2\right)\left(1+\mu_v^2\right) \end{bmatrix}.$$

Then the asymptotic variance matrix is

$$V_\vartheta = \begin{bmatrix} 1 & \rho & 0 & 1-\rho^2 & 0 \\ \rho & 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{\left(1-\rho^2\right)^2}{1+\rho^2} & -\mu_v\frac{\left(1-\rho^2\right)^2}{1+\rho^2} & \frac{\left(1-\rho^2\right)^2}{1+\rho^2} \\ 1-\rho^2 & 0 & -\mu_v\frac{\left(1-\rho^2\right)^2}{1+\rho^2} & \left(1-\rho^2\right)\left(1+\mu_v^2\right) & -\mu_v\left(1-\rho^2\right) \\ 0 & 0 & \frac{\left(1-\rho^2\right)^2}{1+\rho^2} & -\mu_v\left(1-\rho^2\right) & 1-\rho^2 \end{bmatrix},$$

and, consequently,

$$V_\vartheta\Delta = 2\rho^2\frac{1-\rho^2}{1+\rho^2}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mu_v & 0 & -\mu_v \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}.$$

For the second experiment, we extend the method of Anatolyev and Gospodinov (2010) of constructing a joint distribution of mixed discrete and continuous marginals to the cases of the cardinality of the discrete marginal's support higher than two. The joint CDF/CMF is

$$F(u,v) = C(F(u), G(v)),$$

so the PDF/PMF is a derivative with respect to the continuous argument and a difference with respect to the discrete one:

$$\begin{aligned} f(u,v) &= \frac{\partial C}{\partial u}(F(u), G(v)) - \frac{\partial C}{\partial u}(F(u), G(v-1)) \\ &= f_u(u) f_\partial(u,v), \end{aligned}$$

where the second term is

$$f_\partial(u,v) = \left[\frac{\partial C}{\partial w}(w, G(v)) - \frac{\partial C}{\partial w}(w, G(v-1))\right]_{w=F(u)},$$

or

$$\begin{aligned} f_\partial(u,-1) &= \left[\frac{\partial C}{\partial w}(w, q_{-1})\right]_{w=F_u(u)}, \\ f_\partial(u,0) &= \left[\frac{\partial C}{\partial w}(w, 1-q_{+1}) - \frac{\partial C}{\partial w}(w, q_{-1})\right]_{w=F_u(u)}, \\ f_\partial(u,1) &= 1 - \left[\frac{\partial C}{\partial w}(w, 1-q_{+1})\right]_{w=F_u(u)}. \end{aligned}$$

For the FGM copula,

$$\frac{\partial C}{\partial w_1}(w_1, z) = z + \rho(1-2z)w_2(1-w_2),$$

implying the distorted success probabilities

$$\begin{aligned} q_{-1}^C(z) &= q_{-1} + \rho(1-2z)\, q_{-1}(1-q_{-1}), \\ q_0^C(z) &= 1 - q_{-1} - q_{+1} + \rho(1-2z)\left[q_{+1}(1-q_{+1}) - q_{-1}(1-q_{-1})\right], \\ q_{+1}^C(z) &= q_{+1} - \rho(1-2z)\, q_{+1}(1-q_{+1}). \end{aligned}$$

The joint density/mass is

$$f(u,v) = f_u(u) q_{-1}^C(F_u(u))^{1_{\{v=-1\}}} q_0^C(F_u(u))^{1_{\{v=0\}}} q_{+1}^C(F_u(u))^{1_{\{v=+1\}}},$$

and the result follows.

C Appendix: details on empirical illustration

We omit the parameters during the derivations. In the case of only one discrete component, the joint PDF/PMF is

$$f(u,v) = \frac{\partial C}{\partial u}(F_u(u), G_v(v)) - \frac{\partial C}{\partial u}(F_u(u), G_v(v-1)) = f_u(u) f^C(u,v),$$

where the last term is

$$f^C(u,v) = \left[\frac{\partial C}{\partial w}(w, G_v(v)) - \frac{\partial C}{\partial w}(w, G_v(v-1))\right]_{w=F_u(u)}.$$

The Gaussian copula is $C(w,y) = \Phi_2(\Phi^{-1}(w), \Phi^{-1}(y))$, where Φ_2 is CDF of the standard bivariate normal, and Φ^{-1} is inverse to the standard normal CDF. Note the important property:

$$\begin{aligned}
\frac{\partial \Phi_2(x_1,x_2)}{\partial x_1} &= \frac{\partial}{\partial x_1}\int_{-\infty}^{x_1}\int_{-\infty}^{x_2} \phi_2(t_1,t_2)dt_1dt_2 \\
&= \frac{\partial}{\partial x_1}\int_{-\infty}^{x_1}\int_{-\infty}^{x_2} \phi(t_2|t_1)\phi(t_1)dt_1dt_2 \\
&= \frac{\partial}{\partial x_1}\int_{-\infty}^{x_1} \phi(t_1)\left(\int_{-\infty}^{x_2}\phi(t_2|t_1)dt_2\right)dt_1 \\
&= \frac{\partial}{\partial x_1}\int_{-\infty}^{x_1} \phi(t_1)\Phi\left(x_2|t_1\right)dt_1 \\
&= \phi(x_1)\Phi\left(x_2|x_1\right).
\end{aligned}$$

This leads to

$$\begin{aligned}
\frac{\partial C(w,y)}{\partial w} &= \frac{\partial \Phi_2(\Phi^{-1}(w), \Phi^{-1}(y))}{\partial w} \\
&= \left.\frac{\partial \Phi_2(x_1,x_2)}{\partial x_1}\right|_{x_1=\Phi^{-1}(w), x_2=\Phi^{-1}(y)} \cdot \frac{\partial \Phi^{-1}(w)}{\partial w} \\
&= \left.\phi(x_1)\Phi\left(x_2|x_1\right)\right|_{x_1=\Phi^{-1}(w), x_2=\Phi^{-1}(y)} \cdot \left.\frac{1}{\phi\left(x_1\right)}\right|_{x_1=\Phi^{-1}(w)} \\
&= \Phi(\Phi^{-1}(y)|\Phi^{-1}(w)).
\end{aligned}$$

Then,

$$f^C(u,v) = \Phi(\Phi^{-1}(G_v(v))|\Phi^{-1}(F_u(u))) - \Phi(\Phi^{-1}(G_v(v-1))|\Phi^{-1}(F_u(u))).$$

Note that because Φ_2 is bivariate standard normal with correlation coefficient ϱ, we have, by normality of the conditional distributions under joint normality, that

$$\Phi\left(\Phi^{-1}(y)|\Phi^{-1}(w)\right) = \Phi\left(\frac{\Phi^{-1}(y) - \varrho\Phi^{-1}(w)}{\sqrt{1-\varrho^2}}\right),$$

and hence

$$f^C(u,v) = \Phi\left(\frac{\Phi^{-1}(G(v)) - \varrho\Phi^{-1}(F(u))}{\sqrt{1-\varrho^2}}\right) - \Phi\left(\frac{\Phi^{-1}(G(v-1)) - \varrho\Phi^{-1}(F(u))}{\sqrt{1-\varrho^2}}\right).$$

In the case of two discrete components, the joint PDF/PMF is

$$\begin{aligned}
f(u,v_1,v_2) &= \frac{\partial C}{\partial u}(F_u(u), G_1(v_1), G_2(v_2)) - \frac{\partial C}{\partial u}(F_u(u), G_1(v_1-1), G_2(v_2)) \\
&\quad -\frac{\partial C}{\partial u}(F_u(u), G_1(v_1), G_2(v_2-1)) + \frac{\partial C}{\partial u}(F_u(u), G_1(v_1-1), G_2(v_2-1)) \\
&= f_u(u) f^C(u,v_1,v_2),
\end{aligned}$$

where the last term is

$$\begin{aligned}
f^C(u,v_1,v_2) &= \left[\frac{\partial C}{\partial w}(w, G_1(v_1), G_2(v_2)) - \frac{\partial C}{\partial w}(w, G_1(v_1-1), G_2(v_2))\right. \\
&\quad \left. -\frac{\partial C}{\partial w}(w, G_1(v_1), G_2(v_2-1)) + \frac{\partial C}{\partial w}(w, G_1(v_1-1), G_2(v_2-1))\right]_{w=F_u(u)}.
\end{aligned}$$

Consider the 3-dimensional Gaussian copula

$$C(w,y_1,y_2) = \Phi_3(\Phi^{-1}(w), \Phi^{-1}(y_1), \Phi^{-1}(y_2)).$$

Note the property

$$\begin{aligned}
\frac{\partial \Phi_3(x_1,x_2,x_3)}{\partial x_1} &= \frac{\partial}{\partial x_1}\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\int_{-\infty}^{x_3}\phi_3(t_1,t_2,t_3)dt_1dt_2dt_3 \\
&= \int_{-\infty}^{x_2}\int_{-\infty}^{x_3}\left(\frac{\partial}{\partial x_1}\int_{-\infty}^{x_1}\phi_3(t_1,t_2,t_3)dt_1\right)dt_2dt_3 \\
&= \int_{-\infty}^{x_2}\int_{-\infty}^{x_3}\phi_3(x_1,t_2,t_3)dt_2dt_3 \\
&= \int_{-\infty}^{x_2}\int_{-\infty}^{x_3}\phi_2(t_2,t_3|x_1)\phi(x_1)dt_2dt_3 \\
&= \phi(x_1)\int_{-\infty}^{x_2}\int_{-\infty}^{x_3}\phi_2(t_2,t_3|x_1)dt_2dt_3 \\
&= \phi(x_1)\Phi_2(x_2,x_3|x_1),
\end{aligned}$$

which leads to

$$\begin{aligned}
\frac{\partial C(w,y_1,y_2)}{\partial w} &= \frac{\partial \Phi_3(\Phi^{-1}(w),\Phi^{-1}(y_1),\Phi^{-1}(y_2))}{\partial w} \\
&= \left.\frac{\partial \Phi_3(x_1,x_2,x_3)}{\partial x_1}\right|_{x_1=\Phi^{-1}(w),x_2=\Phi^{-1}(y_1),x_3=\Phi^{-1}(y_2)}\cdot\frac{\partial\Phi^{-1}(w)}{\partial w} \\
&= \phi(x_1)\Phi_2(x_2,x_3|x_1)|_{x_1=\Phi^{-1}(w),x_2=\Phi^{-1}(y_1),x_3=\Phi^{-1}(y_2)}\cdot\left.\frac{1}{\phi\left(x_1\right)}\right|_{x_1=\Phi^{-1}(w)} \\
&= \Phi_2(\Phi^{-1}(y_1),\Phi^{-1}(y_2)|\Phi^{-1}(w)).
\end{aligned}$$

Then,

$$\begin{aligned}
f^C(u_1,v_1,v_2) &= \Phi_2(\Phi^{-1}(G_1(v_1)),\Phi^{-1}(G_2(v_2))|\Phi^{-1}(F_u(u))) \\
&\quad -\Phi_2(\Phi^{-1}(G_1(v_1-1)),\Phi^{-1}(G_2(v_2))|\Phi^{-1}(F_u(u))) \\
&\quad -\Phi_2(\Phi^{-1}(G_1(v_1)),\Phi^{-1}(G_2(v_2-1))|\Phi^{-1}(F_u(u))) \\
&\quad +\Phi_2(\Phi^{-1}(G_1(v_1-1)),\Phi^{-1}(G_2(v_2-1))|\Phi^{-1}(F_u(u))).
\end{aligned}$$

As a computational matter, we use the fact that

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}|x \sim \mathcal{N}\left(\mu_\varrho x,\Omega_R\right),$$

where

$$\mu_R = \begin{pmatrix} \varrho_1 \\ \varrho_2 \end{pmatrix},\quad \Omega_R = \begin{bmatrix} 1-\varrho_1^2 & \varrho_0-\varrho_1\varrho_2 \\ \varrho_0-\varrho_1\varrho_2 & 1-\varrho_2^2 \end{bmatrix},$$

and that

$$\Phi_2(y_1, y_2|x) = \frac{1}{2\pi\sqrt{\det \Omega_R}} \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} \exp\left(-\frac{1}{2} \left(\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} - \mu_\varrho x \right)' \Omega_R^{-1} \left(\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} - \mu_\varrho x \right) \right) dz_1 dz_2.$$

References

Anatolyev, S. and N. Gospodinov (2010): "Modeling financial return dynamics via decomposition," *Journal of Business & Economic Statistics*, 28, 232–245.

Azzalini, A. (1985): "A class of distributions which includes the normal ones," *Scandinavian Journal of Statistics*, 12, 171–178.

Azzalini, A., T. Dal Cappello and S. Kotz (2003): "Log-skew-normal and log-skew-t distributions as models for family income data," *Journal of Income Distribution*, 11(3-4), 12–20.

Card, D. (1995): "Using geographic variation in college proximity to estimate the return to schooling," in: L.N. Christofides, E.K. Grant, and R. Swidinsky (editors), *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp.* Toronto: University of Toronto Press, 1995.

Judd, K. (1998): *Numerical Methods in Economics*, MIT Press.

Härdle, W. and E. Mammen (1990): "Comparing nonparametric versus parametric regression fits," *Annals of Statistics*, 21, 1926–1947.

Horowitz, J.L. and V.G. Spokoiny (2001): "An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative," *Econometrica*, 69(3), 599–631.

Massey, F.J. (1951): "The Kolmogorov-Smirnov test for goodness of fit," *Journal of American Statistical Association*, 46(253), 68–78.

Murphy, K. M. and F. Welch (1990): "Empirical age-earnings profiles," *Journal of Labor Economics*, 8(2), 202–229.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," in: *Handbook of Econometrics*, volume 4, chapter 36, 2113–2245.

Vuong, Q.H. (1989): "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, 57(2), 307–333.