Appendix C

Ordinary Least Squares and Poisson Regression Models

by
Luc Anselin
University of Illinois
Champaign-Urbana, IL

This note provides a brief description of the statistical background, estimators and model characteristics for a regression specification, estimated by means of both Ordinary Least Squares (OLS) and Poisson regression.

Ordinary Least Squares Regression

With an assumption of normality for the regression error term, OLS also corresponds to Maximum Likelihood (ML) estimation. The note contains the statistical model and all expressions that are needed to carry out estimation and essential model diagnostics. Both concise matrix notation as well as more extensive full summation notation are employed, to provide a direct link to "loop" structures in the software code, except when full summation is too unwieldy (e.g., for matrix inverse). Some references are provided for general methodological descriptions.

Statistical Issues

The classical multivariate linear regression model stipulates a linear relationship between a *dependent* variable (also called a response variable) and a set of *explanatory* variables (also called independent variables, or covariates). The relationship is stochastic, in the sense that the model is not exact, but subject to random variation, as expressed in an *error* term (also called disturbance term).

Formally, for each observation i, the value of the dependent variable, y_i is related to a sum of K explanatory variables, x_{ih}, with h=1,...,K, each multiplied with a regression *coefficient*, β_h, and the random error term, ε_i:

$$y_i = \sum_{h=1}^{K} x_{ih} \beta_h + \varepsilon_i \qquad \text{(C-1)}$$

Typically, the first explanatory variable is set equal to one, and referred to as the *constant term*. Its coefficient is referred to as the *intercept*, the other coefficients are *slopes*. Using a constant term amounts to extracting a mean effect and is equivalent to using all variables as deviations from their mean. In practice, it is highly recommended to *always* include a constant term.

In matrix notation, which summarizes all observations, i=1,...,N, into a single compact expression, an N *by 1* vector of values for the dependent variable, y is related to an N *by* K matrix of values for the explanatory variables, X, a K *by 1* vector of regression coefficients, β, and an N *by 1* vector of random error terms, ε :

$$y = X\beta + \varepsilon \qquad \text{(C-2)}$$

This model stipulates that on average, when values are observed for the explanatory variables, X, the value for the dependent variable equals $X\beta$, or:

$$E(y \mid X) = X\beta \tag{C-3}$$

where E[|] is the conditional expectation operator. This is referred to as a specification for the conditional mean, conditional because X must be observed. It is a theoretical model, built on many assumptions. In practice, one does not know the coefficient vector, β, nor is the error term observed.

Estimation boils down to finding a "good" value for the β, with known statistical properties. The statistical properties depend on what is assumed in terms of the characteristics of the distribution of the unknown (and never observed) error term. To obtain a Maximum Likelihood estimator, the complete distribution must be specified, typically as a normal distribution, with mean zero and variance, σ^2. The mean is set to zero to avoid systematic under- or over-prediction. The variance is an unknown characteristic of the distribution that must be estimated together with the coefficients, β. The estimate for β will be referred to as b (with b_h as the estimate for the individual coefficient, β_h).

The *estimator* is the procedure followed to obtain an estimate, such as OLS, for b_{OLS}, or ML, for b_{ML}. The *residual* of the regression is the difference between the observed value and the *predicted value*, typically referred to as e. For each observation,

$$e_i = y_i - \sum_{h=1}^{K} x_{ih} \beta_h \tag{C-4}$$

or, in matrix notation, with $\hat{y}=Xb$ as short hand for the vector of predicted values,

$$e=y-\hat{y} \tag{C-5}$$

Note that the residual is *not* the same as the error term, but only serves as an estimate for the error. What is of interest is not so much the individual residuals, but the properties of the (unknown) error distribution. Within the constraints of the model assumptions, some of the characteristics of the error distribution can be estimated from the residuals, such as the error variance, σ^2, whose estimate is referred to as s^2.

Because the model has a random component, the observed y are random as well, and any "statistic" computed using these observed data will be random too. Therefore, the estimates b will have a distribution, intimately linked to the assumed distribution for the error term. When the error is taken to be normally distributed, the regression coefficient will also follow a normal distribution. Statistical inference (significance tests) can be carried out once the characteristics (parameters) of that distribution have been obtained (they are never known, but must be estimated from the data as well). An important result is that OLS is *unbiased*. In other words, the mean of the distribution of the estimate b is β, the true, but unknown, coefficient, such that "on average," the estimation is on target. Also, the variance of the distribution of b is directly related to the variance of the error term (and the values for the X). It can be computed by replacing σ^2 by its estimate, s^2.

An extensive discussion of the linear regression model can be found in most texts on linear modeling, multivariate statistics, or econometrics, for example, Rao (1973), Greene (2000), or Wooldridge (2002).

Ordinary Least Squares Estimator

In its most basic form, OLS is simply a fitting mechanism, based on minimizing the sum of squared residuals or residual sum of squares (RSS). Formally, b_{OLS} is the vector of parameter values that minimizes

$$RSS = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \sum_{h=1}^{K} x_{ih} b_h)^2 \quad \text{(C-6)}$$

or, in matrix notation,

$$RSS = e'e = (y-Xb)'(y-Xb) \quad \text{(C-7)}$$

The solution to this minimization problem is given by the so-called *normal equations*, a system of *K* equations of the form:

$$\sum_{i=1}^{N} (y_i - \sum_{h=1}^{K} x_{ih} b_h) x_{ih} = 0 \quad \text{(C-8)}$$

for *h=1 to K*, or, in matrix notation,

$$X'(y-Xb) = 0 \quad \text{(C-9)}$$
$$X'Xb = X'y \quad \text{(C-10)}$$

The solution to this system of equations yields the familiar matrix expression for b_{OLS}:

$$b_{OLS} = (X' X)^{-1} X'y \quad \text{(C-11)}$$

An estimate for the error variance follows as

$$s^2_{OLS} = \sum_{i=1}^{N} (y_i - \sum_{h=1}^{K} x_{ih} b_{OLS,h})^2 / (N - K) \quad \text{(C-12)}$$

or, in matrix notation,

$$s^2_{OLS} = e'e / (N-K) \quad \text{(C-13)}$$

It can be shown that when the *X* are *exogenous*[1] only the assumption that E[ε]=0 is needed

[1] In practice, this means that each explanatory variable must be uncorrelated with the error term. The easiest way to ensure this is to assume that the *X* are fixed. But even when they are not, this property holds, as long as the randomness in *X* and ε are not related. In other words, knowing something about the value of an explanatory variable should *not* provide any information about the error term. Formally, this means that *X* and ε must be orthogonal, or E[*X*'ε]=0. Failure of this assumption will lead to so-called simultaneous equation bias.

to show that the OLS estimator is *unbiased*. With the additional assumption of a fixed error variance s^2, OLS is also most *efficient*, in the sense of having the smallest variance among all other linear and unbiased estimators. This is referred to as the BLUE (Best Linear Unbiased Estimator) property of OLS. Note, that in order to obtain these properties, no additional assumptions need to be made about the distribution of the error term. However, to carry out statistical inference, such as significance tests, this is insufficient, and further characteristics of the error distribution need to be specified (such as assuming a normal distribution), or asymptotic assumptions need to be invoked in the form of laws of large numbers (typically yielding a normal distribution).

Maximum Likelihood Estimator

When the error terms are assumed to be independently distributed as normal random variables, OLS turns out to be equivalent to ML.

Maximum Likelihood estimation proceeds as follows. First, consider the density for a single error term:

$$\varepsilon \sim N(0, \sigma^2), \text{ or} \tag{C-14}$$

$$f([\varepsilon]_i|s^2) = \frac{1}{\sqrt{2\Pi\sigma^2}} e^{-(1/2)(\varepsilon_i^2/\sigma^2)} \tag{C-15}$$

A subtle, but important, point is that the error itself is not observed, but only the "data" (*y* and *X*) are. We move from a model for the error, expressed in unobservables, to a model that contains observables and the regression parameter by means of a standard "transformation of random variables" procedure. Since y_i is a linear function of ε it will also be normally distributed. Its density is obtained as the product of the density of ε and the "Jacobian" of the transformation, using $\varepsilon_i = y_i - x_i\beta$ (with x_i as the *i*-th row in the X matrix). As it turns out, the Jacobian is one, so that

$$f([y_i|\beta]_i|s^2) = \frac{1}{\sqrt{2\Pi\sigma^2}} e^{-(1/2)((y_i - x_i\beta)^2/\sigma^2)} \tag{C-16}$$

The likelihood function is the joint density of all the observations, given a value for the parameters β and σ^2. Since independence is assumed, this is simply the product of the individual densities from equation C-16. The log-likelihood is then the log of this product, or the sum of the logs of the individual densities. The contribution to the log likelihood of each observation follows from equation C-16:

$$\begin{aligned} &\text{Log } f(y_i|\beta, \sigma^2) = \\ &L_i = -(1/2)\log(2\pi) - (1/2)\log(\sigma^2) - (1/2)[(y_i - x_i\beta)^2/\sigma^2] \end{aligned} \tag{C-17}$$

The full log-likelihood follows as:

$$L = \sum_{i=1}^{N} L_i = -(N/2)\log(2\pi) - (N/2)\log(\sigma^2) - (1/2\sigma^2) \sum_{i=1}^{N} (y_i - x_i\beta)^2 \tag{C-18}$$

or, in matrix notation,

$$L = -(N/2)\log(2\pi) - (N/2)\log(\sigma^2) - (1/2\sigma^2)(y - X\beta)'(y - X\beta) \qquad \text{(C-19)}$$

A Maximum Likelihood estimator for the parameters in the model finds the values for β and σ^2 that yield the highest value for equation C-19. It turns out that minimizing the residual sum of squares (or, least squares), the last term in equations C-18 and C-19, is equivalent to maximizing the log-likelihood. More formally, the solution to the maximization problem is found from the first-order conditions (setting the first partial derivatives of the log-likelihood to zero), which yield the OLS estimator for *b* and

$$s^2_{ML} = \sum_{i=1}^{N} e_i^2/N \qquad \text{(C-20)}$$

or, in matrix notation,

$$s^2_{ML} = e'e/N \qquad \text{(C-21)}$$

Inference

With estimates for the parameters in hand, the missing piece is a measure for the precision of these estimates, which can then be used in significance tests, such as t-tests and F-tests. The estimated variance-covariance matrix for the regression coefficients is

$$Var(b) = s^2(X'X)^{-1} \qquad \text{(C-22)}$$

where s^2 is either s_{OLS}^2 or s_{ML}^2. The diagonal elements of this matrix are the variance terms, and their square root the standard error. Note that the estimated variance using s_{ML}^2 will always be smaller than that based on the use of s_{OLS}^2. This may be spurious, since the ML estimates are based on asymptotic considerations (with a "conceptual" sample size approaching infinity), whereas the OLS estimates use a "degrees of freedom" (*N*-*K*) correction. In large samples, the distinction between OLS and ML disappears (for very large *N*, *N* and *N*-*K* will be very close).

Typically, interest focuses on whether a particular population coefficient (the unknown b_h) is different from zero, or, in other words, whether the matching variable contributes to the regression. Formally, this is a test on the null hypothesis that $b_h = 0$. This leads to a t test statistic as the ratio of the estimate over its standard error (the square root of the *h*,*h* element in the variance-covariance matrix), or

$$t = b_h / \sqrt{s^2(X'X)^{-1}_{hh}} \qquad \text{(C-23)}$$

This test statistic follows a Student t distribution with *N*-*K* degrees of freedom. If, according to this reference distribution, the probability that a value equal to or larger than the t-value (for a one-sided test) occurs is very small, the null hypothesis will be rejected and the

coefficient deemed "significant."[2]

Note that when s_{ML}^2 is used as the estimate for s^2, the t-test is referred to as an "asymptotic" t-test. In practice, this is a standard normal variate. Hence, instead of comparing the t test statistic to a Student t distribution, its probability should be evaluated from the standard normal density.

A second important null hypothesis pertains to all the coefficients taken together (other than the intercept). This is a test on the significance of the regression as a whole, or a test on the null hypothesis that, jointly, $b_h = 0$, for h=2,...,K (note that there are K-1 hypotheses). The F test statistic for this test is constructed by comparing the residual sum of squares (RSS) in the regression to that obtained without a model. The latter is referred to as the "constrained" (i.e., with all the β except the constant term set to zero) residual sum of squares (RSS_C). It is computed as the sum of squares of the y_i in deviation from the mean, or

$$RSS_C = \sum_{i=1}^{N} (y_i - \bar{y})^2 \qquad \text{(C-24)}$$

where $\bar{y} = \sum_{i=1}^{N} y_i/N$. The F statistic then follows as:

$$F = [(RSS_C - RSS) / (K - 1)] / [RSS / (N - K)] \qquad \text{(C-25)}$$

It is distributed as an F-variate with K-1,N-K degrees of freedom.

Model Fit

The most common measure of fit of the regression is the R^2, which is closely related to the F-test. The R^2 departs from a decomposition of the total sum of squares, or the RSS_C from equation C-C-24, into the "explained" sum of squares (the sum of squares of predicted values, in deviations from the mean), and the residual sum of squares, RSS. The R^2 is a measure of how much of this decomposition is due to the "model." It is easily computed as:[3]

$$R^2 = 1 - RSS/RSS_C \qquad \text{(C-26)}$$

In general, the model with the highest R^2 is considered best. However, this may be misleading since it is always possible to increase the R^2 by adding another explanatory variable, irrespective of whether this variable contributes "significantly." The adjusted R^2 (R_a^2) provides a better guide that compensates for "over-fitting" the data by correcting for the number of variables included in the model. It is computed by rescaling the numerator and denominator in equation C-26, as

$$R_a^2 = 1 - [RSS / (N - K)] / [RSS_C / (N - 1)] \qquad \text{(C-27)}$$

---

2 Any notion of significance is always with respect to a given p-value, or Type I error. The Type I error is the chance of making a wrong decision, i.e., of rejecting the null hypothesis when in fact it is true.

3 When the regression specification does not contain a constant term, the value obtained for the R^2 using equation (C-C-26) will be incorrect. This is because the constant term forces the residuals to have mean zero. Without a constant term, the RSS must be computed in the same waysameway as in equation (C-24), by subtracting the average residual $\hat{e} = \Sigma e_i/N$.

For very large data sets, this rescaling will have negligible effect and the R^2 and R_a^2 will be virtually the same.

When OLS is viewed as a ML estimator, an alternative measure of fit is the value of the maximized log-likelihood. This is obtained by substituting the estimates b_{ML} and s_{ML}^2 and into expression C-18 or C-19. With $e = y - Xb_{ML}$ as the residual vector and $s_{ML}^2 = e'e/N$, the log-likelihood can be written in a simpler form:

$$L = -(N/2)\log(2\pi) - (N/2)\log(e'e/N) - (1/2[e'e/N])(e'e) \quad \text{(C-28)}$$

$$= -(N/2)\log(2\pi) - (N/2) - (N/2)\log(e'e/N) \quad \text{(C-29)}$$

Note that the only term that changes with the model fit is the last one, the logarithm of the average residual sum of squares. Therefore, the constant part is not always reported. To retain comparability with other models (e.g., spatial regression models), it is important to be consistent in this reporting. The model with the *highest* maximized log-likelihood is considered to be best, even though the likelihood, as such, is technically not a measure of fit.

Similar to the R_a^2, there exist several corrections of the maximized log-likelihood to take into account potential over-fitting. The better-known measures are the Akaike Information Criterion (AIC) and the Schwartz Criterion (SC), familiar in the literature on Bayesian statistics. They are easily constructed from the maximized log-likelihood. They are, respectively:

$$AIC = -2L+2K, \quad \text{(C-30)}$$
$$SC = -2L+K\log(N) \quad \text{(C-31)}$$

The model with the *lowest* information criterion value is considered to be best.

Poisson Regression

Next, the Poisson regression model is examined.

Likelihood Function

In the Poisson regression model, the dependent variable for observation i (with i=1,...,N), y_i is modeled as a Poisson random variate with a mean λ_i that is specified as a function of a K by 1 (column) vector of explanatory variables x_i, and a matching vector of parameters β. The probability of observing y_i is expressed as:

$$\text{Prob}(y_i) = \frac{e^{-\lambda_i}\lambda_i y_i}{y_i!} \quad \text{(C-32)}$$

The conditional mean of y_i, given observations on x_i is specified as an exponential function of x:

$$E[y_i|x_i] = l_i = e^{x_i'b}, \quad \text{(C-33)}$$

where x_i' is a row vector. Equivalently, this is sometimes referred to as a *loglinear* model, since

$$\ln l_i = x_i'b. \qquad \text{(C-34)}$$

Note that the mean in C-33 is nonlinear, which means that the effect of a change in x_i will depend not only on β (as in the classical linear regression), but also on the value of x_i. Also, in the Poisson model, the mean equals the variance (equidispersion) so that there is no need to separately estimate the latter.

There is a fundamental difference between a classical linear regression model and the specification for the conditional mean in the Poisson regression model, in that the latter does not contain a random error term (in its "pure" form). Consequently, unlike the approach taken for the linear regression, the log-likelihood is not derived from the joint density of the random errors, but from the distribution for dependent variable itself, using C-32. Also, there is no need to estimate a residual variance, as in the classical regression model.

Assuming independence among the count variables (e.g., *excluding* spatial correlation), the log-likelihood for the Poisson regression model follows as:

$$L = \sum_{i=1}^{N} y_i x_i'b - e^{x_i'b} - \ln y_i! \qquad \text{(C-35)}$$

Note that the third term is a constant and does not change with the parameter values. Some programs may not include this term in what is reported as the log-likelihood. Also, it is not needed in a Likelihood Ratio test, since it will cancel out.

The first order conditions, $\partial L/\partial \beta = 0$, yield a system of K equations (one for each β) of the form:

$$\sum_{i=1}^{N} (y_i - e^{x_i'b}) x_i = 0 \qquad \text{(C-36)}$$

Note how this takes the usual form of an orthogonality condition between the "residuals" ($y_i - e^{x_i'b}$) and the explanatory variables, x_i. This also has the side effect that when x contains a constant term, the sum of the predicted values, $e^{x_i'b}$ equals the sum of the observed counts.[4] The system C-36 is nonlinear in β and does not have an analytical solution. It is typically solved using the Newton-Raphson method (see section).

Once the estimates of β are obtained, they can be substituted into the log-likelihood (equation C-36) to compute the value of the maximum log-likelihood. This can then be inserted in the AIC and BIC information criteria in the usual way.

Predicted Values and Residuals

The predicted value, $\hat{y}_i$, is the conditional mean or the average number of events, given the x_i. This is also denoted a λ_i and is typically not an integer number, whereas the observed value y_i

[4] A different way of stating this property is to note that the sum of the residuals equals zero. As for the classical linear regression model, this is not guaranteed without a constant term in the regression.

is a count. The use of the exponential function guarantees that the predicted value is non-negative. Specifically:

$$\hat{\lambda}_i = e^{x_i'\hat{b}} \tag{C-37}$$

The "residuals" are simply the difference between observed and predicted:

$$e_i = y_i - e^{x_i'\hat{b}} = y_i - \hat{\lambda}_i \tag{C-38}$$

Note that, unlike the case for the classical regression model, these residuals are not needed to compute estimates for error variance (since there is no error term in the model).

Estimation Steps

The well known Newton-Raphson procedure proceeds iteratively. Starting from a set of estimates $\hat{\beta}_t$ the next value is obtained as:

$$\hat{\beta}_{t+1} = \hat{\beta}_t - \hat{H}_t^{-1}\hat{g}_t \tag{C-39}$$

where $\hat{g}_t$ is the first partial derivative of the log-likelihood, evaluated at $\hat{\beta}_t$ and $\hat{H}_t$ is the Hessian, or second partial derivative, also evaluated at $\hat{\beta}_t$.

In the Poisson regression model,

$$g = \sum_{i=1}^{N} x_i(y_i - \hat{\lambda}_i) \tag{C-40}$$

$$H = -\sum_{i=1}^{N} \hat{\lambda}_i x_i x_i' \tag{C-41}$$

In practice, one can proceed along the following lines.

1. **Set initial values for parameters**, say $b_0[h]$, for h=1,...,K. One can set $b_0[1]=\bar{y}$, the overall average count as the constant term, and the other $b_0[h]=0$, for h=2,...,K.
2. **Compute predicted values** for each i, the value of $\hat{\lambda}_i = e^{x_i'b0}$.
3. **Compute gradient**, g, using the starting values. Note that $g[h]$ is a K by 1 vector. Each element of this vector is the difference between:

$$O_i = \sum_{i=1}^{N} x_{ih} y_i \tag{C-42}$$

$$N$$

$$P_i = \sum_{i=1} x_{ih}\lambda_i \qquad \text{(C-43)}$$

$$g_i = O_i - P_i \qquad \text{(C-44)}$$

Note that C-42 does not contain any unknown parameters and needs only to be computed once (provided there is sufficient storage). As the Newton-Raphson iterations proceed, the values of *g* will become very small.

4. **Compute the Hessian**, H, using the starting values. H is a *K* by *K* matrix (C-41) that needs to be inverted at each iteration in C-39. It is *not* the *X'X* of the classical model, but rather more like *X' Σ X*, where *Σ* is a diagonal matrix. One way to implement this is to multiply each row of the *X* matrix by $\sqrt{\hat{\lambda}_i}$, e.g., *xs*[*i*][*h*]=*x*[*i*][*h*]*sqrt($\hat{\lambda}$[*i*]), where *xs* is the new matrix (X^*), *i* is the observation (row) and *h* the column of *X*. The Hessian then becomes the cross product of the new matrices, or, $H=X^{*'} X^*$. This needs to be done at each iteration. There is no need to take a negative since the negative in C-41 and in C-39 cancel.

5. **Update the estimate** for the *b*[*h*], say b_1[h] is obtained using the updating equation C-39 except that the product $H^{-1}g$ is added to the initial value. In general, for iteration *t*, the new estimates are obtained as b_{t+1}. After checking for convergence, the old b_t is set to b_{t+1} and inserted in the computation of the predicted values, in step 2 above.

6. **Convergence**. Stop the iterations when the difference between b_{t+1} and b_t becomes below some tolerance level. A commonly used criterion is the norm of the difference vector, or $\Sigma_h (b_{t+1}[h] - b_t[h])^2$. When the norm is below a preset level, stop the iterations and report the last b_t as the result. The reason for not using b_{t+1} is that the latter would require an extra computation of the Hessian needed for inference.

Inference

The asymptotic variance matrix is the inverse Hessian obtained at the last iteration (i.e., using b_t). The variance of the estimates are the diagonal elements, the standard errors their square roots. The asymptotic t-test is constructed in the usual way, as the ratio of the estimate over its standard error. The only difference with the classic linear regression case is that the p-values must be looked up in a standard normal distribution, not a Student t distribution.

Likelihood Ratio Test

A simple test on the overall fit of the model, as an analogue to the F-test in the classical regression model is a Likelihood Ratio test on the "slopes". The model with only the intercept is nothing but the mean of the counts, or

$$\lambda_I = \bar{y} \,\forall \qquad \text{(C-45)}$$

with $\bar{y}=\sum_{i=1}^{N} y_i/N$.

The corresponding log-likelihood is:

$$L_R=-N\bar{y}+\ln(\bar{y})\left(\sum_{i=1}^{N} y_i\right)-\sum_{i=1}^{N}\ln y_i!, \qquad \text{(C-46)}$$

where the *R* stands for the "restricted" model, as opposed to the "unrestricted" model with *K-1* slope parameters. The last term in C-46 can be dropped, as long as it is also dropped in the calculation of the maximized likelihood (C-35) for the unrestricted model (L_U), using $\lambda_i=e^{x_i'\beta_t}$. The Likelihood Ratio test is then:

$$LR=2(L_U-L_R), \qquad \text{(C-47)}$$

and follows a χ^2 distribution with *K*-1 degrees of freedom.

References

Gentle, J. E. (1998). *Numerical Linear Algebra for Applications in Statistics*. Springer-Verlag, New York, NY.

Gentleman, W. M. (1974). Algorithm AS 75: Basic procedures for large, sparse or weighted linear least problems. *Applied Statistics*, 23: 448–454.

Greene, W. H. (2000). *Econometric Analysis, 4th Ed.* Prentice Hall, Upper Saddle River, NJ.

Miller, A. J. (1992). Algorithm AS 274: Least squares routines to supplement those of Gentleman. *Applied Statistics*, 41: 458–478.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C. The Art of Computing (Second Edition)*. Cambridge University Press, Cambridge, UK.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. Wiley, New York, 2nd edition.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.