A Direct Test for Consistency of Random Effects Models that Outperforms the Hausman Test*

Preliminary Version: This paper is under active development. Results and conclusions may change as research progresses.

Justin Esarey
Rice University
Department of Political Science
justin@justinesarey.com

Jacob Jaffe
Rice University
Department of Political Science
jej4@rice.edu

July 12, 2017

Abstract

The Hausman (1978) test is currently recommended by many textbooks in panel data analysis to decide whether fixed effects are required for management of unit heterogeneity or whether more efficient random effects can be used instead (e.g., Cameron and Trivedi, 2005, pp. 717-719; Wooldridge, 2002, pp. 288-291). However, this test has been criticized in simulation studies for both its overrejection of false nulls and underwhelming power (Clark and Linzer, 2012; Sheyatanova, 2004). We propose a more direct test: determine whether the estimated unit intercepts from a fixed effects models are correlated with the unit mean of the model's regressors using an F-test. Simulation results indicate that, when a large number of observations per unit are available, this approach is more powerful and has better size characteristics than the Hausman test and the Mundlak (1978) procedure. Furthermore, models chosen by the F-test procedure have lower bias and are more efficient than models chosen via the Hausman test.

*We thank Jude Hays, Mark Pickup, Vera Troeger, and participants at the 2017 Methods Meeting at the University of Houston and the 2017 Time and Space conference at Texas A&M University for helpful comments and suggestions on earlier drafts of this paper.

Introduction

The analysis of time series cross section (TSCS) data is complicated by the potential presence of unmodelled unit heterogeneity. These complications include bias in estimated effect sizes and artificially deflated uncertainty. Random or fixed effects models can mitigate these problems, but whether it is best to choose a random or fixed effects model depends on facts unknown to the researcher. Using a random effects model when a fixed effect model is appropriate results in biased estimates of effect size; using a fixed effects model when a random effects model is appropriate reduces power and makes it impossible to study influences that only vary between (and not within) units (Wooldridge, 2002, Chapter 10).

When choosing whether to use a fixed or random effects model for the analysis of panel data, standard textbooks recommend using the Hausman (1978) test to guide the decision (e.g., Cameron and Trivedi, 2005, pp. 717-719; Wooldridge, 2002, pp. 288-291). The Hausman test compares the estimated coefficients from a fixed effect model, $\hat{\beta}_{FE}$, to those from a random effects model, $\hat{\beta}_{RE}$. As given by Wooldridge (2002, p. 289, eq. 10.78), the Hausman test statistic is:

$$H = \left[\hat{\beta}_{FE} - \hat{\beta}_{RE}\right]' \left[\widehat{\text{var}}(\hat{\beta}_{FE}) - \widehat{\text{var}}(\hat{\beta}_{RE})\right]^{-1} \left[\hat{\beta}_{FE} - \hat{\beta}_{RE}\right] \tag{1}$$

This statistic is asymptotically distributed χ^2 with degrees of freedom equal to the rank of $\hat{\beta}_{FE}$ under the null hypothesis that the random effects estimator is consistent (Hausman, 1978, p. 1256, Theorem 2.1).

Intuitively, Hausman's approach is to compare the behavior of an inefficient, but consistent estimate $\left(\hat{\beta}_{FE}\right)$ to that of a possibly inconsistent estimate $\left(\hat{\beta}_{RE}\right)$; if they are similar, then both are achieving the same limiting distribution and we can infer that $\hat{\beta}_{RE}$ must be consistent. This test therefore *indirectly* tests the condition that there is no correlation between the regressors and the unit effects, which is necessary in order for random effects models to be consistent.

Unfortunately, prior studies have shown that the Hausman test yields deficient results in practice. Sheyatanova (2004) conducted a thorough simulation study of how the Hausman test performs in panel data, finding that:

> Hausman test over-rejects the null hypothesis if performed based on its asymptotical critical values, when Swamy and Arora and Amemiya methods are used for estimating the random effects model. The Nerlove method of estimation leads to extreme under-rejection of the null-hypothesis.

As another example, Clark and Linzer (2012, pp. 15-18) conduct a simulation study that shows that the Hausman test frequently fails to reject the null hypothesis when there is a moderate level of correlation between a regressor and the unit effects.

An Alternative to the Hausman Test

Rather than use the indirect approach of Hausman (1978), we propose to *directly* test the assumption of correlation between the regressors and the unit effects. Consider an illustrative data generating process (DGP):

$$y_{gi} = \beta_0 + \beta_x x_{gi} + \beta_z z_{gi} + \gamma_g + \varepsilon_{gi} \tag{2}$$

where g indexes units (groups) from $g \in \{1...G\}$, i indexes observations within a group from $i \in \{1...N_g\}$, γ_g is the unit effect specific to group g, x and z are independent variables, and ε is an error term. If there is correlation between x and γ, then $\pi_x \neq 0$ in:

$$\gamma_g = \pi_0 + \pi_x x_{gi} + \pi_z z_{gi} + \psi_{gi} \tag{3}$$

where ψ_{gi} is an error term.

We can test the hypothesis that $\pi_x = \pi_z = 0$ directly if we can obtain feasible estimates of γ_g. In many cases, a consistent fixed effects model produces these estimates, $\hat{\gamma}_g$; for example, an ordinary least squares dummy variable model often produces these estimates as coefficients on the unit specific dummy variables. Thus, we propose:

1. Run a fixed effects (FE) model predicting y_{gi} using all regressors (in the example, x_{gi} and z_{gi}.
2. If the estimated fixed effects $\hat{\gamma}_g$ from the FE model are jointly statistically significant, save these fixed effects and proceed; otherwise, drop the fixed effects and estimate a pooled model.[1]
3. Run an auxiliary regression:

$$\hat{\gamma}_g = \hat{\eta}_0 + \hat{\eta}_x \bar{x}_g + \hat{\eta}_z \bar{z}_g + \hat{\zeta}_g \tag{4}$$

   where $\bar{v}_g$ represents the estimated within-group mean for variable v; for example:

$$\bar{x}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} (x_{gi}) \tag{5}$$

4. Conduct an F-test of the hypothesis that $\hat{\eta}_x = \hat{\eta}_z = 0$.

Note that we propose γ_g be regressed against group mean values of the independent variables, not their individual-level values (as in equation 3). We make this choice because the error terms of observations inside of a unit g may be correlated with one another in ways that are

[1] If $\gamma_g = 0$ for all $g \in \{1...G\}$, small deviations between the true β and the estimated $\hat{\beta}$ could lead to a false finding of correlation between $\hat{\gamma}_g$ and any covariate with strong unit clustering. Testing for the joint statistical significance of $\hat{\gamma}_g$ serves to block this possibility. Thanks to Jude Hays and Vera Troeger for pointing out this possibility. See footnote 4 for more details.

not removed by the fixed effects model in step 1. For example, suppose that in the DGP of equation (2):

$$\gamma_g \sim \phi(\mu = 0, \sigma = 1) \tag{6}$$

$$\mu_g^x \sim U[0, 1] \tag{7}$$

$$x_{ig} \sim \phi(\mu = \mu_g^x, \sigma = 1) \tag{8}$$

$$z_{ig} \sim \phi(\mu = 0, \sigma = 1) \tag{9}$$

A least squares dummy variable model would, under appropriate conditions, be able to accurately recover estimates of γ_g. However, although a regression of γ_g against x_{ig} (a model of equation (3)) would accurately recover no relationship, the error term of this regression ψ_{gi} would be correlated with with x_{gi} by construction and consequently the standard errors would be biased.

Even in a situation like this one, we may be able to assume independence *between* units; this is similar to the logic that motivates use of cluster-robust standard errors in many applied studies (Liang and Zeger, 1986; Rogers, 1993). By definition, $\hat{\gamma}_g$ does not vary within units; therefore, $\eta_x = \pi_x$ and $\eta_z = \pi_z$. However, ζ_g will be uncorrelated with $\bar{x}_g$ when ψ_{ig} is correlated with x_{ig}. Thus, the group-level regression of equation (4) will be unbiased when a model of equation (3) would be biased.

In the auxiliary regression of equation (4):

$$\hat{\gamma}_g = \hat{\eta}_0 + \sum_{j=1}^{J} \hat{\eta}_j \bar{x}_{j(g)} + \hat{\zeta}_g$$

the fact that a standard F-test for $\{\hat{\eta}_1, \hat{\eta}_2, ..., \hat{\eta}_J\}$ is valid (given homoskedasticity of ζ_g) is supported by a textbook econometric proof (see Wooldridge, 2002, Lemmas 3.8-3.9 on pp. 43-44 and Theorem 4.2 on p. 55). We might be concerned that ζ_g is heteroskedastic; for example, different group sizes N_g may lead to differences in the accuracy of $\hat{\gamma}_g$ across groups g.[2] When ζ is heteroskedastic, heteroskedasticity-robust versions of the F-statistic must be used instead of the standard F-test (Wooldridge, 2002, p. 57-58). We examine scenarios that might create heteroskedasticity in our simulation analysis; we also assess the effectiveness of heteroskedasticity-robust versions of our F-test procedure, the Hausman test, and the Mundlak model.

Simulation Evidence

To assess the ability of the Hausman test, F-test, and the Mundlak (1978) procedure to determine the need for a fixed effects model in an applied setting, we conduct simulation study of panel data with and without correlation between regressors and unit effects.[3] All simulations were performed using `R` 3.3.3 (R Core Team, 2017).

Data Generating Process (DGP) and Test Procedures

The baseline simulated DGP for our simulation study is:

$$y_{gi} = \beta_0 + \beta_x x_{gi} + \beta_z z_{gi} + \beta_w w_{gi} + \gamma_g + \varepsilon_{gi} \tag{10}$$

Our study included simulations for $G \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$ and for $N_g \in \{5, 10, 15, 20, 30, 40, 50\}$. We studied $R = 1000$ simulated data sets for each combination of G and N_g. We set β_0, β_x, and β_z all equal to zero and $\beta_w = 1$. Both the group effect γ and the error term ε are drawn from the normal distribution with a mean of 0 and standard deviation equal to 1.

[2] Note that differential measurement error could only create a problem for the efficiency of $\hat{\eta}$ estimates if the accuracy of $\hat{\gamma}_g$ was associated with $\bar{x}_g$.

[3] The basic data generating process and code base for these simulations is taken from the replication code for Esarey and Menger (2017); consequently, much of the text describing the DGP is similar to the corresponding text of that paper.

The regressor x is drawn from the normal density $\phi(\mu = \mu_g^x, \sigma = 0.1)$; that is, groups share a common mean for their x values, μ_g^x, that is different from the mean x for other groups. When there is no correlation between the unit effect and the regressor, $\mu_g^x \sim U[0, 1]$; in this case, a random effects model is appropriate. When there *is* correlation between the unit effect and regressor, such that a fixed effect model is required, $\mu_g^x = 0.35\gamma_g$. The other regressors z and w are both drawn from independent normal densities with a mean of 0 and standard deviation of 1.

After creating a data set with G groups and N_g observations per group, we estimate both fixed and random effects models using the `PLM` package of Croissant and Millo (2008); these models include x, z, and w as regressors. We use the results of these models to run a Hausman test using the `phtest` command and that the test rejects the null (that a random effects model is appropriate) whenever $p < 0.05$. We also implemented the Mundlak (1978) procedure of adding mean values of the regressors to a random-effects model:

$$y_{gi} = \hat{\beta}_0 + \hat{\beta}_x x_{gi} + \hat{\beta}_z z_{gi} + \hat{\beta}_w w_{gi} + \hat{\beta}_{\mu(x)} \bar{x}_g + \hat{\beta}_{\mu(z)} \bar{z}_g + \hat{\beta}_{\mu(w)} \bar{w}_g + \hat{\varepsilon}_{gi} \tag{11}$$

We conduct a Wald test of the joint significance of $\hat{\beta}_{\mu(x)}$, $\hat{\beta}_{\mu(z)}$, and $\hat{\beta}_{\mu(w)}$ and rejected the null (that the random effects model is appropriate) whenever the associated $p < 0.05$.

For our proposed F-test procedure, we save the estimated $\hat{\gamma}$ from our fixed effects model[4] and run a second regression predicted $\hat{\gamma}$ using the unit average values of x, z, and w, corre-

[4]Because we knew that all models in our test included meaningful unit-specific heterogeneity and to ensure a proper comparison with the Hausman test, we did not test the joint significance of the unit effects $\hat{\gamma}$ in our simulations. However, as a direct assessment of whether the problem pointed out in 1, we conducted an auxiliary test setting $N_g = 5$, $G = 50$ and varying the standard deviation of γ_g between 0 and 0.2. The results are featured in Appendix Figure 8. Our analysis found persistent overrejection of a false null hypothesis by the F-test as the standard deviation of $\gamma_g \to 0$; the test rejected at a rate of around 10% despite a nominal $\alpha = 0.05$. By contrast, both the Hausman and Mundlak procedures rejected the null at rates closer to 5%.

sponding to the following model:

$$\hat{\gamma}_g = \hat{\eta}_0 + \hat{\eta}_x \bar{x}_g + \hat{\eta}_z \bar{z}_g + \hat{\eta}_w \bar{w}_g + \hat{\zeta}_g \tag{12}$$

We then record the result of an F-test of the null hypothesis ($\alpha = 0.05$) that $\eta_x = \eta_z = \eta_w = 0$; this should be true whenever a random effects model is appropriate (i.e., there is no correlation between unit effects and the regressors).

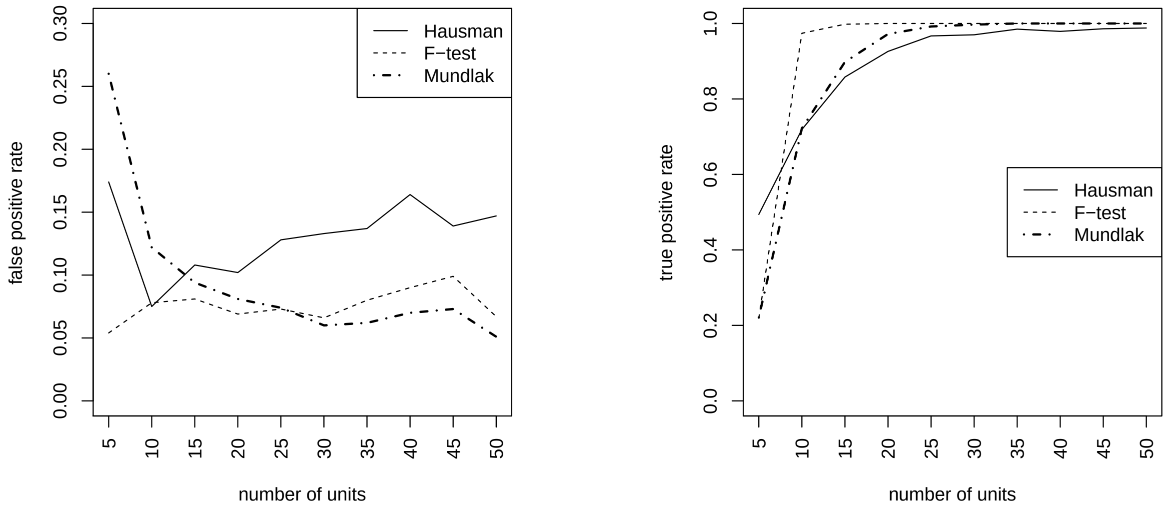
Size and Power Results

The main results of this study pertaining to the size and power of the three candidate test procedures when $N_g = 15$ (i.e., 15 observations per unit) are shown in Figure 1. The left panel of the figure assesses the size of each test (i.e., its propensity to reject the null hypothesis that a random effects model is consistent when this null is true). The right panel assesses the power of the test (i.e., the test's ability to reject the null hypothesis when the null is false and a random effects model is inconsistent).

The left panel indicates that only the F-test procedure rejects true null hypotheses at the prescribed rate of $\alpha = 0.05$ across all values of the number of units G when $N_g = 15$. The Mundlak (1978) procedure rejects the null too often for small values of G, but eventually approaches the correct rejection rate as G approaches 30. The Hausman (1978) test generally rejects the null too often for all values of G.

The right panel in the figure shows that the F-tests has poor power characteristics for a very small number of units, but quickly improves on this performance as G gets larger. Indeed, the F-test is more powerful than the Hausman test for all values of $G > 5$ in our study.

Figure 2 compares the performance of each test, but fixing $G = 50$ and allowing N_g to vary between 5 and 50. As the figure shows, the quality of the F-test results is sensitive to the



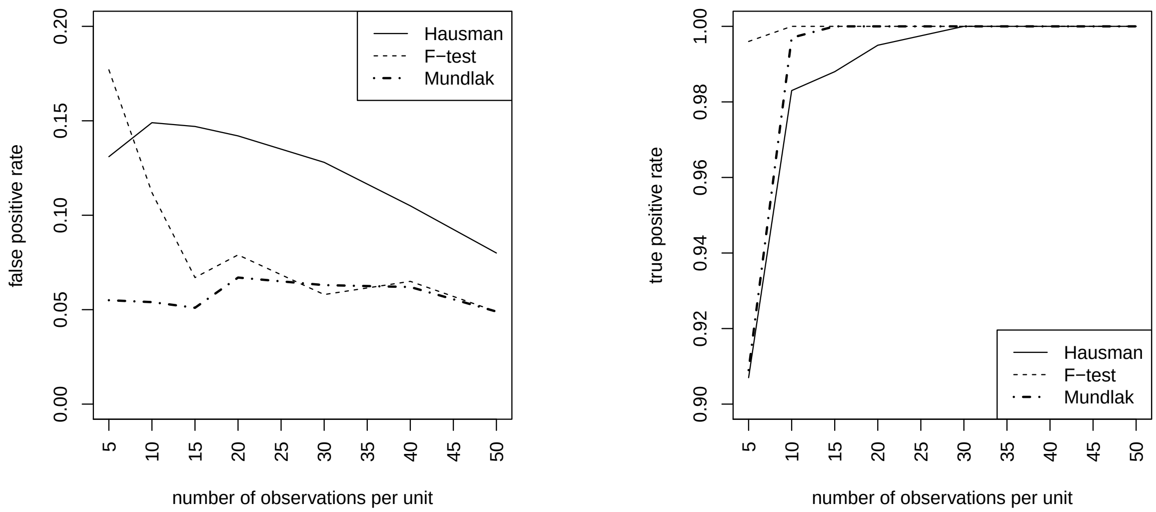
(a) Without Unit/Regressor Correlation

(b) With Unit/Regressor Correlation

Figure 1: Test Size and Power from Cross-Sectional Data Simulations with Unit Heterogeneity, $N_g = 15$

number of observations per unit. When the number of observations per unit is particularly small (less than 15) but the number of units is large (50), the size of the F-test is above its nominal $\alpha = 0.05$ level (though it still performs better than the Hausman test for $N_g > 5$). As expected from Figure 1, the Mundlak procedure does well in this environment.

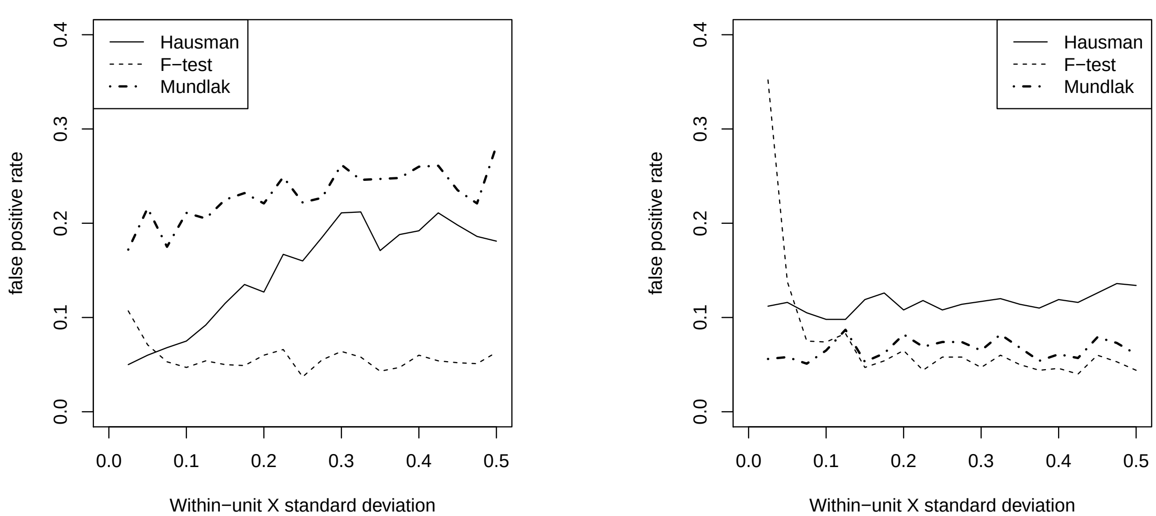
Figure 3 provides insight into why the F-test struggles to maintain appropriate size in certain situations; this Figure repeats the baseline simulations without correlation between the unit effect and regressor (where the random effects model is appropriate), but fixes $N_g = 20$ and varies $\text{Var}[x] \in \{0.025, 0.5\}$. As the figure shows, when $\text{Var}[x] \to 0$ the performance of the F-test drops off, sometimes substantially. The problem is that β_x and γ_g become harder to separately identify as $\text{Var}[x] \to 0$; when $\text{Var}[x] = 0$ we can no longer include x in a model of equation (2) because x and the fixed effects are perfectly collinear. This evidently makes the F-test more prone to choose the fixed effects estimator.



(a) Without Unit/Regressor Correlation

(b) With Unit/Regressor Correlation

Figure 2: Test Size and Power from Cross-Sectional Data Simulations with Unit Heterogeneity, $G = 50$



(a) $G = 6$

(b) $G = 30$

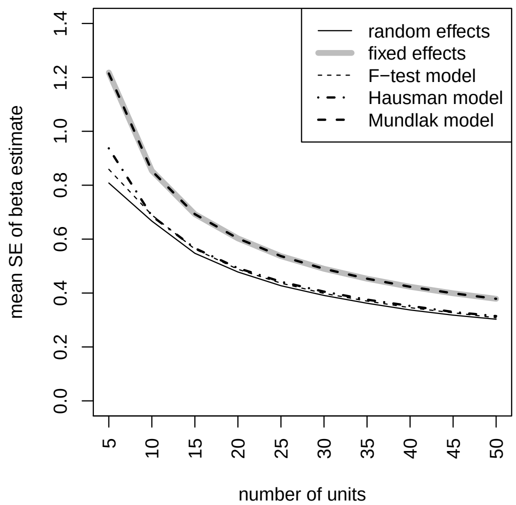
Figure 3: Test Size and Power from Cross-Sectional Data Simulations with Unit Heterogeneity and without Unit/Regressor Correlation, Varying $\text{Var}[x]$, $N_g = 20$

Bias and Efficiency Assessment

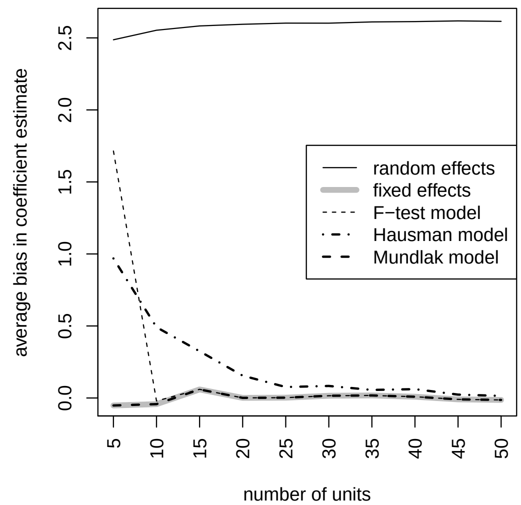
Although assessing the size and power of Hausman, Mundlak, and F-tests is important, the underlying tension in making the choice between random and fixed effects models is a choice between the possible *bias* of a random effects model versus the *inefficiency* of a fixed effects model. Thus, one critical measure of the quality of any of these tests is the extent to which they maximize the efficiency and minimize the bias of $\hat{\beta}$ estimates.

Figure 4b assesses the efficiency and bias of models chosen by each candidate test when each unit contains a limited amount of data ($N_g = 15$). The left panel (Figure 4a) shows the mean standard error of the estimated $\hat{\beta}_x$ in equation (10) for the model chosen by each test when there is no correlation between γ_g and x_{gi}. In this situation, the random effects estimator is unbiased and the most efficient choice. The right panel (Figure 4b) shows the mean bias in $\hat{\beta}_x$ for the model chosen by each test when γ_g and x_{gi} are corrlated; in this scenario, random effects is biased. Note that for the Mundlak procedure, we always assess the bias and efficiency of $\hat{\beta}_x$ estimates for the saturated model in equation (11) regardless of whether the coefficients on the unit mean variables are jointly statistically significant; a "saturated" model includes unit-mean variables for all regressors in the model. We unconditionally assess the Mundlak procedure in order to determine whether a researcher gains any advantage over fixed effects models by automatically including all unit mean terms in a random effects model.

The figure clearly illustrates that random effects is the optimal model to choose when there is no correlation between unit heterogeneity and a regressor, while the fixed effects model is the optimal choice when such correlation exists. The Hausman model chooses the random effects model too frequently when the fixed effects model is required (Figure 4b), resulting in substantial bias that diminishes very slowly as the number of units G increases to 50. The F-test procedure, however, is almost as efficient as choosing random effects by default (in Figure 4a) and yet nearly as good at minimizing bias compared to choosing fixed



(a) Without Unit/Regressor Correlation

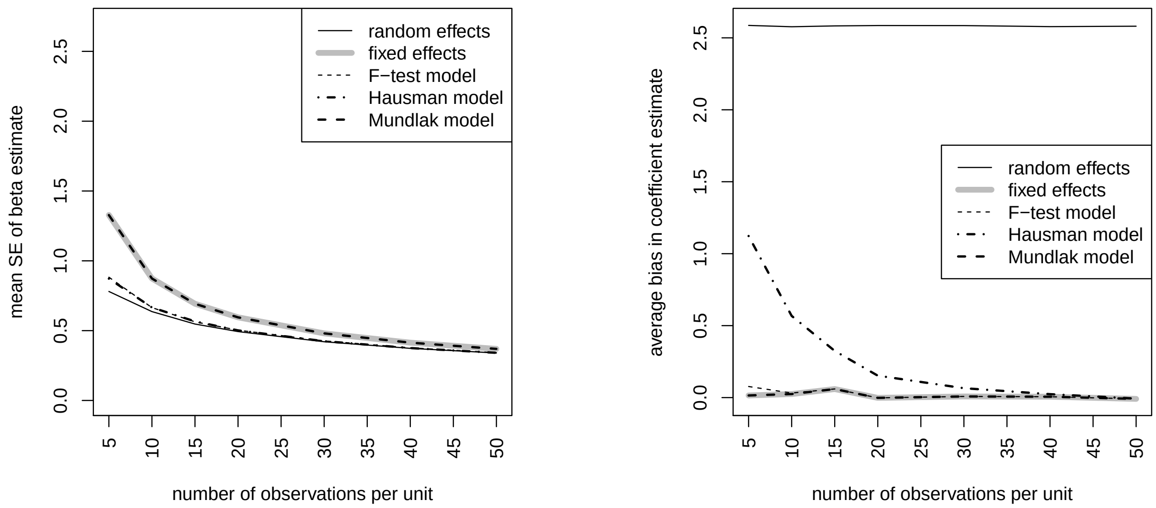


(b) With Unit/Regressor Correlation

Figure 4: Bias and Efficiency from Cross-Sectional Data Simulations with Unit Heterogeneity, $N_g = 15$

effects by default (Figure 4b) for all but the smallest number of clusters (when $G = 5$, the F-test does even worse than the Hausman test). A saturated Mundlak model is essentially equivalent to a fixed effects model in terms of both efficiency and bias consequences.

Figure 5 performs the same analysis as Figure 4, but setting the number of units $G = 15$ and allowing the number of observations per unit N_g to vary. Both the Hausman and F-test models perform essentially the same when efficiency is assessed (Figure 5a), but the F-test minimizes bias much more effectively than the Hausman test (Figure 5b) are assessed. As before, the Mundlak model yields equivalent consequences to using a fixed effects model by default.



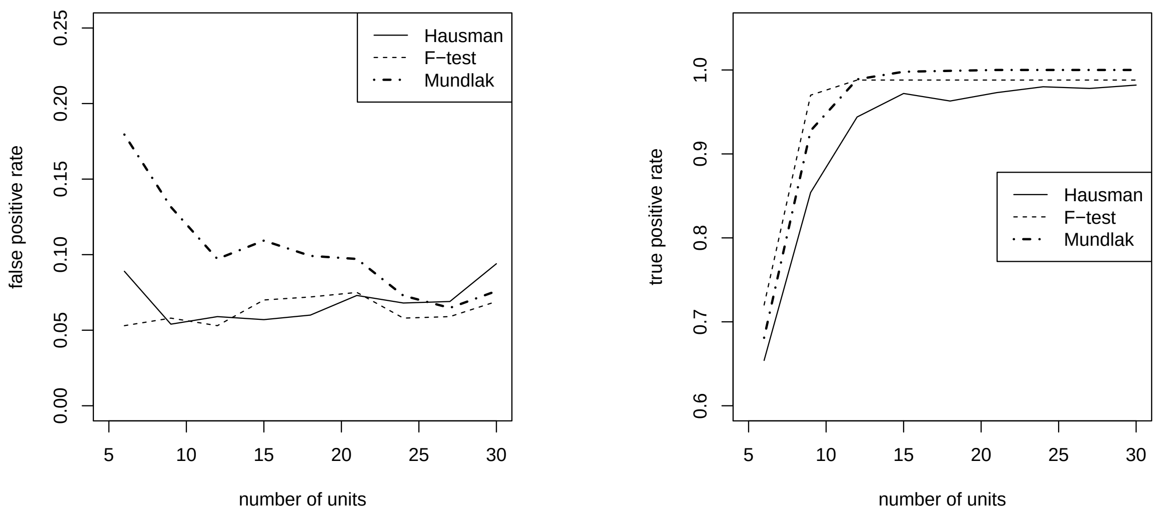
(a) Without Unit/Regressor Correlation (b) With Unit/Regressor Correlation

Figure 5: Bias and Efficiency from Cross-Sectional Data Simulations with Unit Heterogeneity, $G = 15$

Heteroskedasticity-robust testing

It is common in applied data sets for N_g to vary between clusters; as noted before, this is a potential concern if it creates heteroskedasticity in the error term ζ_g in the equation for (4). We therefore examined the performance of the Hausman, Mundlak, and F-tests in an environment with widely varying clusters. To accomplish this, we altered our simulated DGP to create data sets with an equal distribution of clusters with each of $N_g \in \{10, 25, 50\}$. We varied the total number of clusters in the simulated data set to be $G \in \{6, 9, 12, 15, 18, 21, 24, 27, 30\}$ (i.e., all numbers divisible by three). We then implemented all the same test procedures as before with no adjustment for heteroskedasticity.

The results are shown in Figure 6. It is immediately apparent that the substantive interpretation of this figure is similar to that of Figure 1, which depicts the same results for a DGP with identical cluster sizes. From this result, it appears that cluster size heterogeneity



(a) Without Unit/Regressor Correlation

(b) With Unit/Regressor Correlation

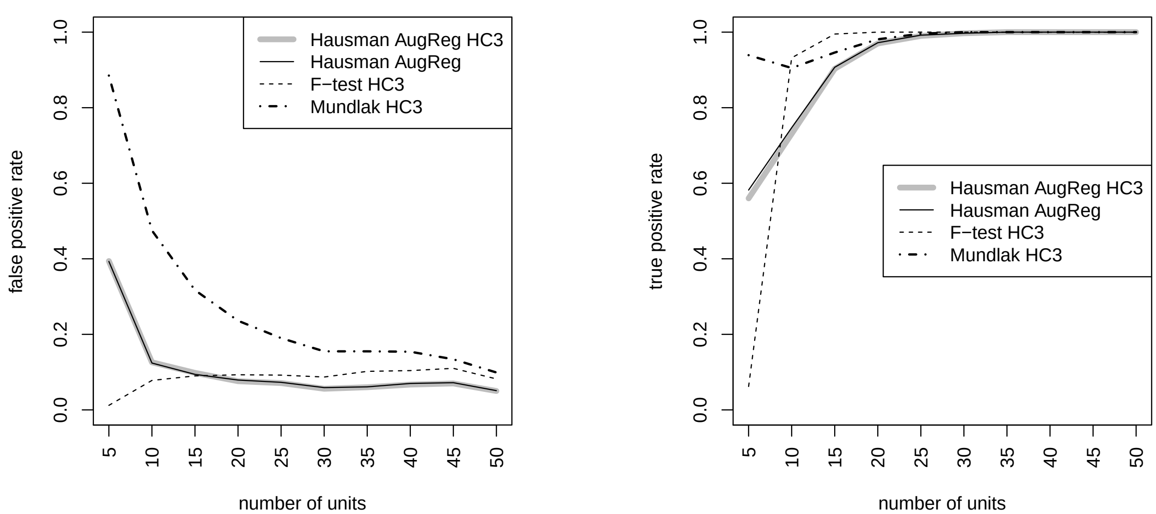
Figure 6: Test Size and Power from Cross-Sectional Data Simulations with Unit Heterogeneity and Varying Unit Sizes, $N_g \in \{10, 25, 50\}$

alone is not sufficient to create worrisome heteroskedasticity in the F-test procedure.

However, using heteroskedasticity-robust procedures for the Hausman test, F-test, and Mundlak (1978) procedures does have some impact on the performance of the procedures we study. To show this, we repeated our initial simulation (in Figures 1 and 2) using alternative forms of the Hausman test, F-test, and Mundlak procedure. For the Hausman test, we used the augmented regression variant of the Hausman test procedure described in Wooldridge (2002, p. 332)[5] that enables us to use HC3-type heteroskedasticity-consistent variance-covariance matrices in the comparison of the fixed and random effects models. For comparison with our prior results that employ the usual χ^2 version of the Hausman test, we also estimate the augmented regression Hausman test without the HC3 VCV. Finally, we use HC3 VCVs when calculating the F-statistics associated with our F-test statistic and the Mundlak (1978) test.

[5]See also Millo (2012).

The results of this reanalysis are shown in Figure 7, which mirrors the earlier Figure 1. The left panel (Figure 7a) shows that using robust standard errors tends to harm the size performance of all the tests, but in different ways. For the F-test, the heteroskedasticity-robust statistics tend to reduce the size for very small numbers of units ($G = 5$); that is, the F-test now rejects the null less than $\alpha = 0.05$ proportion of the time when the null is true. The test's power is negatively impacted for small G as well (as shown in the right panel, Figure 7b). Meanwhile, both the Hausman and Mundlak procedures becomes more *overconfident* for small values of G; that is, they rejects the true null far more than 0.05 proportion of the time. The Hausman test's false positive rate seems to converge toward $\alpha = 0.05$ after $G = 20$ or so, an improvement over the performance of the χ^2 version of the test, but the Mundlak test's size performance is as bad or worse than the Hausman test at every number of units. The Hausman test's power is roughly comparable under heteroskedasticity-robust statistics to its power using the χ^2 version of the test.



(a) Without Unit/Regressor Correlation

(b) With Unit/Regressor Correlation

Figure 7: Test Size and Power from Cross-Sectional Data Simulations with Unit Heterogeneity using HC3 VCV, $N_g = 15$

Conclusion

Our simulation studies confirm earlier results (Sheyatanova, 2004; Clark and Linzer, 2012) showing that the Hausman (1978) test struggles to maintain appropriate size and power characteristics under realistic conditions. For social scientists, the upshot is that researchers may find it difficult to make informed decisions about whether to use a fixed or random effects model on any particular data set.

We find that an alternative F-test procedure outperforms both the Hausman test and another test first suggested by Mundlak (1978). Moreover, it is simple to perform: researchers need only regress the estimated unit effects (obtained by running a fixed effects model) against the unit mean values of all independent variables; if an F-test rejects the null hypothesis of no correlation between regressors and unit effects, then a fixed effects model is required.

Our alternative is not without limitation: it relies upon a researcher's ability to obtain consistent estimates of the unit effects. This is not always possible. For example, it is trivially true that correlation between the unit effects and any variable that does not vary within unit cannot be assessed by our test. Our simulations also reveal that the F-test procedure can be misleading when variables are included that are close to not varying within units. As another example, estimating fixed effects in a probit model results in inconsistent estimates when $G \to \infty$ and N_g is constant (Wooldridge, 2002, p. 484); our test would be inappropriate in this scenario, at least when N_g is small.

However, when our procedure *is* feasible, it outperforms both the Hausman and Mundlak procedures in both controlling size and maximizing power in our simulations. It also tends to produce more efficient and less biased estimates compared to these other procedures, at least when N_g is large. Furthermore, although researchers may not be able to tell whether the Hausman test will be effective in their data set, we believe that the limitations of our procedure are straightforward enough such that an analyst will know whether it is appropriate to apply.

References

Cameron, A. Colin and Pravin K. Trivedi. 2005. *Microeconometrics: methods and applications*. Cambridge: Cambridge University Press.

Clark, Tom S. and Drew Linzer. 2012. "Should I Used Fixed or Random Effects?" Online. URL: `https://datajobs.com/data-science-repo/Fixed-Effects-Models-[Clark-and-Linzer].pdf` accessed 1/6/2017.

Croissant, Yves and Giovanni Millo. 2008. "Panel Data Econometrics in R: The plm Package." *Journal of Statistical Software* 27(2).

Esarey, Justin and Andrew Menger. 2017. "Practical and Effective Approaches to Dealing with Clustered Data." *Political Science Research and Methods* (conditionally accepted). URL: `https://goo.gl/4NGDem`.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica* 46(6):1251–1271. URL: `http://www.jstor.org/stable/1913827` accessed 1/6/2017.

Liang, Kung-Yee and Scott L. Zeger. 1986. "Longitudinal data analysis using generalized linear models." *Biometrika* 73(1):13–22.

Millo, Giovanni. 2012. "Hausman test error solve." Online. URL: `https://stat.ethz.ch/pipermail/r-help/2012-October/339401.html` accessed 7/11/2017.

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46(1):69–85. URL: `http://www.jstor.org/stable/1913646` accessed 1/6/2017.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org/*

Rogers, William. 1993. "Regression standard errors in clustered samples." *Stata Technical Bulletin* 13:19–23.

Sheyatanova, Teodora. 2004. The Accuracy of the Hausman Test in Panel Data: a Monte Carlo Study. Master's thesis Örebro University. URL: `http://oru.diva-portal.org/smash/get/diva2:805823/FULLTEXT01.pdf` accessed 1/6/2017.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Appendix: Assessment of Spurious Correlation between Unit Mean Regressors and Unit Effects γ

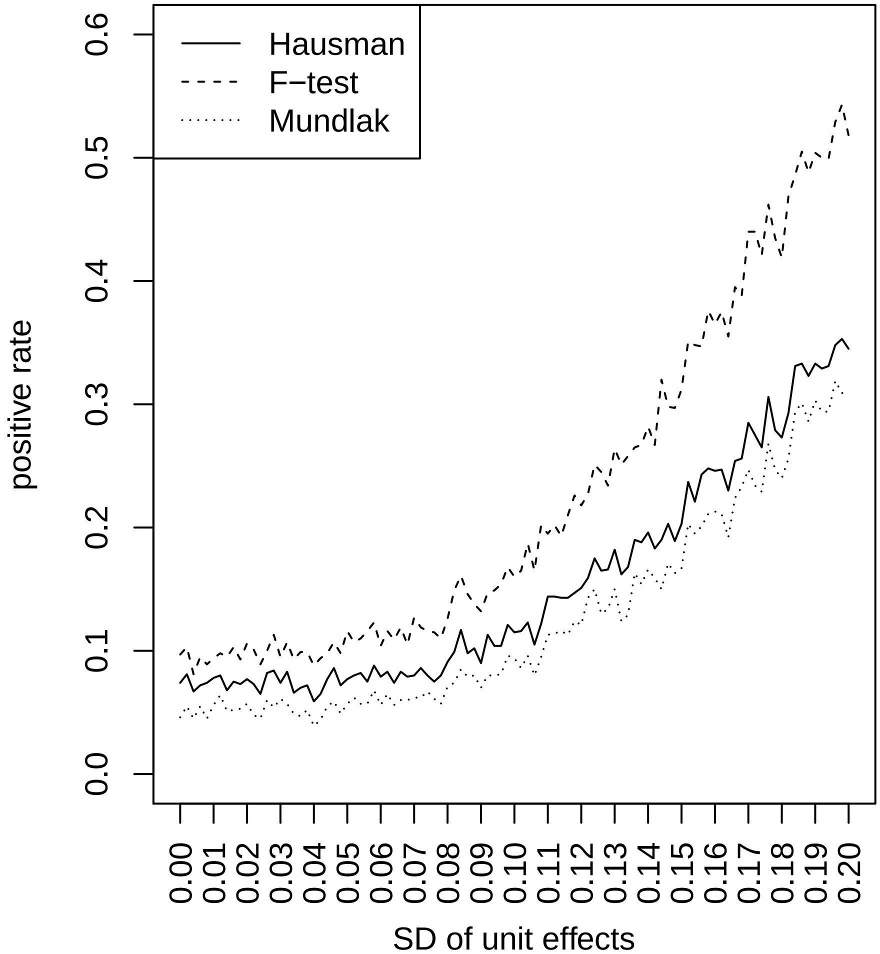


Figure 8: Test Null Rejection Rates from Cross-Sectional Data Simulations with Unit Heterogeneity, $N_g = 5$, $G = 30$