A Hausman test for non-ignorability

Michael Bücker[a,*], Walter Krämer[a], Matthias Arnold[a]

[a] *Fakultät Statistik, TU Dortmund, D-44221 Dortmund, Germany*

Abstract

We propose a Hausman test for non-ignorability of missing data, which is a common problem in empirical economics. A representative case in point is the modelling of defaults in the consumer credit industry, where estimation is based solely on customers who have been granted a credit. Using an empirical likelihood approach, we show that generalized linear models can still be consistently estimated even if dependent variables are not missing at random, and derive a Hausman test by comparing this estimator to the standard one.

Keywords: Hausman test, Missing data, Empirical likelihood, Reject inference, Credit scoring, Logistic regression.

JEL classification: C12, C2, G24

1. Introduction and summary

It has long been known that maximum likelihood estimation runs into all sorts of problems in the case of missing data (see e.g. Hsiao 1980 for an early reference). The representative case of interest is a logistic regression of defaults in the consumer credit industry, where N customers apply for credit, but $N-n$ are refused, and no information concerning default is available for the latter.

A model often considered is for $i = 1, \dots, N$

$$P(Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{\beta}) := \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i})}. \quad (1)$$

The fact that $y_{n+1}, \dots, y_N$ are missing, i.e. that estimation is based only on the first n data points, does not matter if $y_{n+1}, \dots, y_N$ are missing at random (MAR) in the sense of Rubin (1976), i.e. if the distribution for the response variable Y, given all the relevant exogenous variables $\boldsymbol{X}$ of the model, is the same whether Y is observed or not,

$$P(Y|\boldsymbol{X}, R) = P(Y|\boldsymbol{X}), \quad (2)$$

[*] Corresponding author; tel.: +49-231-755-3127; fax: +49-231-755-5284.

Email addresses: buecker@statistik.tu-dortmund.de (Michael Bücker), walterk@statistik.tu-dortmund.de (Walter Krämer), arnold@statistik.tu-dortmund.de (Matthias Arnold)

where R is a binary random variable indicating the missingness of Y, say $R = 1$ if Y is observed and $R = 0$ if it is not.

However, if data are missing not at random (MNAR) in the Rubin (1976) sense, i.e. if

$$P(R = 1|\boldsymbol{X}, Y) \neq P(R = 1|\boldsymbol{X}). \tag{3}$$

ML-estimates for $\boldsymbol{\beta}$ will be inconsistent. There is some evidence that this is the norm rather than the exception in consumer credit scoring, see e.g. Crook (1999).

Although the general assumption of ignorable missingness is not testable (Manski, 2003; Jaeger, 2006), a straightforward test suggests itself in the present situation. Extending the empirical likelihood approach of Qin et al. (2002), we compare an estimator proposed by Bücker and Krämer (2011) to the standard ML-estimator and show that the resulting Hausman test has considerable power to detect deviations from ignorability.

2. The model and main results

We consider iid data of the type $(Y_i, \boldsymbol{X}_i, R_i)$ $(i = 1, \dots, N)$, where Y_i denotes the response, $\boldsymbol{X}_i$ $(k \times 1)$ is a vector of regressors, and $R_i = 1$ if Y_i is observed and $R_i = 0$ if Y_i is missing. Let $F(y, \boldsymbol{x})$ be the joint distribution function of $(Y, \boldsymbol{X})$ (no parametric model is needed for this), let

$$w(y, \boldsymbol{x}, \boldsymbol{\theta}) := P(R = 1|Y, \boldsymbol{X}, \boldsymbol{\theta})$$

be some parametric model for observability, let $W := P(R = 1)$, and consider the following semiparametric likelihood for $\boldsymbol{\theta}$, W, and F:

$$L_N(\boldsymbol{\theta}, W, F) = \left[\prod_{i=1}^{n} w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) dF(y_i, \boldsymbol{x}_i)\right] \cdot (1 - W)^{N-n}. \tag{4}$$

Maximizing this function under the constraints

$$\begin{aligned} p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \left[\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}\right] = 0, \\ \text{and} \quad \sum_{i=1}^{n} p_i \left[w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right] = 0, \end{aligned} \tag{5}$$

where $p_i = dF(y_i, \boldsymbol{x}_i) = F(y_i, \boldsymbol{x}_i) - F_{-}(y_i, \boldsymbol{x}_i)$, and profiling for all values of p_i, we obtain new weights

$$p_i = \frac{1}{n\left[1 + \boldsymbol{\lambda}_1^{\top}(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + \lambda_2(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W)\right]},$$

where $\boldsymbol{\lambda}_1$ and λ_2 are Lagrange multipliers, which we use to reweight the likelihood derived from (1) to obtain

$$L_n^{\star}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \hat{p}_i f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}). \tag{6}$$

The conventional ML-estimator $\hat{\boldsymbol{\beta}}$, which ignores all missings, is the solution to (6) without the weights $\hat{p}_i$. In Bücker and Krämer (2011), we show that, under mild regularity conditions, the modified ML-estimator $\tilde{\boldsymbol{\beta}}$ is weakly consistent and

$$\sqrt{N}\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\boldsymbol{V}\right),$$

where $\boldsymbol{\beta}_0$ denotes the true value of $\boldsymbol{\beta}$.

A major drawback of the proposed estimation method is the non-identifiability in the case of too many covariates in the missing data process. More precisely, the parameter $\boldsymbol{\theta}$ of the missing data process must not have length larger than $k+1$ since the number of free parameters must not exceed the number of estimation equations in (5). Therefore, if $w(y,\boldsymbol{x},\boldsymbol{\theta})$ is a logistic regression model and the missingness depends on Y we can only identify the parameters of $k-1$ covariates in addition to the intercept and the parameter of Y.

If the y's are missing at random, the conventional ML-estimator $\hat{\boldsymbol{\beta}}$ is asymptotically normal and efficient with covariance matrix $\boldsymbol{V}^{\star}$, say, so

$$\sqrt{N}\left(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\right)\xrightarrow{d}\mathcal{N}(\mathbf{0},\boldsymbol{V}-\boldsymbol{V}^{\star})$$

(Hausman, 1978). This follows from the fact that the difference $\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}$ must be asymptotically uncorrelated with the modified ML-estimator $\tilde{\boldsymbol{\beta}}$ due to the efficiency of the conventional ML-estimator $\hat{\boldsymbol{\beta}}$. In the MNAR case, however, $\tilde{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}_0$ whereas $\hat{\boldsymbol{\beta}}$ is inconsistent, so the statistic

$$h:=N\left(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\right)^{\top}\left(\boldsymbol{V}-\boldsymbol{V}^{\star}\right)^{-}\left(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}\right), \tag{7}$$

where $(\quad)^{-}$ denotes some generalized inverse, provides a consistent test of the MAR null hypothesis.

Under the null and some regularity conditions, h is asymptotically χ^2, with degrees of freedom equal to the rank of $\boldsymbol{V}-\boldsymbol{V}^{\star}$. This leads to all sorts of complications in applications where $\boldsymbol{V}-\boldsymbol{V}^{\star}$ is singular, but the estimate for $\boldsymbol{V}-\boldsymbol{V}^{\star}$ that is used in finite samples for the statistic (7) has full rank nevertheless (Krämer and Sonnberger, 1986). Also, the estimate of the differences of the covariance matrices can fail to be positive-definite (Schreiber, 2008). We propose to estimate the finite sample covariance matrix of $\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}$ via the bootstrap. For another application of the bootstrap in the Hausman context see Wong (1996).

3. Some Monte Carlo evidence

Tables 1 and 2 provide some Monte Carlo results for our Hausman test for linear and logistic regressions. For the former we generate an iid sample of N observations of two regressors X_1 and $X_2\sim\mathcal{N}(3,4)$. The response is defined as $Y_i=\beta_0+\beta_1X_{1,i}+\beta_2X_{2,i}+\varepsilon_i$, where the iid noise ε_i is standard Gaussian $(i=1,\ldots,N)$, $\beta_0=2$, and $\beta_1=\beta_2=1$. For the case of the logistic regression

Table 1: Empirical rejection frequencies of the Hausman test for various percentages of missing data in a linear regression model

a) MAR ($\theta_1 = 0$)

	θ_0 (resulting percentage of missings in parentheses)			
nominal significance level α	1 (77.5)	2.5 (57.5)	4 (35.2)	5 (22.5)
0.01	0.026	0.025	0.015	0.018
0.05	0.070	0.064	0.058	0.053
0.10	0.121	0.115	0.092	0.100

b) MNAR ($\theta_1 = -1$)

	θ_0 (resulting percentage of missings in parentheses)			
nominal significance level α	8 (73.0)	10 (58.1)	12 (41.9)	15 (20.8)
0.01	0.158	0.645	0.951	0.992
0.05	0.319	0.820	0.981	0.997
0.10	0.422	0.882	0.985	0.998

we have again two regressors X_1 and $X_2 \sim \mathcal{N}(0,4)$. The dependent variable is $Y_i \sim \mathcal{B}(1, \pi_i)$, where $\pi_i = \exp(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i})/[1 + \exp(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i})]$ $(i = 1, \ldots, N)$, where $\beta_0 = 2$ and $\beta_1 = \beta_2 = -1$. In both settings the observability of Y is governed by

$$w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{\exp(\theta_0 + \theta_1 y_i + \theta_2 x_{1,i})}{1 + \exp(\theta_0 + \theta_1 y_i + \theta_2 x_{1,i})} \quad (i = 1 \ldots, N).$$

We let $\theta_2 = -1$ in all experiments. The MAR-case corresponds to $\theta_1 = 0$, the MNAR case corresponds to $\theta_1 = -1$. We run 1,000 simulations for various values of the crucial parameter θ_0 which determines the proportion of missing data (the larger θ_0, the smaller the proportion of missing y's).

The tables show that our test keeps the nominal significance level quite well; it is only lightly oversized in most cases. Also, the results reveal a nontrivial power of the test procedure, with the power in the logistic regression being superior to the one observed in the linear model. Also, the fewer data are missing the higher the power of the test.

4. Discussion

The Hausman test for ignorability developed here can be applied to arbitrary parametric models of the dependency of Y on $\boldsymbol{X}$. Future research could include a generalization of the method so that parameters for all covariates are identifiable in the missingness model. Additional applications for the test can be imagined like clinical studies with nonrespondents or further inquiries where missing response occurs.

Table 2: Empirical rejection frequencies of the Hausman test for various percentages of missing data in a logistic regression model

a) MAR ($\theta_1 = 0$)

	θ_0 (resulting percentage of missings in parentheses)			
nominal significance level α	-2 (77.5)	-0.5 (57.5)	0.5 (42.5)	2 (22.5)
0.01	0.010	0.017	0.018	0.028
0.05	0.067	0.059	0.063	0.076
0.10	0.116	0.100	0.110	0.125

b) MNAR ($\theta_1 = -1$)

	θ_0 (resulting percentage of missings in parentheses)			
nominal significance level α	-2 (85.8)	0 (61.5)	1.5 (38.1)	3 (18.4)
0.01	0.265	0.956	0.996	1.000
0.05	0.481	0.989	1.000	1.000
0.10	0.596	0.995	1.000	1.000

Acknowledgements

Financial support from Deutsche Forschungsgemeinschaft (SFB 823, Project A1) is gratefully acknowledged. We are indepted Jing Qin for helpful discussions and comments.

References

Bücker, M., Krämer, W., 2011. Reject inference in consumer credit scoring with nonignorable missing data. Tech. Rep. 1/2011, SFB 823, TU Dortmund.
URL `http://www.statistik.tu-dortmund.de/sfb823-dp.html`

Crook, J., 1999. Who is discouraged from applying for credit? Economics Letters 65, 165–172.

Hausman, J. A., 1978. Specification tests in econometrics. Econometrica 46 (6), 1251–1271.

Hsiao, C., 1980. Missing data and maximum likelihood estimation. Economics Letters 6, 249–253.

Jaeger, M., 2006. On Testing the Missing at Random Assumption. Vol. 4212/2006 of Lecture Notes in Computer Science. Springer, Berlin / Heidelberg, pp. 671–678.

Krämer, W., Sonnberger, H., 1986. Computational pitfalls of the Hausman test. Journal of Economic Dynamics and Control 10 (1-2), 163–165.

Manski, C. F., 2003. Partial Identification of Probability Distributions. Springer, Berlin.

Qin, J., Leung, D., Shao, J., 2002. Estimation with survey data under nonignorable nonresponse or informative sampling. Journal of the American Statistical Association 97 (457), 193–200.

Rubin, D. B., 1976. Inference and missing data. Biometrika 36 (3), 581–592.

Schreiber, S., August 2008. The Hausman test statistic can be negative even asymptotically. Journal of Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik) 228 (4), 394–405.

Wong, K., 1996. Bootstrapping hausman's exogeneity test. Economics Letters 53, 139–143.