

Abstract

Hypothesis testing is the customary instrument for analysing the empirical validity of an economic theory. This theory is reduced to a hypothesis which is tested in a statistical model. Hypothesis testing is thus an important tool for conducting statistical inference in economic models. Instead of providing an incomplete overview of this literature, we provide a somewhat hands on discussion of testing in which we show how an economic theory is tested in a statistical model. We therefore begin with the discussion of the basic results on hypothesis testing and later focus on some recent developments that have improved testing in commonly used economic models such as the linear instrumental variables regression model. We use a real economic example to illustrate the main findings.

Testing Hypothesis testing is the customary instrument for analysing the empirical validity of an economic theory. This theory is reduced to a hypothesis which is tested in a statistical model. Hypothesis testing is thus an important tool for conducting statistical inference in economic models. An impressive literature has therefore emerged which discusses tests of economic hypotheses. Instead of providing an incomplete overview of this literature, we provide a somewhat hands on discussion of testing in which we show how an economic theory is tested in a statistical model. We therefore begin with the discussion of the basic results on hypothesis testing and later focus on some recent developments that have improved testing in commonly used economic models. We use a real economic example to illustrate the main findings.

When testing an economic hypothesis, we want our test results to hold generally and not to be affected by highly specific assumptions on the statistical model, like, for example, the distribution of the disturbances. Under these general conditions, the finite sample distributions of the involved test statistics are unknown. The realized values of the test statistics are then confronted with critical values that result from the large sample distributions of the statistics under our hypothesis of interest. Since this is currently the common approach to testing, our discussion is solely from the large sample perspective.

We illustrate the tests of an economic theory by testing for a unit demand price elasticity for the demand for oranges. We use data on the demand for oranges in the US from 1910-1959. The data results from Nerlove and Waugh (1961) and is also used in Berndt (1991, p. 417-420). The demand equation is specified as

$$\log(P_t) = \alpha + \gamma \log(Q_t) + \beta \log(RI_t) + \varepsilon_t, \quad t = 1, \dots, T, \quad (1)$$

with Q_t the traded quantity, P_t the price of oranges and RI_t the real disposable income. The other available series are current (AC_t) and past advertisement (AP_t , averaged over the last ten years). We test the hypothesis of a unit demand price elasticity $H_0 : \gamma = -1$ against the alternative hypothesis $H_1 : \gamma \neq -1$.

1 The trinity of tests

Most tests result from one of the three main principles for constructing a test statistic: Wald, likelihood ratio (LR) and Lagrange multiplier (LM) or score which are often referred to as the trinity of tests, see *e.g.* Engle (1984). When $p(y; \theta)$ is the joint density of the $T \times 1$ data vector y and we want to test the hypothesis $H_0 : \theta = \theta_0$ on the $m \times 1$ vector of parameters θ against the alternative hypothesis $H_1 : \theta \neq \theta_0$, these three statistics read:

$$\begin{aligned} \text{Wald}(\theta_0) &= T(\hat{\theta} - \theta_0)'V(\hat{\theta})^{-1}(\hat{\theta} - \theta_0) \\ \text{LR}(\theta_0) &= 2 \left[L(y; \hat{\theta}) - L(y; \theta_0) \right] \\ \text{LM}(\theta_0) &= \frac{1}{T} s(y; \theta_0)' I(\theta_0)^{-1} s(y; \theta_0), \end{aligned} \quad (2)$$

with $L(y; \theta)$ the logarithm of the likelihood or joint density $p(y; \theta)$, $L(y; \theta) = \log(p(y; \theta))$; $s(y; \theta)$ is the score, $s(y; \theta) = \frac{\partial}{\partial \theta} L(y; \theta)$; $\hat{\theta}$ is the maximum likelihood estimator under H_1 so $s(y; \hat{\theta}) = 0$, $V(\theta)$ is the covariance matrix of $\hat{\theta}$ and $I(\theta)$ is the information matrix ($V(\theta) = I(\theta)^{-1}$),

$$I(\theta) = -E \left[\frac{1}{T} \frac{\partial^2}{\partial \theta \partial \theta'} L(y; \theta) \right]. \quad (3)$$

The three statistics provide different measures of the relative distance between H_0 and H_1 . Under sufficient regularity conditions which ensure the consistency and asymptotic normality of $\hat{\theta}$, all three statistics converge under H_0 to the same $\chi^2(m)$ distributed random variable when the sample size becomes large, see *e.g.* Newey and MacFadden (1994, Th. 9.2). When these regularity conditions hold, usage of a specific statistic is typically a matter of computational ease. The Wald and LM statistics only analyze the model under H_1 or H_0 resp. so one could have a preference for either one of them given the computational effort to analyze the model under H_0 or H_1 . The LR statistic involves the analysis of the model under H_0 and H_1 and is therefore more demanding to compute than the Wald and LM statistics. When conducting tests on only one element of θ , the so-called *t*-statistic is often used which equals the square root of the Wald statistic and has a large sample normal distribution under H_0 .

2 Significance, size and power

When we test H_0 , we specify a significance level of $100 \times (1 - \alpha)\%$ which sets the probability that we reject H_0 while it is true equal to α . The critical value associated with this significance level is then such that the probability mass of the large sample distribution of the statistic under H_0 above the critical value equals α . We then reject H_0 with $(1 - \alpha) \times 100\%$ significance when the realized value of the statistic exceeds the $(1 - \alpha) \times 100\%$ critical value.

Another manner to test H_0 with $100 \times (1 - \alpha)\%$ significance is by using the *p*-value associated with the realized value of the statistic. The *p*-value equals the probability mass of the large sample distribution of the statistic under H_0 that lies above the realized value of the statistic. Hence, we reject H_0 with $100 \times (1 - \alpha)\%$ significance when the *p*-value is less than α .

Tests of $H_0 : \theta = \theta_0$ with $100 \times (1 - \alpha)\%$ significance for a range of values of θ_0 can be used to obtain the $100 \times (1 - \alpha)\%$ confidence set of θ . The $100 \times (1 - \alpha)\%$ confidence set of θ contains all values of θ_0 for which $H_0 : \theta = \theta_0$ is not rejected with $100 \times (1 - \alpha)\%$ significance and therefore contains the true value of θ with probability $1 - \alpha$.

Besides computational issues there are several other reasons to prefer a specific statistic especially when it is unclear if the regularity conditions, that imply the large sample distribution of the statistic under H_0 , hold. Examples of such other reasons are invariance to transformations of the parameters, observed size of the statistics and discriminatory power.

Especially in models that are non-linear in the parameters, it is appealing to use a statistic whose specification is invariant to non-linear transformations of the parameters so it does not depend on the specification of the model. This property is violated by the Wald statistic but satisfied by the LR and LM statistics, see *e.g.* Dagenais and Dufour (1991) and Dufour (1997). Hence, it is better to use either the LR or LM statistic in such models.

The specification of the significance level of the test intends to control the *Type I error* or probability that we reject H_0 while it holds. The rejection frequency under H_0 , to which we refer as the size of the test, should therefore coincide with α . Because we use the large sample distribution of the statistic under H_0 instead of the unknown finite sample distribution to obtain the critical value of the test, this is, however, not the case. The statistic whose size properties dominate those of the others is then typically preferred. The size properties of the different statistics can often be improved by computing the critical values using the bootstrap instead of the large sample distribution, see *e.g.* Horowitz (2001).

The *Type II error* of the statistic is the probability of not rejecting H_0 while it is false. We thus prefer statistics that minimize the *Type II error*, or, put differently, maximize the discriminatory power, while preserving an adequate size. When the likelihood function is known, the Neymann-Pearson Lemma implies that the LR statistic is the most powerful statistic for testing a point null hypothesis, like $H_0 : \theta = \theta_0$, against a point alternative, like $H_2 : \theta = \theta_2$. For composite alternatives, like $H_1 : \theta \neq \theta_0$, there is typically no statistic that is the most powerful one in all cases.

3 Tests in the linear regression model

To construct test statistics for the linear demand equation (1), we assume that the disturbances are independently and identically distributed so we can estimate the parameters using least squares¹:

$$\log(P_t) = \underbrace{-6.19}_{(1.66)} - \underbrace{0.79}_{(0.11)} \log(Q_t) + \underbrace{0.92}_{(0.23)} \log(RI_t) + \hat{\varepsilon}_t. \quad (4)$$

To compute the Wald, LR and LM statistic that test for a unit demand price elasticity, we further assume the disturbances to be independently identically normally distributed with mean zero. The specifications of the three tests then read

$$\text{Wald}(\theta_0) = \frac{\text{RSSR} - \text{USSR}}{\text{USSR}/T}, \quad \text{LR}(\theta_0) = T \log \left[\frac{\text{RSSR}}{\text{USSR}} \right] \quad \text{and} \quad \text{LM}(\theta_0) = \frac{\text{RSSR} - \text{USSR}}{\text{RSSR}/T}, \quad (5)$$

which test $H_0 : \theta = \theta_0$ and where USSR is the unrestricted sum of squared residuals $\hat{\varepsilon}_t$, *i.e.* the residuals under H_1 , and RSSR is the restricted sum of squared residuals, *i.e.* the residuals under H_0 . Under H_0 and when the disturbances are independently identically distributed with mean zero and a finite fourth order moment all three statistics converge to the same $\chi^2(1)$ distributed random variable when the sample size gets large.

The expression of the LM statistic in (5) is such that we can compute it as well using an auxiliary regression of the restricted residuals under H_0 on all explanatory variables. The expression of $\text{LM}(\theta_0)$ is then such that it equals T times the R^2 of this regression.

The values of the statistics that test for a unit demand price elasticity that result from (5) read²

$$\text{Wald}(-1) = 3.62 > \text{LR}(-1) = 3.50 > \text{LM}(-1) = 3.38. \quad (6)$$

All three statistics are smaller than the 95% critical value of 3.84 that results from the large sample distribution of the statistics under H_0 , which is the $\chi^2(1)$ distribution, so we do not reject the unit demand price elasticity with 95% significance. The Wald statistic that tests for a unit demand price elasticity exceeds the value of the LR statistic which again exceeds the value of the LM statistic. This result always holds for tests on the parameters of the linear regression model and is not a result of the involved data.

4 Specification tests

Estimation of the demand equation (1) by least squares as in (4) presumes that the traded quantity is exogenous since least squares leads to an inconsistent estimator when the traded quantity is endogenous. When the traded quantity is endogenous, we need to use an estimator that remains consistent in that

¹The standard errors are reported below the parameter estimates.

²The value of the Wald statistic in (6) is computed using (5) but could alternatively have been computed using the least squares estimator and standard error that are reported in (4) since $3.62 \approx \left(\frac{-0.79+1}{0.11} \right)^2$.

case, like, for example, the two stage least squares (2SLS) estimator or the limited information maximum likelihood (LIML) estimator, see *e.g.* Theil (1953) and Hood and Koopmans (1953).

Exogeneity or endogeneity of the traded quantity lead to different specifications of the statistical model for the demand for oranges. A test for the appropriate specification of the model is the Durbin-Wu-Hausman (DWH) statistic which tests the difference between two estimators, $\hat{\theta}$ and $\tilde{\theta}$, one of which, $\hat{\theta}$, is efficient and consistent in the model under the null hypothesis but not in the model under the alternative hypothesis while the other estimator, $\tilde{\theta}$, is consistent both in the model under the null and alternative hypothesis, see Durbin (1954), Wu (1973) and Hausman (1978):

$$\text{DWH}(\theta) = T(\tilde{\theta} - \hat{\theta})' \left[V(\tilde{\theta}) - V(\hat{\theta}) \right]^{-} (\tilde{\theta} - \hat{\theta}), \quad (7)$$

with $V(\tilde{\theta})$ and $V(\hat{\theta})$ the covariance matrices of $\tilde{\theta}$ and $\hat{\theta}$ and $[\dots]^{-}$ is the generalized inverse operator. Under sufficient regularity conditions, the DWH statistic converges under H_0 to a $\chi^2(m)$ distributed random variable with m the minimum of the number of elements of θ and the rank of the matrix $V(\tilde{\theta}) - V(\hat{\theta})$.

Using the current and past advertisement variables as instruments, we computed the DWH statistic to test the null hypothesis of exogeneity of the traded quantity against the alternative hypothesis of endogeneity using both the 2SLS and LIML estimators:

$$\begin{aligned} \text{DWH}_{2\text{SLS}}(\theta) &= T(\tilde{\theta}_{2\text{SLS}} - \hat{\theta}_{\text{LS}})' \left[V(\tilde{\theta}_{2\text{SLS}}) - V(\hat{\theta}_{\text{LS}}) \right]^{-} (\tilde{\theta}_{2\text{SLS}} - \hat{\theta}_{\text{LS}}) = 4.99, \\ \text{DWH}_{\text{LIML}}(\theta) &= T(\tilde{\theta}_{\text{LIML}} - \hat{\theta}_{\text{LS}})' \left[V(\tilde{\theta}_{\text{LIML}}) - V(\hat{\theta}_{\text{LS}}) \right]^{-} (\tilde{\theta}_{\text{LIML}} - \hat{\theta}_{\text{LS}}) = 4.96. \end{aligned} \quad (8)$$

Both statistics exceed the 95% critical value of 3.84 of the large sample distribution of the DWH statistic under H_0 which is a $\chi^2(1)$ distribution. Hence, we reject with 95% significance that the traded quantity is exogenous. This implies that we have to account for the endogeneity of the traded quantity when we test the unit demand price elasticity hypothesis.

5 Tests in the linear instrumental variables regression model

To accomodate the endogeneity of the traded quantity, we test the demand price elasticity in a linear instrumental variables regression model

$$\begin{aligned} y_t &= x_t \beta + w_t' \gamma + \varepsilon_t \\ x_t &= z_t' \pi + w_t' \delta + v_t, \end{aligned} \quad (9)$$

where y_t and x_t are the endogenous variables, w_t is a $k_w \times 1$ vector that contains the included exogenous variables, z_t is a $k_z \times 1$ vector that contains the instruments and ε_t and v_t are the disturbances. The instruments z_t are uncorrelated with the structural disturbances ε_t . We assume that the vectors of disturbances $\begin{pmatrix} \varepsilon_t \\ v_t \end{pmatrix}$, $t = 1, \dots, T$, are independently identically distributed. The variables for the demand for oranges are such that $y_t = \log(P_t)$, $x_t = \log(Q_t)$, $w_t = \begin{pmatrix} 1 \\ \log(RI_t) \end{pmatrix}$ and $z_t = \begin{pmatrix} \log(AC_t) \\ \log(AP_t) \end{pmatrix}$.

The structural parameter β is typically our parameter of interest and we therefore partial out w_t from the model by replacing all remaining variables by the residuals that result from regressing them on w_t :

$$\begin{aligned} \hat{y}_t &= \hat{x}_t \beta + \varepsilon_t \\ \hat{x}_t &= \hat{z}_t' \pi + v_t, \end{aligned} \quad (10)$$

with \hat{y}_t , \hat{x}_t and \hat{z}_t the residuals that result from regressing y_t , x_t and z_t resp. on w_t .

We want to test a hypothesis on the structural parameter β , $H_0 : \beta = \beta_0$, like, for example, that of a unit demand price elasticity. We discuss some statistics that can be used for this purpose most of which belong to the trinity of tests.

Wald statistics. Using either the 2SLS or LIML estimator, we can test H_0 using a Wald statistic:

$$\begin{aligned}\text{Wald}_{2\text{SLS}}(\beta_0) &= T(\hat{\beta}_{2\text{SLS}} - \beta_0)'V(\hat{\beta}_{2\text{SLS}})^{-1}(\hat{\beta}_{2\text{SLS}} - \beta_0) \\ \text{Wald}_{\text{LIML}}(\beta_0) &= T(\hat{\beta}_{\text{LIML}} - \beta_0)'V(\hat{\beta}_{\text{LIML}})^{-1}(\hat{\beta}_{\text{LIML}} - \beta_0),\end{aligned}\quad (11)$$

with $\hat{\beta}_{2\text{SLS}}$ the 2SLS estimator: $\hat{\beta}_{2\text{SLS}} = \left(\hat{\pi}' \sum_{t=1}^T \hat{z}_t \hat{x}_t\right)^{-1} \hat{\pi}' \sum_{t=1}^T \hat{z}_t \hat{y}_t$, $\hat{\pi} = \left(\sum_{t=1}^T \hat{z}_t \hat{z}_t'\right)^{-1} \sum_{t=1}^T \hat{z}_t \hat{x}_t$, and $\hat{\beta}_{\text{LIML}}$ the LIML estimator:

$$\hat{\beta}_{\text{LIML}} = \arg \min_{\beta} \frac{\left[\sum_{t=1}^T \hat{z}_t(\hat{y}_t - \hat{x}_t\beta)\right]' \left[\sum_{t=1}^T \hat{z}_t \hat{z}_t'\right]^{-1} \left[\sum_{t=1}^T \hat{z}_t(\hat{y}_t - \hat{x}_t\beta)\right]}{\sum_{t=1}^T (\hat{y}_t - \hat{x}_t\beta)^2 - \left[\sum_{t=1}^T \hat{z}_t(\hat{y}_t - \hat{x}_t\beta)\right]' \left[\sum_{t=1}^T \hat{z}_t \hat{z}_t'\right]^{-1} \left[\sum_{t=1}^T \hat{z}_t(\hat{y}_t - \hat{x}_t\beta)\right]}.\quad (12)$$

Both Wald statistics converge under H_0 and a number of regularity conditions which rule out zero values of π to a $\chi^2(1)$ distributed random variable when the sample size gets large, see *e.g.* Newey and MacFadden (1994). The assumption of a non-zero value of π for the large sample distribution implies that finite sample distributions of both Wald statistics depend on the value of π . The actual size of the Wald statistics can therefore deviate considerably from the assumed *Type 1 error* which makes these statistics unreliable for usage in practice, see *e.g.* Nelson and Startz (1990), Bound *et. al.* (1995), Dufour (1997) and Staiger and Stock (1997). The bootstrap can not be used to overcome these size distortions, see *e.g.* Horowitz (2002).

Anderson-Rubin statistic. Anderson and Rubin (1949) construct a test for H_0 by substituting the equation of \hat{x}_t into the equation of \hat{y}_t :

$$\hat{y}_t - \hat{x}_t\beta_0 = \hat{z}_t'\varphi + u_t,\quad (13)$$

with $\varphi = \pi(\beta - \beta_0)$ and $u_t = \varepsilon_t + v_t(\beta - \beta_0)$. Under H_0 , φ equals zero and a test for H_0 can therefore be conducted using a test for a zero value of φ . Anderson and Rubin (1949) propose to use the *F*-statistic that tests for a zero value of φ in (13) for this purpose. This *F*-statistic is commonly referred to as the Anderson-Rubin (AR) statistic.

When the disturbances are independently and identically distributed with finite fourth order moments, the AR statistic converges under H_0 to a $\chi^2(k_z)/k_z$ distributed random variable when the sample size gets large. This large sample distribution of the AR statistic does not depend on the value of π which makes the AR statistic a more reliable statistic for practical purposes than the Wald statistics in (11). A disadvantage of the AR statistic is that its large sample distribution is proportional to a χ^2 distribution with a degrees of freedom parameter that equals the number of instruments while we conduct a test on only one parameter. This reduces the discriminatory power of the AR statistic when the number of instruments is large which is often the case.

LM statistic. When we assume the vector of disturbances $\begin{pmatrix} \varepsilon_t \\ v_t \end{pmatrix}$ to be independently identically normally distributed with mean zero, we can construct the likelihood function and therefore also the LM statistic for testing H_0 , see Kleibergen (2002):

$$\text{LM}(\beta_0) = \frac{1}{\tilde{\sigma}_{\varepsilon\varepsilon}} \left(\sum_{t=1}^T \hat{z}_t \tilde{\varepsilon}_t\right)' \tilde{\pi}(\beta_0) \left[\tilde{\pi}(\beta_0)' \left(\sum_{t=1}^T \hat{z}_t \hat{z}_t'\right) \tilde{\pi}(\beta_0)\right]^{-1} \left(\sum_{t=1}^T \hat{z}_t \tilde{\varepsilon}_t\right),\quad (14)$$

with $\tilde{\varepsilon}_t = \hat{y}_t - \hat{x}_t\beta_0$,

$$\begin{aligned}\tilde{\pi}(\beta_0) &= \left(\sum_{t=1}^T \hat{z}_t \hat{z}_t'\right)^{-1} \sum_{t=1}^T \hat{z}_t \left[\hat{x}_t - \tilde{\varepsilon}_t \frac{\tilde{\sigma}_{v\varepsilon}}{\tilde{\sigma}_{\varepsilon\varepsilon}}\right], \\ \tilde{\sigma}_{\varepsilon\varepsilon} &= \frac{1}{T-k_z} \sum_{t=1}^T (\tilde{y}_t - \tilde{x}_t\beta_0)^2, \\ \tilde{\sigma}_{v\varepsilon} &= \frac{1}{T-k_z} \sum_{t=1}^T \tilde{x}_t(\tilde{y}_t - \tilde{x}_t\beta_0)\end{aligned}\quad (15)$$

and where \tilde{x}_t and \tilde{y}_t are the residuals that result from regressing x_t and y_t on w_t and z_t . When the disturbances (ε_t) are independently identically distributed and have finite fourth order moments, the LM statistic converges to a $\chi^2(1)$ distributed random variable when the sample size increases. The convergence of the LM statistic does not depend on the value of π . The large sample distribution of the LM statistic under H_0 is therefore typically a rather accurate approximation of the finite sample distribution and this approximation can even be further improved by using the bootstrap, see Kleibergen (2004).

LR statistic Under identically independently normal distributed disturbances, Moreira (2003) constructs the likelihood ratio statistic to test H_0 :

$$\text{LR}(\beta_0) = \frac{1}{2} \left[\text{AR}(\beta_0) - r(\beta_0) + \sqrt{(\text{AR}(\beta_0) + r(\beta_0))^2 - 4r(\beta_0)(\text{AR}(\beta_0) - \text{LM}(\beta_0))} \right], \quad (16)$$

with $\text{AR}(\beta_0)$ k_w times the AR statistic that tests H_0 :

$$\text{AR}(\beta_0) = \frac{1}{\tilde{\sigma}_{\varepsilon\varepsilon}} \left(\sum_{t=1}^T \hat{z}_t \tilde{\varepsilon}_t \right)' \left(\sum_{t=1}^T \hat{z}_t \hat{z}_t' \right)^{-1} \left(\sum_{t=1}^T \hat{z}_t \tilde{\varepsilon}_t \right), \quad (17)$$

and $r(\beta_0)$ a statistic that tests for a zero value of π under H_0 , so by using $\tilde{\pi}(\beta_0)$,

$$r(\beta_0) = \frac{1}{\tilde{\sigma}_{vv,\varepsilon}} \tilde{\pi}(\beta_0)' \left(\sum_{t=1}^T \hat{z}_t \hat{z}_t' \right) \tilde{\pi}(\beta_0), \quad (18)$$

where $\tilde{\sigma}_{vv,\varepsilon} = \tilde{\sigma}_{vv} - \frac{\tilde{\sigma}_{v\varepsilon}^2}{\tilde{\sigma}_{\varepsilon\varepsilon}}$ and $\tilde{\sigma}_{vv} = \frac{1}{T-k_z} \sum_{t=1}^T \hat{x}_t^2$. Moreira (2003) shows that, when the disturbances are independently identically distributed with finite fourth order moments, the large sample distribution of $\text{LR}(\beta_0)$ under H_0 is conditional on the value of $r(\beta_0)$. We thus need to use a different critical value to determine the significance of a realized value of the LR statistic for every value of $r(\beta_0)$. When $r(\beta_0)$ is zero, the large sample distribution of $\text{LR}(\beta_0)$ is identical to a $\chi^2(k_w)$ distribution while it equals the $\chi^2(1)$ distribution for large values of $r(\beta_0)$. Besides the dependence on $r(\beta_0)$, the large sample distribution of $\text{LR}(\beta_0)$ does not depend on π which makes $\text{LR}(\beta_0)$ a trustworthy statistic for practical purposes. Andrews *et. al.* (2005) show that the LR statistic is the most powerful statistic of those statistics whose large sample distributions do not depend on π and are invariant under transformations of the model.

The unit demand price elasticity. We test for a unit demand price elasticity using each of the above statistics.

$$\begin{aligned} \text{Wald}_{2\text{SLS}}(-1) &= 191 & \text{Wald}_{\text{LIML}}(-1) &= 178 \\ \text{AR}(-1) &= 73.7 & \text{LM}(-1) &= 67.3 & \text{LR}(-1) &= 69.1. \end{aligned} \quad (19)$$

The value of $r(\beta_0)$ is 174 which makes the large sample distribution of $\text{LR}(-1)$ given $r(\beta_0)$ identical to the large sample distribution of $\text{LM}(-1)$ which is a $\chi^2(1)$. The large sample distribution of $\text{AR}(-1)$ is a $\chi^2(2)$ distribution and the large sample distributions of the Wald statistics are $\chi^2(1)$ distributions while assuming a non-zero value of π .

All statistics reject the hypothesis of a unit demand price elasticity with 95% significance. This shows the importance of accounting for the endogeneity of the traded quantity since this hypothesis was not rejected in the linear regression model that assumes the traded quantity to be exogenous. The values of the statistics whose large sample distributions are not influenced by the value of π , *i.e.* $\text{AR}(-1)$, $\text{LM}(-1)$ and $\text{LR}(-1)$, are all of the same order of magnitude while the Wald statistics are much larger. This indicates the different behavior of these statistics and we recommend not to use these Wald statistics.

6 More general specifications

We discussed the trinity of tests in a linear model estimated using either least squares or instrumental variables. The Wald, LM and LR statistics extend to a large variety of models which are possibly non-linear in the parameters and have unknown likelihood functions. The expression of the Wald statistic is such that it can be applied to any estimator which has a normal large sample distribution and for which a consistent estimator of the asymptotic variance exists. The LM test is applicable in any model where the estimators solve a first order condition. The LR test is based on the difference of an objective function under H_0 and H_1 , a specification that allows it to accommodate more general statistical models, see *e.g.* Engle (1984) and Newey and MacFadden (1994). In these general settings, like, for example, the generalized method of moments, the large sample distribution of the Wald statistic is often affected by nuisance parameters while the large sample distributions of the LM and LR statistics remain robust to the value of these nuisance parameters, see *e.g.* Kleibergen (2005). Hence, the LM and LR statistics often provide more reliable tests.

Frank Kleibergen³

References

- [1] Anderson, T.W. and H. Rubin. 1949. Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations. *The Annals of Mathematical Statistics*, **21**:570–582.
- [2] Andrews, D.W.K., M.J. Moreira and J.H. Stock. 2005. Optimal Invariant Similar Tests for Instrumental Variables Regression. *Econometrica*. Forthcoming.
- [3] Berndt, E.R. 1991. *The Practice of Econometrics: Classic and Contemporary*. Addison-Wesley.
- [4] Bound, J., D.A. Jaeger and R. Baker. 1995. Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association*, **90**:443–450.
- [5] Dagenais, M.G. and J.-M. Dufour. 1991. Invariance, Nonlinear Models, and Asymptotic Tests. *Econometrica*, **59**:1601–1615.
- [6] Dufour, J.-M. 1997. Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models. *Econometrica*, **65**:1365–388.
- [7] Durbin, J. 1954. Error in Variables. *Review of the International Statistical Institute*, **22**:23–32.
- [8] Engle, R.F. 1984. Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics. In Z. Griliches and M.D. Intrilligator, editor, *Handbook of Econometrics, Volume 2*. Elsevier Science (Amsterdam).
- [9] Hausman, J.A. 1978. Specification Tests in Econometrics. *Econometrica*, **46**:1251–1272.
- [10] Hood, W.C. and T.C. Koopmans. 1953. *Studies in Econometric Method*, volume 14 of *Cowles Foundation Monograph*. Wiley (New York).

³Department of Economics, Brown University, Box B, Providence, RI 02912, United States and Department of Quantitative Economics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands, Email: Frank_Kleibergen@brown.edu. Homepage: http://www.econ.brown.edu/fac/Frank_Kleibergen.

- [11] Horowitz, J.L. 2001. The Bootstrap. In J.L. Heckman and E. Leamer, editors, *Handbook of Econometrics*, volume 5, chapter 52, pages 3159–3228. Elsevier Science B.V.
- [12] Kleibergen, F. 2002. Pivotal Statistics for testing Structural Parameters in Instrumental Variables Regression. *Econometrica*, **70**:1781–1803, 2002.
- [13] Kleibergen, F. 2004. Expansions of GMM statistics that indicate their properties under weak and/or many instruments and the bootstrap. Brown University, Working Paper.
- [14] Kleibergen, F. 2005. Testing Parameters in GMM without assuming that they are identified. *Econometrica*, **73**:1103–1124.
- [15] Moreira, M.J.,. 2003. A Conditional Likelihood Ratio Test for Structural Models. *Econometrica*, **71**:1027–1048.
- [16] Nelson, C.R. and R. Startz. 1990. Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator. *Econometrica*, **58**:967–976.
- [17] Nerlove, M. and F.V. Waugh. 1961. Advertising without Supply Control: Some Implications of a Study of the Advertising of Oranges. *Journal of Farm Economics*, **43**:813–837.
- [18] Newey, W.K. and D. McFadden. 1994. Large Sample Estimation and Hypothesis Testing. In R. Engle and D. McFadden, editor, *Handbook of Econometrics, Volume 4*, chapter 36, pages 2113–2148. Elsevier Science B.V.
- [19] Staiger, D. and J.H. Stock. 1997. Instrumental Variables Regression with Weak Instruments. *Econometrica*, **65**:557–586.
- [20] Theil, H. 1953. Estimation and Simultaneous Correlation in Complete Equation Systems. *Mimeographed Memorandum of the Central Planning Bureau, The Hague*.
- [21] Wu, D.-M. 1973. Alternative Tests of independence between stochastic regressors and disturbances. *Econometrica*, **41**:733–750.