

NOTE T: SINGULARITY, MULTICOLLINEARITY, ACCURACY AND OTHER MATRIX PROBLEMS

SINGULARITY

Singularity occurs when one or more columns (or rows) of the X matrix are either identical or differ by a constant times the other column or row. Near singularity occurs when the data in one row (or column) is very close to a multiple of another row (or column).

It is not an issue when there is only one X variable. It occurs when there are two or more X variables.

As Reymont and Joreskog (1996) say, “By finding the rank of a data matrix, much can be learned about the complexity of the contained data. The rank will give the dimensionality of the matrix, which determines the number of linearly independent vectors necessary to span the space containing the vectors of the matrix, The eigenvectors and eigenvalues of a matrix will not only determine the rank, but will also yield a set of linearly independent basis vectors”. From the basis vectors, a new $X'X$ and $X'Y$ matrix can be constructed which will give the correct regression of Y on the linearly independent X variables (which may not be directly related on a 1:1 basis with the identified variables)

The only adequate way to determine singularity and near singularity is to examine the eigenvalues. Excel does not have a function that will return a vector of eigenvalues for the $X'X$ matrix. This is a real weakness in Excel. However, the interpretation of an eigenvalue vector with regard to singularity is generally beyond the scope of introductory statistics.

MATRIX SINGULARITY PROBLEMS:

EXCEL 2000

LINEST in Excel 2000 is not able to work when the input data has singularity. It will stop with meaningless values in the output. This is discussed in KBAs 209326 and 828533. Excel 2003 LINEST however is resistant to singularity problems (see KBA 828533) and when it occurs, it will automatically delete the singular columns from the analysis and reset the degrees of freedom to reflect this.

By doing CORREL on the data, columns that have correlation values of 0.999999 or greater, indicate a singular or near singular data input matrix. Excel 2000 matrix inversion is fairly robust to near singularity, and will calculate a linear regression with one pair of columns that correlate to 0.999999 or more. Excel 2003 LINEST will not find singularity when high correlation values occur. The relationship has to be exact (or when in the QR triangularization, the normalized diagonal value is smaller than $5E-15$.)

The method given in the next paragraph is a weak method, but it is the only way to get some rank information out of the X matrix using only Excel provided functions. Excel does not have the capabilities to fix the problem.

Checking to see if there is a singularity problem:

1. Put the X matrix on a separate worksheet.
2. Create the transpose matrix of X using the TRANSPOSE matrix function (X')
3. Form the X'X matrix by multiplying the transposed matrix times the original matrix using the MMULT matrix function (X'X).
4. Note: b and c can be done in one equation
=MMULT(TRANSPOSE(B2:G17),B2:G17).
5. Create the inverse matrix of the X'X matrix using the MINVERSE matrix function. $(X'X)^{-1}$. If Excel gives an error code, then X is definitely singular. A singular symmetrical matrix here is the same as dividing by zero.
6. Obtain the determinant of both X'X and $(X'X)^{-1}$ using the MDETERM function in a single cell. It returns a single value.
 - a. Let $a = \text{MDETERM}(X'X)$ and $b = \text{MDETERM}(((X'X)^{-1}))$.
 - b. If the inversion matrix $(X'X)^{-1}$ is correct, a should equal 1/b
 - c. The LRE value of the 1/b value with reference to the a value is an indicator of rank or singularity problems.

EXCEL 2003

The new algorithm in LINEST detects singularity, and automatically deletes the offending variable.

MULTICOLINARITY:

Co-linearity occurs when data columns are very closely correlated. That is the correlation coefficients are close to one, but not exactly one (plus or minus). If it were exactly one, it would be a case of singularity.

The inherent problem with co-linearity is that regression coefficient values become very dependent on the distribution and values of the errors. Given a reference column of data (and a value of a reference regression coefficient), by adding a new data column that is co-linear with the reference column, the resulting regression coefficients will spread away from the reference value, as co-linearity increases, with the sum of these two coefficients approaching the reference value asymptotically. The result with co-linearity is that some coefficient values that appear to be unrealistic. Co-linearity also affects the values of the other coefficients.

Co-linearity check:

1. Select an empty block of cells, m x m.
2. In the first cell enter =CORREL(B{n+5}:B{2n+5},B{n+5}:B{2n+5})
3. Formula copy across m cells.
4. Formula copy down m rows.
5. The block now contains the correlation coefficients of all variables, with a 1 down the diagonal and values above the diagonal equal to values below the diagonal.

6. Look for cells off the diagonal that have values close to 1. A value of 0.99 and larger, is considered close to 1. Excel matrix inversion is fairly robust to near singularity, and will calculate a linear regression with one pair of columns that correlate to 0.999999 or more.
7. Coefficient values for variable closely correlated (0.99 and higher) variables may be “screwy” and have high standard errors. There is nothing you can do about this except to change your model to reduce the effects.

ESTIMATING THE ACCURACY OF THE RESULTS:

COMPUTING METHODS:

In general the methods to solve linear regression problems are:

Input matrix inversion and LU decomposition

QR decompositions, using Householder transformations and plane rotations

QR Gram-Schmidt Algorithm

Classical

Modified

LINEST in Excel 2000 uses input matrix inversions and LU decomposition. Excel 2003 uses a QR decomposition using Householders transformations to get a triangular matrix for solution of coefficient values.

Matrix arithmetic theory (see Stewart 1995 and 1998) goes into the deeper mathematical aspects. It is based on the concept of matrix norms (every non-singular matrix has a norm, and there are many types of norms). The errors in regression coefficients due to the finite arithmetic used, is reflected in the concept that the actual output results are different from exact arithmetic results. This difference is unknown since exact arithmetic is not possible, but by applying theory here, the actual computer results can be expressed as being equal to exact arithmetic on an augmented matrix that differs from the true input matrix(s). Theory allows estimates to be made of the differences between the actual matrix and the augmented matrix in terms of norm values.

ESTIMATING ACCURACY (EXCEL 2000)

The basis is that the accuracy of the solution depends on the accuracy of the inversion of the $X'X$ covariance matrix. The figure of merit used is the LRE value of the determinant of the $(X'X)^{-1}(X'X)$ matrix. If the inversion is precise, this (hat) matrix is the identity matrix, and the identity matrix has a determinant of one. If the off-diagonal values are large enough, the determinant will be different from one. The LRE value of the difference from one is a measure of the accuracy of the regression results. The process of obtaining the LRE value of the determinant is given below.

1. Put the X matrix on a separate worksheet.
2. Create the transpose matrix of X using the TRANSPOSE matrix function (X')
3. Form the $X'X$ matrix by multiplying the transposed matrix times the original matrix using the MMULT matrix function ($X'X$).

4. Note: b and c can be done in one equation
=MMULT(TRANSPOSE(B2:G17),B2:G17) /goes in cells B25:G30/
5. Create the inverse matrix of the X'X matrix using the MINVERSE matrix function. $(X'X)^{-1}$. If Excel gives an error code, then X is definitely singular. A singular symmetrical matrix here is the same as dividing by zero.
6. Create the product matrix by multiplying the inverse matrix $(X'X)^{-1}$ by $(X'X)$ using the MMULT matrix function (Q).
7. Note: d and e can be done in one equation.
8. =MMULT(MINVERSE(B25:G30),B25:G30) /goes in cells B35:G40/
9. Examine Q for the magnitude of the off diagonal cells, and how close the diagonal cells are to 1.
10. If the matrix looks to be reasonably close to a unity matrix, continue and get the determinant of the Q matrix.
11. Obtain the determinant using the MDETERM function in a single cell. It works on a symmetrical matrix and returns a single value.
12. Calculate the LRE value with reference to 1.0.

There are no guidelines on what is an acceptable LRE value here. It really depends of how well the linear equation model fits the data.

This limit is not a hard limit, since larger matrices will have larger off diagonal terms. This is due to the limitations of IEEE 64 bit floating point arithmetic. The determinant of the $(X'X)^{-1} * (X'X)$ product when given an LRE calculation with reference to 1, is a crude estimate of the LRE value of the coefficients.

For example from the Longley data set: The additive was applied to the centered data.

Table N-1: Coefficient and Determinant LRE Values

Variable	Original	Centered	1000 Additive	100,000 Additive	10,000,000 Additive
Intercept	8.26	12.41	-	-	-
GNP Deflator	7.57	12.80	8.03	3.76	0
Gross National Product	7.78	13.36	9.24	4.13	0.68
Unemployment	8.36	13.61	9.61	4.75	1.40
Military Employment	8.60	12.49	9.83	5.01	1.69
Population	7.40	13.17	8.86	3.56	0
Year	8.27	13.17	9.26	4.87	1.76
Determinant	10.83	13.55	9.62	5.12	3.60

The determinant LRE value runs somewhat higher than the LRE values of the coefficients. For a rule-of-thumb, it would suggest that if the LRE value of the

determinant is less than 5, the coefficient vector should be considered inaccurate and not be used.

ESTIMATING ACCURACY (EXCEL 2003)

Under Excel 2003, there is no way to estimate the accuracy of the results from LINEST. Stewart making some general observations says that generally the QR methods tend to give more accurate results but the difference from normal equations (the edge) is small. However the QR methods are much more stable. There are conditions where the normal equations are not even positive definite. (Stewart, 1995, p 318)

Excel 2000 uses the normal equations. Excel 2003 uses a QR method. However for the example given in KBA 828533, internal values to 15 digits are given with regard to applying LINEST 2003 to a simple problem. Stewarts (1995) QR algorithm 1.11 gives identical internal values.