

REGRESSION WITH PANEL DATA

The following notes are designed for multiple regressions in which observations on a *set* of sample units are made for many time periods. A good illustration is provided by a set of sports teams that are being studied over a number of consecutive seasons. If Y_{it} denotes, say the percent of games won by team i during season t , then one might want to explain these winning percentages in terms of a number of attributes ($x_{(1)it}, \dots, x_{(k)it}$) of teams in each of these seasons. Here one might consider a regression of the form:

$$(1) \quad Y_{it} = \beta_0 + \sum_{j=1}^k x_{(j)it} \beta_j + \varepsilon_{it}, \quad i = 1, \dots, n, t = 1, \dots, T$$

Because this is time series data, it is of course desirable check for autocorrelation. But there is a difficulty here since the full data set is not strictly ordered by time (i.e., in each time period one has samples for all teams). However, there are two approaches that one can take.

First if there are only a few time periods, say $T \leq 5$, then there is too little data to obtain reliable estimates of autocorrelation. Hence the best way to proceed in this case is to simply treat time periods as a standard *categorical variable*. Here one can use time period $t = 1$ as the “benchmark” period and introduce $T - 1$ dummy variables, ($\delta_s : s = 2, \dots, T$), one for each of the remaining time periods. Model (1) will then take the form:

$$(2) \quad Y_{it} = \beta_0 + \sum_{j=1}^k x_{(j)it} \beta_j + \sum_{s=2}^T \delta_s(it) \alpha_s + \varepsilon_{it}, \quad i = 1, \dots, n, t = 1, \dots, T$$

where $\delta_s(it) = 1$ if $t = s$ and $\delta_s(it) = 0$ otherwise (and where α_s is the slope parameter for δ_s). This is exactly the same as any other categorical variable, and will help to distinguish time periods by allowing a different intercept for each period. If there is a belief that some of the x -variables may have different slopes in some time periods, then one can also introduce interaction terms of the form ($x_j \cdot \delta_s$). Here the implicit assumption is that different time periods affect only the *conditional mean* of Y , and not the correlations among unobserved residuals.

But if there are sufficient time periods to allow meaningful estimation of autocorrelation effects, say $T > 5$, then one should attempt to do so. Unfortunately JMP provides no explicit procedure for doing so (unlike JMP’s “parent” software, SAS). But it is still possible to approximate the simplest of these standard procedures in the following way (which assumes that you have already read the Autocorrelation Notes on the web).

Step 0. Order your data in JMP by units, $i = 1, \dots, n$, i.e.,

Y	$X_{(1)}$	$\dots \dots$	$X_{(k)}$
y_{11}	$x_{(1)11}$	$\dots \dots$	$x_{(k)11}$
\vdots	\vdots		\vdots
y_{1T}	$x_{(1)1T}$	$\dots \dots$	$x_{(k)1T}$
y_{21}	$x_{(1)21}$	$\dots \dots$	$x_{(k)21}$
\vdots	\vdots		\vdots
y_{2T}	$x_{(1)2T}$	$\dots \dots$	$x_{(k)2T}$
\vdots	\vdots		\vdots
y_{n1}	$x_{(1)n1}$	$\dots \dots$	$x_{(k)n1}$
\vdots	\vdots		\vdots
y_{nT}	$x_{(1)nT}$	$\dots \dots$	$x_{(k)nT}$

Notice that by doing so, the only rows that do not follow the previous row in time are the first rows of each sample unit (team). But since the Durbin-Watson statistic and ρ estimate only involve *pairs* of time elements, $(t-1, t)$, almost all row pairs can thus serve to estimate these quantities. Hence the present “approximation” idea is simply to treat this entire data set of $(n \times T) - 1$ row pairs as being ordered in time (and implicitly ignore the effect of the n row-pairs which are not).

Step 1. Proceed with Step 1 of the standard procedure on page 12 of the Autocorrelation notes. If the p-value of the Durbin-Watson test is above .05, then the autocorrelation problems can safely be ignored. If not then continue to Step 2 below.

Step 2. Again this is the same as Step 2 on page 12. Set $\hat{\rho}$ equal to the *Autocorrelation value* reported in JMP.

Step 3. To carry out a two-stage regression using this estimate, $\hat{\rho}$, the procedure is somewhat more complex. Here the *transformed variables* are of the form:

$$(3) \quad \hat{z}_{it} = y_{it} - \hat{\rho} y_{i,t-1} \quad , \quad t = 2, \dots, T, \quad i = 1, \dots, n$$

$$(4) \quad \hat{w}_{(j)it} = x_{(j)it} - \hat{\rho} x_{(j)i,t-1} \quad , \quad t = 2, \dots, T, \quad j = 1, \dots, k, \quad i = 1, \dots, n$$

As an illustration of how to construct this transformed data set in JMP, we consider the following simple illustration, with a single explanatory variable, x , and with parameters $\hat{\rho} = .5$, $n = 3$, and $T = 3$ (where the small values of n and T are used only for purposes of illustration).

Row	T	r	Y	X	Z	W
1	3	0.5	85	33	•	•
2	3	0.5	51	23	8.5	6.5
3	3	0.5	93	48	67.5	36.5
4	3	0.5	13	57	•	•
5	3	0.5	36	12	29.5	-16.5
6	3	0.5	45	7	27	1.0
7	3	0.5	22	55	•	•
8	3	0.5	44	87	33	59.5
9	3	0.5	37	77	15	33.5

Here the first two columns, **T** and **r**, denote the number of time periods, T , and the value of $\hat{\rho}$, respectively, and are inserted in order to allow general expressions for the formulas in the **Z** and **W** columns discussed below. The actual data is in the **Y** and **X** columns, where for example the first three rows of column **Y** correspond to (y_{11}, y_{12}, y_{13}) and so on. The corresponding transformed variables are given respectively by **Z** and **W**. Here we focus on the transformation of **Y** into **Z**. The first row of column **Z** is blank (since there is no predecessor) and the second row is given by

$$(5) \quad y_{12} - \hat{\rho} y_{11} = 51 - (.5)85 = 8.5$$

The transformation of **X** into **W** is similar. One can of course calculate these transformed values outside JMP and insert them by hand. But a much more efficient procedure is to construct a JMP formula for carrying out this procedure. One possibility is given by the following conditional “if” statement:

$$\text{If} \left(\begin{array}{l} \text{Modulo}(\text{Row}(), \text{T}) == 1 \Rightarrow \\ \text{else} \qquad \qquad \qquad \Rightarrow \text{Y} - \text{r} * \text{Lag}(\text{Y}, 1) \end{array} \right)$$

Basically, the top line identifies the rows corresponding to the first period, $t = 1$, for each unit. In particular the **Modulo** function determines those row numbers that are equal to 1 plus some nonnegative multiple of **T** (i.e., which satisfy $\text{Row} = m * \text{T} + 1$ for some number $m = 0, 1, \dots, n - 1$). In the present example, this yields precisely rows 1, 4, and 7. These rows are then set to “blank”. In all other cases, the bottom line performs the calculation in (5). By clicking on **Formula**, you will see that functions **Row** and **Lag** are both inside the category “**Row**”, and the function **Modulo** is inside “**Numeric**”. (You can also obtain this formula directly by opening the file **Panel_Example.jmp** in the class directory and copying the formula for the **Z** column). The corresponding formula for the **W** column is obtained by simply replacing **Y** with **X** in the formula expression above.

Steps 4 and 5. Finally, using this transformed data set (including similar transformations of all other explanatory variables), Steps 4 and 5 are exactly the same as on page 12. Notice in particular that JMP will automatically discard the rows with empty values (as in rows 1, 4, and 7 above), so that no additional manipulations are required to do so.

If the results of this procedure still yield very significant autocorrelation, it is in principle possible to repeat the above procedure on the transformed data set (essentially producing “second differences” from these transformed “first differences”). However, the beta estimates obtained are usually much less reliable after two transformations.