PROJECT TEMPLATE ON LOSS DEVELOPMENT: MULTICOLLINEARITY

(The attached PDF file has better formatting.)

The calendar year, accident year, and development year are correlated.

Jacob: Why are they correlated? Exposure growth, inflation, and loss payment pattern are distinct (unrelated) phenomena. The exposure growth is in *deflated* (inflation-adjusted) dollars, and the loss payment pattern is also in *deflated* dollars.

Rachel: Exposure growth, inflation, and loss payment pattern are distinct. But calendar year, accident year, and development year are correlated: $AY_k + DY_j = CY_{j+k}$

Intuition: If we know the row (accident year) and the column (development year), we know the exact cell, so we know the diagonal (calendar year).

Jacob: Is this perfect multicollinearity?

Rachel: If the accident year, development year, and calendar year ß's are constant, this is perfect multicollinearity, and the solution to the regression analysis is not unique. We eliminate one of the independent variables before running the regression.

Jacob: Can you give a numerical example of this multicollinearity?

Rachel: You may not understand the figures yet, but they make sense once you start the project. Suppose

- The development period trend is –0.20.
- The accident year trend is +0.05.
- The payment period trend is +0.08.

Take heed: These are logarithms of trends, so they are additive. They are continuously compounded trend rates.

- Along any accident year row, the observed trend is the sum of the development period and payment period trends: –0.20 + 0.08 = –0.12.
- Down any development period column, the observed trend is the sum of the accident year and payment period trends: +0.05 + 0.08 = 0.13.

If we use the observed trends for the rows and columns, the trends are –0.12 for development period, +0.13 for accident year, and zero for calendar year. We get the correct combined trend for the cells.

- If we move down a diagonal from the *upper right to the lower left*, the payment year does not change: the trend is the *negative* of the development period trend plus the

accident year trend, or the accident year trend *minus* the development period trend: –(–0.20) + 0.05 = +0.25 = –(–0.12) + (0.13)

- If we move down a diagonal from the *upper left to the lower right*, the accident year and the development period both increase by one and the payment year increases by two: the trend is the negative of the development period trend plus the accident year trend plus twice the payment period trend: –0.20 + 0.05 + 2 × 0.08 = +0.01 = –0.12 + 0.13 + 2 × 0 = +0.01.

Jacob: How do we solve this problem?

Rachel: We discard the accident year trend and use a measure of business volume by accident year.

- This measure is in real terms, such as policies issued, lives insured, cars covered, or number of claims.
- If the measure is in nominal dollars, such as premium volume, we divide by an inflation index.

We adjust the paid loss dollars to base measure of volume.

Illustration: Suppose an insurer covers 10,000 cars in the base accident year, 20X0; 11,000 cars in 20X1, and 10,600 cars in 20X2. The incremental paid loss dollars in the first two development periods are

	0 to 1 Year	*1 to 2 Years*
20X0	100,000	80,000
20X1	105,000	86,000
20X2	110,000	90,000

We do not estimate an accident year trend. Instead, we divide the 20X1 and 20X2 paid loss dollars by the change in the number of cars:

	0 to 1 Year	*1 to 2 Years*
20X0	100,000	80,000
20X1	95,455	78,182
20X2	103,774	84,906

Illustration: $105,000 / (11,000 / 10,000) = $95,455.

Jacob: Do the illustrative worksheets show the measure of volume?

Rachel: This accident year adjustment does not use regression techniques. The illustrative worksheets begin after this adjustment. They assume the same volume of business in each accident year.

- We have no accident year variable; just development year and calendar year.
- This eliminates the perfect multicollinearity, making the regression analysis easier.
- We have just 50% multicollinearity between development year and calendar year.

{*Note:* The Equifas software uses varying β's and all three independent variables.}

Jacob: If we have the same measure of business volume, such as policies or cars insured, are the expected paid losses the same in each accident year?

Rachel: We still have an inflation trend. If inflation is 10% per annum, the expected paid losses in 20X1 are 10% greater than the paid losses in 20X0.

Jacob: If we have the same measure of business volume and we deflate the dollars of loss, are the paid losses the same in each accident year?

Rachel: The *expected* paid losses are the same; the actual paid losses have a random error term. The regression analysis forecasts the expected paid losses.

{Summary: The simulated data in the project templates are logarithms of incremental paid losses. No adjustments are needed. To apply this reserving method to an insurer's actual paid losses, adjust the data three ways: cumulative to incremental, dollars to logarithms, and divide by an exposure index.

Jacob: What order do we use for these three adjustments?

Rachel: Any order is fine. The simplest order is

- convert cumulative to incremental values by taking first differences
- divide by the measure of volume to eliminate the accident year dimension
- take logarithms to convert a multiplicative model to an additive model }

Jacob: What data do we simulate for the student project? If we use actual insurance data with the three adjustments, what do we have?

Rachel: We simulate a table of two dimensions and ½ × N × (N+1) data points. N = 10 gives 55 data points. The two dimensions are not rows and columns.

- One dimension is the development year, which is the column in the table.
- The other dimension is the calendar year, which is the diagonal in the table.

From these 55 data points we form a matrix of 55 rows by 3 columns.

- The first two columns are the calendar year and development year.
- The third column is the logarithm of the incremental paid losses.

We form a multiple regression model and derive ordinary least squares estimators for α, β_1, and β_2. We forecast the remaining 45 data points: the cells in the lower right triangle, or calendar years 10 through 18.

Take heed: The illustrative worksheets use several more columns to simulate a random draw from a normal distribution.

- The separate columns are easier for new users to understand.
- Experienced Excel users can combine the random draw with the observed value of Y.

{The following section was written by a second course instructor. It reviews some of the items above and adds another perspective.}

MULTICOLLINEARITY

Jacob: The development period affects the calendar year. The calendar year is a diagonal, so it equals the column plus the row (for an origin of zero) or the column + the row – 1 (for an origin of 1). Does this cause a problem?

Rachel: This causes multicollinearity.

- With *three* independent variables of development period, accident year, and calendar year, we have perfect collinearity.
- If we use just the development period and the calendar year, we have a positive but not perfect correlation.

We use the variances of the ordinary least squares estimators to verify that results are reasonable. To determine the correlation of the two independent variables, we work out the standard deviations of X_1 and X_2 and their covariance.

Jacob: Do we use the formulas in the textbook to adjust the standard errors of the ordinary least squares estimators for multicollinearity?

Rachel: The adjustment is done by the *REGRESSION* add-in. We explain the effects here so you can compare Excel's results with the formulas in the textbook.

Jacob: To use the formula in the textbook, we need the correlation of the independent variables. Can we use Excel to work out the correlation?

Rachel: We can use the Excel CORREL built-in function, or we can use intuition. We know that calendar year = accident year + development period.

- The correlation of calendar year with accident year + development period is +100%.
- Accident year and development period are symmetrically related to the calendar year.
- ⇛ The correlation of calendar year with accident year or development period is +50%.

Jacob: How does multicollinearity affect the regression? Even with multicollinearity, the ordinary least squares estimators are unbiased. Why be concerned with multicollinearity?

Rachel: Multicollinearity increases the variances of the ordinary least squares estimators.

- We estimate the geometric decay and inflation rate from the observed data.
- Multicollinearity makes the estimators less accurate.

Jacob: How large is this effect?

Rachel: We divide the variance of the β parameters by $(1 - r^2) = 1 - 50\%^2 = 0.75$; this raises the variance by 33.3%. The multicollinearity does not invalidate the regression, but it makes the ordinary least squares estimators less efficient.

{Jacob's question below is asked by almost every candidate. It takes a while to distinguish the index of a dimension with the trend of that dimension.

- The index along rows is development period; the trend is loss payment decay.
- The index along diagonal is calendar year; the trend is inflation.

Each index varies from 0 to Z–1. The regression uses the cumulative trend:

- Along rows: index × trend = cumulative loss payment decay.
- Along diagonals: index × trend = cumulative inflation.

In a loss triangle, rows and diagonals have a 50% correlation.}

Jacob: The estimators are geometric payment decay and inflation.

- The geometric decay depends on the loss payment pattern in real dollars.
- The inflation rate depends on the price index.

These estimators are not correlated. Why is there multicollinearity?

Rachel: Multicollinearity refers to the independent variables, not to the coefficients.

~ The independent variables are calendar year and development year.
~ The slope coefficients are inflation and geometric payment decay.

The calendar year = development period + accident year. These are correlated.

Jacob: For development period = 0, the calendar year may be 0 to 14. Any development period may have 15 calendar years. Why is there multicollinearity?

Rachel: Look at the cells in the *historical* loss *triangle* of *observations*.

- Low calendar years are associated with low development periods in the loss triangle. If calendar year = 0, the development period is 0; if calendar year = 1, the development period is either 0 or 1.
- High development periods are associated with high calendar years. If development period = 14, the calendar year is 14; if development period = 13, the calendar year is 13 or 14.

Jacob: This shows that X_1 (development period) and X_2 (calendar year) are positively correlated. Are the beta parameters β_1 and β_2 also positively correlated?

Rachel: The true beta parameters are scalars. The ordinary least squares estimators are random variables. You mean to ask about the estimators.

Along any accident year row, the development period and the calendar year change by the same amount from cell to cell. Increasing one regression coefficient and decreasing the other has no net effect. This causes a *negative* correlation of the estimators.

Rule: If the independent variables are positively correlated, the estimators of the beta parameters are negatively correlated.

Illustration: Suppose PL(dp, cy) is the logarithm of the paid loss by development period and calendar year, β_1 is the development period trend and β_2 is the calendar year trend. PL(0,0) = α, and PL(1,1) = α + Z. If we add k to the development period trend and subtract k from the calendar year trend, PL(0,0) = α and PL(1,1) = α + Z.

Jacob: Doesn't the regression separate the trends? Why care about multicollinearity?

Rachel: The goal of the regression analysis is to estimate the trends. The accuracy of the results depends on the stochasticity. If $\sigma = 0$, the standard error of each β parameter is zero, and the ordinary least squares estimator is the true trend. If σ is high, the standard error of each β is high (and this is compounded by the multicollinearity). The ordinary least squares estimators are unbiased, but they are inefficient.

Jacob: Should we comment on the multicollinearity in the student project?

Rachel: To test hypotheses, we use *t* values. The multicollinearity affects the variance of the ordinary least squares estimators, so it affects the *t* values.

~ For the student project, you can use the *t* values produced by Excel.
~ If you check the *t* values by hand, adjust for multicollinearity.