



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

Stata tip 137: Interpreting constraints on slopes of rank-deficient design matrices

Demetris Christodoulou
University of Sydney
Sydney, Australia
demetris.christodoulou@sydney.edu.au

A rank-deficient design matrix of explanatory variables \mathbf{X} is not of full-column rank when there is one or more linear dependencies, meaning that $\mathbf{X}'\mathbf{X}$ is singular and its inverse does not exist; thus, there is no unique solution to $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Rank deficiency is sometimes referred to as “perfect collinearity”.

There are two ways to enable the use of \mathbf{X} in regression analysis. If there is one linear dependency, then the standard approach is to reduce the dimension of \mathbf{X} by identifying a zero-parameter constraint on one of its columns. This is the default treatment in Stata and other software, that is, to arbitrarily remove one column of \mathbf{X} .

The alternative approach is to expand \mathbf{X} by adding an extra column through the identification of a linear constraint across the parameters. Then, \mathbf{X} can be used in constrained least squares via the Stata command `cnsreg`. Both approaches yield the same fully identified model, but the interpretation of their estimated coefficients depends entirely on the imposed constraint.

Much of the relevant econometric literature focuses on the identification of rank-deficient matrices of mutually exclusive binary variables and the interpretation of their intercepts, a problem so ubiquitous and well understood that it has earned its colloquial moniker of “dummy variable trap”. The interpretation of constrained intercepts is indeed elementary because it is a simple matter of weighted constants.

However, as I discuss in Christodoulou (2018), a more rigorous discussion on the effect on slope coefficients of a rank-deficient \mathbf{X} seems to evade the literature. When \mathbf{b} involves slope coefficients, the reduction of \mathbf{X} by imposing zero-parameter constraints or the expansion of \mathbf{X} by imposing linear constraints amounts to an imposition of a structural relation on the parameters to be estimated. The interpretation of the constrained slopes then becomes conditional on the validity of the structural constraint.

Consider the question of how capital investment in operating assets affects sales revenue in fixed asset-intensive firms. Companies with high stakes in tangible assets rely on capital investment to boost revenue, but the more the assets are used in operations, the more their value is depleted and needs to be replenished. The economic transactions describing this relation are captured by the accounting identity

$$\text{ppe}_{it-1} + \text{cpx}_{it} - \text{dep}_{it} \equiv \text{ppe}_{it} \quad (1)$$

or equivalently stated as $\Delta \text{ppe}_{it} \equiv \text{cpx}_{it+1} - \text{dep}_{it+1}$, where ppe_t is the stock in property plant and equipment, cpx_{t+1} is new capital expenditure, and dep_{t+1} is depreciation plus other events that may deplete assets, such as the sale of assets. Naturally, increases in

capital stock are expected to bring more sales. Indeed, one could argue that the variation in the period's sales could be explained by the average capital investment used from t to $t + 1$.

The following simple simulation generates data that describe this scenario:

```
. set type double
. set seed 1234
. set obs 10000
number of observations (_N) was 0, now 10,000
. generate id = _n
. generate ppe0 = rnormal(4.5,1)
. generate cpx1 = ppe0*0.08 + rnormal(0.5,0.2)
. generate dep1 = (ppe0+cpx1)*0.06 + rnormal(0.3,0.1)
. generate ppe1 = ppe0 + cpx1 - dep1
. generate sales = 0.25 + ((ppe0+ppe1)/2)*0.1 + rnormal(0,0.1)
```

The parameters of the random normal distributions are selected so that they appear somewhat realistic, considering that these are a result of log-transformations from originally log-normally distributed variables.

Let's say that someone is interested in learning how much revenue would change if a company decides to spend more in new capital expenditure and, at the same time, how depletion would affect sales, conditional of course on the capital investment stock. Then, the variation of sales revenue could be written as a function of the structural relation of (1) plus a random-error term:

$$\text{sales}_{it} = a + b_1\text{ppe}_{it-1} + b_2\text{cpx}_{it} + b_3\text{dep}_{it} + b_4\text{ppe}_{it} + \epsilon_{it} \quad (2)$$

Given the rank deficiency in **X**, Stata will estimate this regression by imposing a zero-parameter restriction to one of the explanatory variables and also issue a warning that a variable was omitted because of perfect collinearity:

```
. regress sales ppe0 cpx1 dep1 ppe1, noheader
note: dep1 omitted because of collinearity
```

	sales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	ppe0	.0500818	.0095262	5.26	0.000	.0314086	.068755
	cpx1	.0073934	.0106667	0.69	0.488	-.0135154	.0283023
	dep1	0	(omitted)				
	ppe1	.0470199	.0100562	4.68	0.000	.0273077	.0667321
	_cons	.2559735	.0060155	42.55	0.000	.2441819	.2677651

Stata decided to drop the variable **dep1**, but this could have been another explanatory variable; for example, changing the seed to 1235 would drop **cpx1**. The interpretation of the remaining estimated slope parameters depends on the validity of the zero-parameter restriction on $b_3 = 0$ on **dep1**, a highly doubtful assumption even with real data. Let's see what happens when we estimate all competing specifications with zero-parameter constraints, that is, each time omitting one explanatory variable:

```

. quietly regress sales cpx1 dep1 ppe1
. estimates store ppe0_0
. quietly regress sales ppe0 dep1 ppe1
. estimates store cpx1_0
. quietly regress sales ppe0 cpx1 ppe1
. estimates store dep1_0
. quietly regress sales ppe0 cpx1 dep1
. estimates store ppe1_0
. estimates table ppe0_0 cpx1_0 dep1_0 ppe1_0, se(%5.4f) stats(rmse ll)

```

Variable	ppe0_0	cpx1_0	dep1_0	ppe1_0
cpx1	-.04268835 0.0055		.00739343 0.0107	.05441336 0.0051
dep1	.05008178 0.0095	.00739343 0.0107		-.04701994 0.0101
ppe1	.09710171 0.0012	.05441336 0.0051	.04701994 0.0101	
ppe0		.04268835 0.0055	.05008178 0.0095	.09710171 0.0012
_cons	.25597348 0.0060	.25597348 0.0060	.25597348 0.0060	.25597348 0.0060
rmse	.09990187	.09990187	.09990187	.09990187
ll	8848.2841	8848.2841	8848.2841	8848.2841

legend: b/se

Note how the magnitudes of the estimated slopes switch place depending on the variable that is omitted from estimation. This is because each restriction sways estimation so that the collection of all estimated slopes remains parallel to the null vector that describes the linear dependency (for an illustration, see figure 1 in Christodoulou [2018]). This sort of behavior makes any discussion on marginal effects entirely meaningless.

Such ad hoc imposed constraints, whose only purpose is to enable mere estimation, are dangerous practices when applied on rank-deficient design matrices involving slope coefficients. A zero-parameter restriction on a slope suggests a zero marginal effect, and in this case such restrictions are simply untenable.

Another way to enable estimation is to expand \mathbf{X} by imposing a linear constraint that specifies a structural relation across all parameters. For example, one could suggest that (2) behaves like a homogeneous function of some degree. The structure of the data does not allow us to estimate which degree this is, so we need to assume the degree as a constraint. We could claim that the slope coefficients add to some fixed c , thus effectively imposing a homogeneous function of degree c , meaning that a fixed change in all independent variables would change the dependent variable by that value raised to the power of c .

For example, for $c = 0$, a fixed change would result in no change in the dependent variable. For $c = 1$, a fixed change would result in a linear change in the dependent variable, or what the economists call a “constant return to scale” within the right

context; for $c < 1$, we have decreasing returns to scale, and for $c > 1$ we have increasing returns to scale. Consider the following examples:

```
. constraint define 1 ppe0 + cpx1 - dep1 - ppe1 = 0
. quietly cnsreg sales ppe0 cpx1 dep1 ppe1, collinear constraint(1)
. estimates store c0
. constraint define 1 ppe0 + cpx1 - dep1 - ppe1 = 0.75
. quietly cnsreg sales ppe0 cpx1 dep1 ppe1, collinear constraint(1)
. estimates store c0p75
. constraint define 1 ppe0 + cpx1 - dep1 - ppe1 = 1
. quietly cnsreg sales ppe0 cpx1 dep1 ppe1, collinear constraint(1)
. estimates store c1
. constraint define 1 ppe0 + cpx1 - dep1 - ppe1 = 1.25
. quietly cnsreg sales ppe0 cpx1 dep1 ppe1, collinear constraint(1)
. estimates store c1p25
. estimates table c0 c0p75 c1 c1p25, se(%5.4f) stats(rmse ll)
```

Variable	c0	c0p75	c1	c1p25
ppe0	.04746796 0.0028	.23496796 0.0028	.29746796 0.0028	.35996796 0.0028
cpx1	.00477961 0.0045	.19227961 0.0045	.25477961 0.0045	.31727961 0.0045
dep1	.00261382 0.0073	-.18488618 0.0073	-.24738618 0.0073	-.30988618 0.0073
ppe1	.04963375 0.0031	-.13786625 0.0031	-.20036625 0.0031	-.26286625 0.0031
_cons	.25597348 0.0060	.25597348 0.0060	.25597348 0.0060	.25597348 0.0060
rmse	.09990187	.09990187	.09990187	.09990187
ll	8848.2841	8848.2841	8848.2841	8848.2841

legend: b/se

The option `collinear` tells Stata to keep perfectly collinear variables, thus ensuring reporting of all estimated coefficients.

Note how the addition of all estimated slopes is always the same, at $\hat{b}_1 + \hat{b}_2 + \hat{b}_3 + \hat{b}_4 = 0.10449514$, regardless of the imposed constraint. This is the same constant to the addition of the coefficients as with the estimates with zero-parameter constraints, as above. Regardless of the constraint, the coefficients must add up to the same constant.

The coefficients are simply scaled up or down by a fixed amount as c changes. This means that because the constraints are needed for identification, the rank-deficient nature of the data does not allow one to say which structural constraint is most appropriate. One must assume it.

The model with the homogeneous function of degree zero, with $c = 0$, can also be fit using the Moore–Penrose pseudoinverse (for example, see Mazumdar, Li, and Bryce [1980]; Searle [1984]), using the `pinv()` Mata function as follows:

```
. mata:
_____ mata (type end to exit) _____
: y = st_data(.,("sales"))
: X = st_data(., ("ppe0" ,"cpx1", "dep1", "ppe1"))
: n = rows(X)
: X = X,J(n,1,1)
: XpXi = pinv(quadcross(X,X))
: b = XpXi*quadcross(X,y)
: end

. mata: transposeonly(b)
              1              2              3              4              5
1 | .0474679606   .0047796107   .0026138174   .0496337539   .2559734849
```

These are identical coefficients to those reported in the table just above under the heading *c0*, in that order. Using the Moore–Penrose pseudoinverse, we can recover every other solution that is parallel to the null vector. Given that there are four coefficients, $k = 4$, then the imposition of an assumed degree for the homogeneous function c must be equally allocated across the k coefficients. For instance, for $k = 0.75$, it holds that

```
. mata: b[1] + 0.75/4, b[2] + 0.75/4, b[3] - 0.75/4, b[4] - 0.75/4
              1              2              3              4
1 | .2349679606   .1922796107   -.1848861826   -.1378662461
```

and similarly for any other c . Similarly, because the Moore–Penrose pseudoinverse gives the solution for $c = 0$, we could use this result to see what would be the set of estimates for any given zero-parameter restriction. Here is the case of the zero-parameter restriction on the coefficient of ppe_{it-1} , which is the same as that reported in the first column of the first estimates table above:

```
. mata: b[1] - b[1], b[2] - b[1], b[3] + b[1], b[4] + b[1]
              1              2              3              4
1 | 0   -.0426883499   .050081778   .0971017145
```

Finally, an important note about standard errors—they remain the same across all specifications. As shown in Greene and Seaks (1991), the individual standard errors of regressions involving rank-deficient design matrices are no longer informative. We cannot speak of coefficient-specific statistical significance. For example, in the first table of estimates reported, the coefficient on *cpx1* appears as statistically insignificant with a p -value of 0.488. This standard-error estimate is of course nonsensical, given the nature of the simulated data. In specifications with rank-deficient design matrices, we can speak only about the fit of the overall model, as in the root mean squared error and the estimated log likelihood, which remain identical regardless of the type of constraint. There is no such thing as coefficient significance.

In Christodoulou and McLeay (2014, 2019), we use Stata to explain how this lack of insight has proven to be an acute problem in financial research that relies on inputs from the rank-deficient accounting data matrix of articulated financial statements. Accounting data, governed by a double-entry data-generating process whereby a transaction is recorded twice, is purposefully designed to be rank deficient of order one. This is a matter of structural nonidentification and requires the additional specification of a suitable constraint to enable estimation. If the constraint is arbitrarily imposed, then inference is entirely useless.

Acknowledgment

I acknowledge the useful comments by an anonymous reviewer.

References

- Christodoulou, D. 2018. The accounting identity trap: Identification under stock-and-flow rank deficiency. *Applied Economics* 50: 1413–1427. <https://doi.org/10.1080/00036846.2017.1363860>.
- Christodoulou, D., and S. McLeay. 2014. The double entry constraint, structural modeling and econometric estimation. *Contemporary Accounting Research* 31: 609–628. <https://doi.org/10.1111/1911-3846.12038>.
- . 2019. The double entry structural constraint on the econometric estimation of accounting variables. *European Journal of Finance* 25: 1919–1935. <https://doi.org/10.1080/1351847X.2019.1667847>.
- Greene, W. H., and T. G. Seaks. 1991. The restricted least squares estimator: A pedagogical note. *Review of Economics and Statistics* 73: 563–567. <https://doi.org/10.2307/2109587>.
- Mazumdar, S., C. C. Li, and G. R. Bryce. 1980. Correspondence between a linear restriction and a generalized inverse in linear model analysis. *American Statistician* 34: 103–105. <https://doi.org/10.1080/00031305.1980.10483009>.
- Searle, S. R. 1984. Restrictions and generalized inverses in linear models. *American Statistician* 38: 53–54. <https://doi.org/10.1080/00031305.1984.10482873>.