SIMULATION FOR STUDENT PROJECT ON LOSS DEVELOPMENT.

(The attached PDF file has better formatting.)

We simulate paid loss triangles in Excel. The on-line course covers the distribution of the error terms, but it does not teach how to simulate data.

- We provide the cell formulas to set up the Excel worksheets.
- The illustrative worksheets are fully documented. Copy them to your machine, replicate the illustration, and then modify the parameters for your student project.

Take heed: We simulate the logarithms of the paid losses, not the dollars of paid losses.

This posting gives instructions for completing the *Excel worksheets* for the student project on regression analysis for loss reserving. It explains which statistical techniques are used for each hypothesis, and how you can change the example for your student project.

Take heed: This project template provides cell formulas to form residual plots and a VBA macro that automates the process. No special Excel expertise is required.

- The instructions are detailed: call-outs, text boxes, and comments on the illustrative worksheets plus PDF files that explain the Excel techniques.
- Some instructions are for candidates with little Excel experience. If you know Excel well, you can do some steps more efficiently. We mention alternatives for many regression techniques: VBA macros, Excel built-in functions, the *REGRESSION* add-in, or cut-and-paste. Use whichever method is best for you.
- Separate postings explain different parts of the student project: background, overview, residual plots, dummy variables, etc. Some topic are explained in two places.
- The illustrative worksheets are an example; they are not *your* student project. The PDF files explain what you change for your student project.
- If you take CAS examinations, this student project is excellent exam preparation for the stochastic reserving section on Exam 6 (soon to be moved to Exam 8).

Jacob: We have illustrative worksheet and VBA macros. What are we graded on?

- Choose parameters, perform the analysis, and write up your conclusions. The student project teaches you to analyze regression analyses when the parameters vary.
- The illustrative worksheets provide cell formulas and macros. Your time is spent on interpreting the regression output and using dummy variables or other independent variables to offset a change in a regression coefficient.
- The student project is graded on your interpretation of the statistical tests and the methods used to adjust for the changes in the regression coefficients.
- The illustrative worksheets have low stochasticity (low σ) so the ordinary least squares estimators are about the same as the simulation parameters. You use a higher σ for the student project, showing that you can interpret more realistic scenarios.

- We use Excel because most candidates are familiar with it. You can use other spreadsheets or statistical software (such as SAS, Minitab, or "R") if you wish.
- You simulate data for each part of the student project. Verify the results by comparing the ordinary least squares estimators with the simulation parameters.

Jacob: The project template is complex? In what order should we proceed?

- Do each step first with no stochasticity ($\sigma = 0$) to be sure your method is correct.
- Replicate the illustrative worksheets with a low σ. If $\alpha = 10$ and $\beta_1 = -0.25$ in the illustrative worksheet, use $\alpha = 12$ and $\beta_1 = -0.35$. Then choose different parameters (such as 20 years or a smooth change in β_1) for your student project.
- The illustrative spreadsheets use variable names from a two column table. If you have not used names, read the on-line help (or any Excel manual) for CREATE NAMES and DEFINE NAMES. You don't need names to run a regression, but it makes the work clearer.
- The residual plots from the regression add-in do not have the information needed for the student project. Form residual plots with the cell formulas or VBA macros in the illustrative worksheets. Use the charting tools or the chart wizard. Label the axes so that the course instructor reviewing your project can verify what you have done.

Each step can be done alternative ways. Choose a method and its parameters.

Illustration: We simulate normal errors in three steps:

- Excel's built-in function *RAND* give random numbers in a uniform distribution on (0,1).
- Excel's built-in function *NORMSINV* converts to a standard normal distribution ($\sigma = 1$).
- Multiplication by σ gives a normal distribution with a different standard deviation.

You may use other methods:

- Combine the steps in a single cell formula: =sigma * normsinv(rand()).
- Use the *DATA ANALYSIS* add-in, which gives random numbers from various distributions.

{The project template discusses the stochasticity of the simulation. Your write-up should show that you understand how stochasticity affects the simulation and your anaysis.}

STOCHASTICITY

Background: Stochasticity refers to the value of σ. The dialogue below discusses this topic. Revise the σ parameter as you work through the student project.

Jacob: The instructions speak of low and high stochasticity. What values are low vs high?

Rachel: The low and high stochasticity depend on the values of the regression parameters.

- If $\alpha = 4$, $ß_1 = -2\%$, and $ß_2 = +1\%$, a σ of 0.05 is high stochasticity.
- If $\alpha = 20$, $ß_1 = -25\%$, and $ß_2 = +15\%$, a σ of 0.05 is moderate stochasticity.

The student project shows how the value of σ affects the regression results. The terms high vs low stochasticity are not precise.

- To test if your cell formulas are correct, use $\sigma = 0.01$.
- For actual reserve estimates, σ might be 0.25 or 0.50.

Jacob: What values of σ should we use?

Rachel: Choose $\sigma = 0$ for your first run, to make sure your method is correct. If you turn off the stochasticity ($\sigma = 0$), you can verify the regression results with simple algebra. The ordinary least squares estimators equal the simulation parameters.

Then choose $\sigma = 0.01$. The ordinary least squares estimators are about the same as the simulation assumptions, so you can verify that the regression is correct. The patterns do not change much from the no-stochasticity scenario.

- If the patterns change materially, you have probably made an error.
- Check the variance of the ordinary least squares estimators with the textbook formulas.
- The estimates should be within 2 or 3 standard deviations of the simulated parameters.
- If the discrepancy is larger, check your work.

Then choose an σ large enough for the simulation output to fluctuate moderately. Use trial and error to get a reasonable value. If you choose $\sigma = 0.01$, the data are not realistic. If you choose $\sigma = 0.50$, the stochasticity overwhelms the patterns.

The student project shows the effects of stochasticity.

- We don't specify values; you must examine the results and choose appropriately.
- Actual casualty reserves have high stochasticity that overwhelms inflation changes.
- Use a moderate σ so that the patterns in the residual plots can be seen.

Jacob: What is a reasonable value of σ for the final student project?

Rachel: Choose a value for moderately fluctuating results but enough stability to interpret the regression. A σ that adds stochasticity but does not change the pattern is sufficient.

Take heed: We do not specify a value for σ so that you try several values and see how they affect your regression analysis.

MATRIX SIZE AND HETEROSCEDASTICITY

{Loss development can be used for projects on parameter stability, heteroscedasticity, forecasting, and other topics.

- The project template gives extensive guidance and illustrative worksheets for a student project on parameter stability.
- Several postings on the discussion forum suggest other topics.

After working through the illustrative worksheets, you might design a project on another topic. Do not hesitate to develop an alternative student project.

Illustration: Several postings discuss how stochasticity may differ by development period or calendar year. The VBA macro that forms residual plots creates column charts of the standard deviations of residuals by development period and calendar year. Your student project may similar heteroscedastic data, use residual plots to identify heteroscedasticity, and then correct for heteroscedasticity using the techniques in the textbook.}

Jacob: What size triangles do the illustrative worksheets use?

Rachel: The illustrative spreadsheet uses 15 development periods and 15 accident years, for a total of 225 cells.

- The upper left triangle of observations (including the middle diagonal from the lower left corner to the upper right corner) has ½ × 15 × 16 cells = 120 cells.
- The remaining 105 cells are forecasts.

For low σ, the estimators are accurate. As σ rises, the estimators become less accurate.

Take heed: You can choose another size for your student project.

- If you use a 30 by 30 square, with four times as many observations, use lower geometric decay (such as 15%) and inflation rates (such as 10%).
- With lower inflation rates, you may have to reduce the value of σ in your simulations.

Jacob: Is there an optimal size for the paid loss triangles?

Rachel: The loss triangle is the upper left part of a square.

- Use a large enough square for the ordinary least squares estimators to be accurate.
- If the square is too large, the stochasticity and geometric decay cause problems if the data are homoscedastic.

Jacob: What do you mean by “the stochasticity and geometric decay cause problems”?

Rachel: Consider a 30 × 30 year loss triangle with ½ × 30 × 31 = 465 observations.

If α (the logarithm of incremental paid losses) = 10 in development period 0, σ = 200%, β_1 (development period trend) = –25% per annum, and β_2 (inflation) = 0%:

- The expected value is 10 – 30 × 25% = 2.5 after 30 periods.
- Since σ = 2, we get some negative logarithms and very small paid losses.

Jacob: Is this a problem with our regression analysis or is it an attribute of loss reserving?

Rachel: Classical regression analysis assumes the variance of the error terms is constant. If the Y values range widely, this assumption is not realistic. If the incremental paid losses are $10 million in the first 12 months (time 0 to time 1) and $1,000 in the last 12 months (time 29 to time 30), the standard deviation of the paid losses can not be $100,000 in both.

- If α is 15% of the expected value, it is $1.5 million in the first 12 months and $150 in the last 12 months.
- In practice, α rises with the development period. It may be 20% in the first 12 months and 75% in the last 12 months.

Classical regression analysis assumes the data are homoscedastic to simplify the work.

- If the expected Y values do not vary much, this assumption is usually reasonable.
- For loss reserving, the expected values decline to zero as development progresses. The data are always heteroscedastic.

Jacob: Does this create a problem for the student project?

Rachel: On the contrary; this creates opportunity for student projects on heteroscedasticity.

- For the project template on parameter stability, use values for α, β_1, β_2, and σ so that the standard deviation is never too large a percentage of the expected value.
- For a student project on heteroscedasticity, you may use several assumptions:

Take heed: Each assumption below is important for accurate loss reserving. Any one of these assumptions can form a student project on heteroscedasticity. All three assumptions are true, but don't design a project that is too complex to complete.

- The standard deviation of the error term is proportional to the price level. As inflation raises the price level 10%, the standard deviation of the error term rises 10%.
- The variance of the error term is proportional to exposures. As development reduces the loss reserves 20%, the variance of the error term falls 20%.
- The constant of proportionality rises with the development period. Claims that settle slowly have more fluctuation in settlement values.

All three relations affect the standard error. Examine each assumption separately. You may pick one assumption, see its effects on residual plots, and modify the regression analysis to increase its efficiency.

SIMULATION IN EXCEL

{Some candidates have never before used simulation and prefer to use actual data. But a statistics course focuses on uncertainty (stochasticity). The student project teaches you

- How to simulate data in Excel.
- How the stochasticity inherent in the simulation affects the regression results.}

Jacob: Why do we simulate values? Why not use given data?

Rachel: We simulate for several reasons:

- If we specify data for the project template, candidates have the same figures and we can't ensure independent work. By simulating, all candidates have different data. Even if candidates use the same techniques, the data differ and the results differ. Candidates can even discuss their work with each other.
- By simulating, you can verify the work in each step. Use a σ of zero to check the procedure, a low σ (such as 0.01) to verify that the technique gives the expected result, and a moderate σ for the student project itself.
- Simulating show the relation of σ to estimates and forecasts. Even unbiased estimators may not be useful if they are inefficient.

Jacob: Why not use just the moderate σ?

Rachel: Starting with zero or low stochasticity saves time. If we start with moderate σ, we can't tell if the results differ from the expected because of random fluctuations or errors.

Jacob: With good instructions we don't expect to make errors.

Rachel: Everyone makes errors. With high stochasticity, you won't see the error at first and you waste time. With zero stochasticity, you can check the results and find the errors.

If you don't get results you expect, turn off the stochasticity and run the simulation again.

SIMULATION

Jacob: Do we need simulation software?

Rachel: We use Excel functions; no simulation software is needed.

Jacob: How do we simulate values in Excel?

Rachel: The simulation has two parts: the expected values and the stochasticity.

Expected values: Choose a value for α, such as 10. With zero-based arrays, α is the logarithm of the paid losses when development period = 0 and calendar year = 0.

- The logarithm of the paid loss in other cells is $\alpha + ß_1 \times X_1 + ß_2 \times X_2$.
- Form a triangle with these figures, using the assumed β parameters.

With a σ of zero, the regression analysis has no uncertainty. Use this scenario to make sure the computations (the Excel formulas) are correct.

Jacob: Where is this on the illustrative worksheet?

Rachel: Look at the spreadsheets for STABLE RATES and UNSTABLE RATES. Column C shows the development period and Column D shows the calendar year.

Jacob: Where is the accident year?

Rachel: The accident year is not shown explicitly, since it is not used in the regression. But the observations are organized by accident year:

- The first 15 observations are accident year 0 and development periods 0–14. Calendar year = accident year + development period, so the calendar years are 0–14.
- The next 14 observations are accident year 1, development periods 0–13, and calendar years 1–14.

Jacob: Are you saying that if we include an accident year dimension its β would be zero?

Rachel: If the independent variables were not perfectly correlated, its expected β would be zero. The actual β would differ from zero, but the estimated value is spurious. In most cases, it would not be statistically significant. In the regression here, the independent variables are perfectly correlated. If the ß's are constant, we don't get a unique solution.

Take heed: To avoid the problems of perfect multicollinearity, we don't use an accident year dimension.

Jacob: Why is there one less development period for the second accident year?

Rachel: The date of the reserve analysis is the end of calendar year 14.

- For the first accident year, this is development period 14.
- For the second accident year, this is development period 13.

Jacob: I notice that Column E, a third independent variable, is the calendar year *squared*. But exponents are not linear functions. Doesn't exponentiating violate the assumptions of linear regression?

Rachel: The regression coefficients are *linear functions of the parameters*. The variables are not restricted.

- Variables of X_2 and $(X_2)^2$ are fine. They are two separate (but correlated) variables.
- Coefficients of β_2 and $(\beta_2)^2$ violate the assumptions of linear regression.

Jacob: Why would we use the square of the calendar year? Inflation might increase or decrease, but it does not vary with the square of the calendar year.

Rachel: If inflation changes linearly, cumulative inflation varies with the square of the calendar year.

Jacob: Is the same true for the development period trend?

Rachel: Yes. You can use Column E for the square of the development period to simulate a varying loss payment pattern.

Jacob: Does Column F on the illustrative worksheets show the Y values?

Rachel: Column F shows non-stochastic (expected) Y values, based on the assumptions for α, $ß_1$, and $ß_2$, and the values of X_1 and X_2.

Jacob: Must we use this accident year order for the observations? Can we order them by development period or by calendar year instead?

Rachel: Any order is fine.

- The illustrative worksheet uses an accident year by development period order.
- To form residual plots by development period (or calendar year), you may find it easier to order the data points by development period (or calendar year).

Jacob: To re-order the observations, do we start over and simulate again?

Rachel: Excel can sort the observations in any order you wish. But be careful.

- Before sorting, select the entire range: both independent and dependent variables. If you select only the variable by which you are sorting, the other columns are not sorted, and the data points are mis-matched.
- The simpler version of the VBA macro assumes a residual matrix in accident year by development period order. If you change the order, use the alternative VBA macro.

Take heed: If you do not use accident year by development period order, you must change a parameter in the VBA macro creating the residual plots from *TRUE* to *FALSE*.

Jacob: With no uncertainty (if $\sigma = 0$), won't the ordinary least squares estimators always be the simulation parameters? What's the purpose of simulating with $\sigma = 0$?

Rachel: If the trends (the regression coefficients) are stable (= constant), the ordinary least squares estimators are the simulation parameters. If they are not stable, the ordinary least squares estimators are a weighted average, and the residual plot reflects the changes in the simulation parameters. The scenario with $\sigma = 0$ helps you interpret the residual plots.

Summary: We use a triangle of 120 observations, with two independent variables. Focus on the *indexing* of the observations: the calendar year and development period indices for each cell. If you associate loss payments with the wrong calendar years or development years, you may not get proper results.