PROJECT TEMPLATE LOSS DEVELOPMENT: RESIDUAL PLOTS.

(The attached PDF file has better formatting.)

{Statisticians don't just formulate and test hypotheses with algebraic formulas. They devote much effort to interpreting graphs: scatter plot, time series, residuals, and correlations. The SOA views the student projects as more than book-learning. Candidates learn to format and interpret statistical charts.}

Jacob: The course modules do not cover residual plots. How do we form them? Are they part of Excel's *REGRESSION* add-in?

Rachel: Select RESIDUAL PLOTS on the REGRESSION menu in the DATA ANALYSIS add-in.

- The Excel plots show the residuals by each dimension (development period or calendar year), not their averages or variances.
- Other statistical packages show averages and variances, but we use Excel for the student projects since most candidates are familiar with it.

Your student project analyzes averages and standard deviations in the residual plots. You format your own plots with the needed information. Compare your plots with the output of the *REGRESSION* add-in to verify your results.

Jacob: The residual plots are too small to read clearly.

Rachel: You can enlarge the residual plot by dragging one of the corner points.

Jacob: What does the *residual output* show? (The table of residuals, not the plot.)

Take heed: You use the residual output to format your own residual plots. Understand the definition of the residual so that you can interpret the slope of the residual plot.

Rachel: Excel's *RESIDUAL OUTPUT* shows the fitted value, the residual, and the standardized residual for each observation (if you request these on the regression screen). If you place the output of the REGRESSION add-in on a new worksheet, this table appears in columns A through D starting on row 26. (Depending on your version of Excel, the output may appear in other columns or rows.) If you place the output on the same spreadsheet, the rows and columns depend on your choice of the upper left cell. Below are three rows from this table.

Observation	*Predicted Y*	*Residuals*	*Standard Residuals*
1	10.00346	0.06873	0.71250
2	9.90262	0.09186	0.95232
3	9.80178	-0.01594	-0.16522

We simulated with $\alpha = 10$, $\beta_1 = -25\%$, $\beta_2 = 15\%$, and $\sigma = 0.10$.

Jacob: This output gives $\alpha = 10.00346$ and slightly different ß's.

Rachel: These are the ordinary least squares estimators; they are close to the simulation parameters but not exactly the same, since $\sigma > 0$.

In this table, two thirds of the residuals are less than 0.10 (in absolute value), since $\sigma = 0.10$. The residuals are zero if $\sigma = 0$. To see the effect of stochasticity, choose $\sigma = 0.01$, 0.1, and 1.0 in three simulations. The ordinary least squares estimators are further away (on average) from the true coefficients as σ increases, and the residuals are larger.

Jacob: We simulated with $\alpha = 10$. The first observation has $X_1 = 0$ and $X_2 = 0$. The predicted Y is 10.00346. Shouldn't the residual by –0.00346? Why is the residual 0.06873?

Rachel: The residual is the observed Y minus the predicted Y. The observed Y for the first observation is 10.07219. The residual is 10.07219 – 10.00346 = 0.06873.

Jacob: So the residual is the error term? Or should we say the realization of the error term?

Rachel: Not exactly. The error term for the first observation is 10.07219 – 10 = 0.07219. The regression analysis does not give the exact α, so the residual differs slightly from the error term.

Jacob: Do we use residuals or standardized residuals?

Rachel: Use either residuals or standardized residuals to test parameter stability; use standardized residuals to test heteroscedasticity. The illustrative workbook uses residuals.

Jacob: We simulated with $\sigma = 0.10$. If the residual is 0.06873, shouldn't the standardized residual be 0.06873 / 0.10 = 0.68730?

Rachel: The standardized residual uses the ordinary least squares estimator for the standard deviation, which is slightly less than 0.10 (by random fluctuations).

Jacob: The residual output shows the residuals by observations. They are not related to the calendar year or development period. 15 development periods and 15 accident years give ½ × 15 × 16 = 120 observations on rows 26 through 145. They are labeled 1 through 120; the independent variables for each observation are not shown. How do we use these?

Rachel: We examine the residuals as a function of development period or calendar year.

- We need a matrix of residuals by development period and calendar year, from which we calculate means and variances and form bar and line charts.
- Once we have this matrix, we use Excel's built-in functions for means and variances, and the chart wizard for graphs.

FORMATTING THE RESIDUAL OUTPUT

The residual output is not formatted ideally for the student project on parameter stability.

- The *REGRESSION* add-in does not know what hypotheses we want to test. It orders the residuals by observation, and it shows the predicted Y value, but it does not show the values of the independent variables or the actual Y value.
- We need a matrix of residuals by development period and calendar year. We convert the *table of residuals* to a *matrix* several ways.

Take heed: If you are an experienced Excel user, you may not need this guidance. If you use VBA macros, you can skip manual sorting and cell functions, and just use the macro.

(1) Manual Sorting

The observations are in accident year by development period order: the row-column order used by reserving actuaries. See the background posting on paid loss triangles.

Re-order the observations by development period or calendar year to compute means and variances for each development period or calendar year.

- You need not re-run the regression.
- Copy the values of X_1 and X_2 (from the data worksheet) into columns E and F (on the regression output worksheet).
- Sort the residual output by one of these columns.

    - X_1 and X_2 are the explanatory variables for the *REGRESSION* add-in.
    - The residual output shows the observations in the same order as the input data.
    - Your residual output now has five columns: observation, fitted Y, residual, X_1, X_2
    - Give column headers to the two explanatory variables (cal yr; dev per).

We need the mean residual for all observations with the same calendar year (or the same development period).

Select all five columns, all the data rows, and the row of headers. Copy these to a blank worksheet. Select all the cells. Click on *TOOLS* → *SORT* (pre-Excel 2007) or *DATA* → *SORT* (Excel 2007). Sort by calendar year or development period. Compute the average and the standard deviation for observations with the same calendar year or development period. Arrange these means (or standard deviations) in a column or row and form residual plots.

- If you are not familiar with Excel, this is the easiest method.
- This is a tedious method: you spend much time computing means by development period and calendar year.

Take heed: The time spent re-sorting the observations and computing means can be better spent on statistical analysis.

(2) Cell Formulas (*INDEX* and *IF* Functions)

Index Function: Compute a column of co-ordinates for the observations and use Excel's *INDEX* built-in function to form the matrix. The cell formulas are in the illustrative worksheet.

- Copy the values of X_1 and X_2 (from the data worksheet) into columns E and F (on the regression output worksheet).
- Form column G as a combination of Columns E and F.
- Format the residual matrix using Excel's INDEX function and IF function.

We show an example on the STABLE RATES worksheet. Candidates who are proficient at Excel and its *INDEX* built-in function but don't use VBA may prefer this method.

Intuition: We look up the residual based on the calendar year and development period. The Excel *VLOOKUP*, *INDEX*, and *MATCH* built-in functions look up values based on a single base, such as calendar year alone or development period alone. We combine the calendar year and development period into a single index in Column G.

Take heed: If you are taking also the time series on-line course, compare the time series project template on daily temperature, which looks up the average daily temperature by month and day. The illustrative worksheet combines month and day into a single index to the use *VLOOKUP* function.

Take heed: Excel 2007 allows more efficient use of these built-in functions. The illustrative worksheet can be used with Excel 97 or later.

Take heed: If you prefer the *VLOOKUP* built-in function, re-arrange the columns so that the M×N index column is the first column in the table.

Column G uses an MxN format: development period M and calendar year N becomes the index value MxN. This is a character string, not a formula.

The residual matrix on the illustrative worksheet is in Cells L44:Z58.

Take heed: We added blank rows to the illustrative worksheet for call-outs and comments. On your student project, the residual matrix may have L29 as the top-left cell.

The illustrative worksheet uses

- Column K as the development period index.
- Row 43 as the calendar year index.
- Column G as the development period by calendar year index.
- Column C as the residuals.

Take heed: Copy the formula in Cell L44 to your worksheet. Change the cell formulas to match the structure of your worksheet. *Illustrations:*

- If your development period index is in Column J and your calendar year index is in Row 28, replace the absolute reference $K by $J and the reference $29 by $28.
- If your residuals are in Column D and your calendar year by development period index is in Column H, replace the absolute reference $C:$C by $D:$D and the reference $G:$G by $H:$H.

Take heed: Range names make the formulas more general.

- Name the column of residuals (Column C) as *nmResiduals*.
- Name the column of indices (Column G) as *nmIndices*.

Use the names in the cell formulas (and VBA code), not the absolute references. If you use names, adding or deleting columns won't affect the formulas. You can move tables to different parts of the worksheet without affecting the formulas. Excel replaces relative cell references if you change the structure of the worksheet, but not everything comes out the way you want.

Take heed: If you use names, make them work-sheet names, not work-book names. The default is work-book names. You may use several scenarios in your student project. If you duplicate the work-sheet for another scenario,

- Workbook names refer to the original worksheet.
- Worksheet names refer to the new worksheet.

The illustrative worksheets use cell address instead of names so that new Excel users can follow the cell formulas. If you are familiar with Excel names, use them.

THE CELL FORMULAS IN THE RESIDUAL MATRIX

The residual matrix has values for the cells with observations: that is, cells whose calendar year is at least as great as the development period. The formula in each cell of the matrix is the same, after adjustment of relative references.

Take heed: You want empty cells, not zeros in the cells with no observation. The zeros distort the computed means and standard deviations.

Consider the formula for calendar year 2 and development period 1 in Cell N45:

=IF(N$43>=$K45,*INDEX*($C:$C,*MATCH*($K45&"x"&N$43,$G:$G,0)),"")

Take heed: If you use names instead of cell addresses, the formula is

=IF(N$43>=$K45,*INDEX*(nmResiduals,*MATCH*($K45&"x"&N$43,nmIndices,0)),"")

The formula says =IF(N$43>=$K45, [use this formula] ,""), which means:

If the value in Row 43 and the same column [N$43]
is at least as great as
the value in Column K and the same row [$K45]

- then use the formula beginning with the term *index*;
- otherwise, make the cell empty.

The formula has two pieces. The *MATCH* built-in function says:

Consider a string with three pieces, all of which are characters:

- the development period for this cell
- “x”
- the calendar year for this cell

The three pieces are concatenated by ampersands (&).

Match this string with the string values in Column G. The "0" at the end of the function means we need an exact match. The result is the row number for development period 1 and calendar year 2.

Illustration: For calendar year 2 and development period 1, the string is "1x2". The *MATCH* built-in function returns the row number for the string "1x2" in Column G.

The *INDEX* built-in function says: “Take the residual in Column C corresponding to the row number returned by the *MATCH* function.” This residual is placed in the residual matrix.

Illustration: For development period 1 and calendar year 2, the *MATCH* function gives Row 61. This row in Column C gives a residual of 0.027764 (rounded to 0.0278).

Take heed: The cell formulas are general: they work for a 20 × 20 loss triangle just as for a 15 × 15 loss triangle. They use absolute references for the columns of residuals and indices. You can replace the absolute references by range names to further generalize the cell formula. We have not done this in the illustrative worksheet because the range names may confuse candidates who don't use names.

Take heed: If you are familiar with these built-in functions, you can modify the cell formulas for other sizes of the paid loss triangle. If you have not used these functions before, make changes one at a time. It may take an hour to convert the cell formulas to other sizes of the paid loss triangle, but its an hour well spent.

Take heed: Traditional loss triangles are accident year by development period or accident year by calendar year. The matrix in the illustrative worksheets is development period by calendar year.

Take heed: You can use the *VLOOKUP* built-in function instead of the *MATCH* and *INDEX* built-in functions. Some candidates are more familiar with *VLOOKUP*. To use *VLOOKUP*

- Place the column of co-ordinates before the column of residuals and move cells to the right to make room for them. The *lookup value* must be the *first column* in the *lookup array*. The index built-in function does not specify the order of the columns.
- The co-ordinates are in Column C and the residuals in Column D.
- The *vlookup table* is Columns C and D. If the first row is Row 22, the last row is Row 141, and the *vlookup table* is C22:D141.
- Give the *vlookup table* a name, such as *ResidualLookup*.
- The formulas in the cells of the residual matrix are =("M×N", ResidualLookup, 2, 1). The four parameters are

1. "M×N" are the co-ordinates of the matrix cell that we look up.
2. ResidualLookup is the vlookup table.
3. The values returned are in Column 2 of the table.
4. The final parameter (1) specifies an exact match, not the closest match. If no exact match exists, the cell formula has an error. Fix the error; don't use the wrong value.

Take heed: Excel provides many ways to match values. Excel 2007 allows indexing with multiple columns, and previous versions could do the same with nested IF statements. If you are proficient with Excel, you can form the residual matrix several ways.

Take heed: You can copy the cell formulas from the illustrative worksheet. If you don't change the cell formulas, the upper left corner of the matrix should be in the same place relative to the residual output. Copy the cell formula in the upper left cell to your worksheet, and then copy this formula to the rest of your matrix.

Jacob: If I am not familiar with Excel, and I do not understand the built-in functions and macros, what should I do?

Rachel: Ideally, learn Excel. But you can complete the student project with a 15 by 15 loss triangle. Copy the illustrative worksheets and select other values for α, β_1, β_2, and σ.

Most cell formulas are already in the illustrative worksheet. Make column headers (0 to 14) and row labels (0 to 14). Specify if these are calendar years or development periods (or whatever you use). Use the design in the illustrative worksheet as your base.

Form means and standard deviations for each row and column. The number of observations differs in each row or column. If the unused cells have nothing and appear blank on the screen, Excel ignores them for the averages and standard deviations. If the unused cells have zeros, Excel includes them in the averages and standard deviations.

RESIDUAL PLOTS

Jacob: Do we do the analysis by examining the plots or the means and variances?

Rachel: We analyze parameter stability by looking at the means. We examine

- The slopes of the line segments in the residual plot.
- The change in the average residual from one calendar year (or development period) to the next.

These two items are equivalent. The graphic (chart, plot) helps you see the change. The shape of the residual plot reflects the change in the regression parameter.

Copy the residual plot to your write-up. Show the values on the residual plot or on a table, so that you can compute the slopes of the line segments. Excel charts allow you to show the values of all the points in a series or individual points in a series.

Jacob: How does σ affect the residual plot?

Rachel: If σ is low, the graph is clear.

- If the residual plot is a horizontal line, the regression coefficient is constant.
- If the residual plot is a V or a parabola, the regression coefficient is not constant.

You can simulate other shapes with two or more changes in the parameter.

Jacob: Do we check if the differences in the slopes are statistically significant? Suppose the slope is +10% if the first ten years and –8% in the last five years. Should we use an an ANOVA test to determine if the slope really changes?

Rachel: You need not do so. If you can't tell from the chart or the figures that the inflation rate is not constant, your σ is too high relative to the change in the inflation rate.

Jacob: Do we use sample variances or population variances?

Rachel: Use the sample variances. We don't know the means; we estimate them from the observations. We simulated figures, so they are a sample.

Illustration: A 15 × 15 loss triangle gives 15 residuals for calendar year = 14. Divide the sum of squares by (15 – 1) for the variance.

Jacob: When I try this, some of the means and variances are unusual; what causes this?

Rachel: We have few data points for some development periods and calendar years.

- If we have only one residual, such as for calendar year 0 or development period 14, we can't form the sample variance.
- If there are only 2, 3 or 4 residuals, such as development periods 11, 12, and 13, the estimated means and variances have large standard errors.

Every so often, a simulation gives unexpected results that don't confirm your hypothesis. The cause may be an error in your work or random fluctuation in the figures.

The quickest solution is to simulate again with a lower σ.

- If you again get strange results, you may have made an error.
- If you get the expected results, your previous result was a random fluctuation.

Jacob: If these means and variances are too volatile, how can we tell if our work is correct?

Rachel: Start with a low σ, such as 0.01, so the standard errors are low. The student project shows the dependence of statistical conclusions on the accuracy of the regression. Use a low σ to check your technique; use a realistic σ afterward.

- If you are comfortable with Excel, you can use a large loss triangle, such as 30 by 30 instead of 15 by 15. With a larger loss triangle, you can see the pattern in the residual plot even if σ is high.
- If you are not experienced with Excel, use a 15 by 15 matrix. The residual plots are harder to read, but you can check your work with the illustrative worksheets.

Take heed: The cell formulas and VBA macros work for loss triangles of any size. Using larger loss triangles should not be a problem.

- If you use cell formulas, copy them to a larger matrix.
- If you use the VBA macro in pre-2007 versions of Excel, you must enable the macros for the workbook. If you copy the illustrative worksheet, Excel will ask if you want to enable the macros. Answer *yes*.
- Excel 2007 has more stringent safeguards against macros. Some macros are viruses. Unless the macro is digitally signed or from a trusted source (or you turn off macro security), Excel 2007 won't allow you to play the macro. It won't even allow to save the macro, unless you specify that you want a macro-enabled workbook. Use the *SAVE AS* command, and specify a macro-enabled workbook.

Take heed: Be careful about downloading macros from the internet, unless you know the author. The chance that a disgruntled actuarial candidate might post a virus on the NEAS discussion forum may be one in a million, but these extreme events might happen.

{The VBA macro forms residual plots for you. If you use cell formulas or different data, you must form the residual plots manually. Even if you use residual plots from the VBA macro, form the plots manually at least one time, so that you understand what they represent.}

Jacob: Once we have the means and standard deviations, how do we form the plots?

Rachel: Select the means or standard deviations.

- If you are new to Excel, use the chart wizard.
- If you are more experienced, press F11.

{The chart wizard has been revised for Excel 2007. The chart gallery is now on the ribbon.}

Use line charts for means and column charts for standard deviations. Add data labels to specific points so you can easily compute slopes.

- To add data labels to all the points, select the data series (such as the line or the bars), right-click, and choose add data labels.
- To add data labels to one points, select the data series (such as the line or the bars), wait a second, then click on one point to select it, right-click, and choose add data label.

Take heed: If you don't wait long enough, Excel interprets the two clicks as a double click. Start again and wait several seconds before the second click.

Check if the means or variances increase or decrease. With high stochasticity, the means and variances are dispersed. Do this first with low stochasticity for each step. Once you know what to look for, use higher stochasticity.

Jacob: How do we check if we formed the residual plots correctly?

Rachel: We form residual plots so we see the slopes of the line segments and differences in the variances. The *REGRESSION* add-in also forms residual plots, but without the means and variances. They are scatter plots, not line graphs or bar graphs. Drag a corner of the residual plot formed by the *REGRESSION* add-in so you see all the points.

- Check if your means are the averages of the figures in each column.
- Check if your variances reflect the dispersion of the figures in each column.

Jacob: With low stochasticity, what should we expect?

Rachel: With low stochasticity, the ordinary least squares estimators are close to the simulation parameters.

Illustration: On one illustrative worksheet, the ordinary least squares estimators are $\alpha \approx 10$, $ß_1 \approx -0.25$, and $ß_2 \approx 0.15$. The ordinary least squares estimators are off by less than 1%, since $\sigma = 0.01$. The R^2 is almost one, and the standard errors are very low.

Verify the standard errors in your regression output by the formulas in the textbook. Since σ is small, the figures in your output should be very close to the formula values.

Summary: We form residual plots showing the average residual at each X value, for either X_1 or X_2. With low stochasticity and a stable trend, the residual plot is a horizontal line at the X axis.

EXCEL'S RESIDUAL PLOTS

Jacob: Excel forms residual plots. Why don't we use these plots instead of forming new ones, either by a VBA macro or by the chart wizard (or chart functions)?

Rachel: The residual plots from the *REGRESSION* add-in are scatter plots.

- For a student project on parameter stability (this project template), we use plots of the *mean* (average) residual by the independent variable.
- For a student project on heteroscedasticity, we use plots of the standard deviations of the residuals by the independent variable.

Compute the means or standard deviations of the residuals. Form line charts from the means and column charts from the standard deviations, not scatter plots.

- The VBA macro shows what the residual plots should look like. You can use the residual plots from the macro or you can form your own charts.
- Check the residual plots with the graphs from the *REGRESSION* add-in.

The residual plots from the *REGRESSION* add-in are small.

- Enlarge the residual plots from the *REGRESSION* add-in so you see all the residuals for a given value of the independent variable.
- From the vertical axis, compute the mean residual for these points.
- For a more exact average, place the cursor over any of the points to see a quick-tip of the exact values.

Illustration: For a calendar year index of 3, the residual plot from the *REGRESSION* add-in shows four points, such as (3, 0.65), (3, 0.15), (3, –0.25), and (3, 0.45).

- The mean residuals at CY = 3 is ¼ × (0.65 + 0.15 – 0.25 + 0.45) = 0.250.
- Your residual plot should have the point (3, 0.250).

USING RESIDUAL PLOTS

Jacob: How do we use the residual plots?

Rachel: With constant β parameters and low stochasticity, the line connecting the average residuals is flat at the horizontal axis.

Take heed: Stochasticity distorts the residual plot. If σ is high, the residual plot may be a jagged line. The standard deviation of the mean is $\sigma / \sqrt{N}$, where N is the number of observations for the mean.

Illustration: We examine the residual plot for a loss triangle of five years. Calendar years 0, 1, 2, 3, and 4 have 1, 2, 3, 4, and 5 observations. The standard deviations of the mean residuals are $\sigma/\sqrt{1}$, $\sigma/\sqrt{2}$, $\sigma/\sqrt{3}$, $\sigma/\sqrt{4}$, and $\sigma/\sqrt{5}$. We may see patterns in the residual plot stemming from random fluctuations in the values.

Change the inflation rate from a constant rate, such as 15%, to a higher figure in one set of years and a lower figure in another set of years. Examine the residual plots first with no stochasticity or little stochasticity; then use moderate stochasticity.

- With one discrete change, the residual plot is easy to read.
- With a smoothly progressing change, the residual plot shows the effect, but it is harder to read, particularly when the stochasticity is high.

The student project can use either a discrete or a continuous change. These instructions are general; we show an example below.

~ If you are familiar with Excel, use both types of changes. The spread-sheets for the two changes are similar. After simulating one change, it is easy to simulate the other.
~ If you have trouble with the simulation, residual plots, or regression analysis, focus on one of the two types of changes.

Jacob: How do we see the effect of stochasticity on the residual plot and other aspects of the regression analysis?

Rachel: We vary σ. If the inflation rate is 15% and σ is low, the estimated inflation rate might be 14% or 16%. With the same figures and a higher σ, the estimated inflation rate might be 12% or 18%. The error is magnified in the forecasts. Use a low stochasticity to make sure you understand the method; then increase the stochasticity to work the project.

Jacob: Even with high stochasticity, the estimators are unbiased. We expect the same results. How does stochasticity affect the analysis?

Rachel: With low stochasticity, the residual plots show if the parameters are stable. With high stochasticity, the residual plots are hard to read, and we can't always tell if the parameters are stable.