STUDENT PROJECT DOCUMENTATION: WHAT TO INCLUDE IN YOUR SUBMISSION

Updated: September 28, 2012

(The attached PDF file has better formatting.)

This posting is a guide for candidates who can do the statistical techniques but have trouble with written work. Many student projects submitted in past semesters are excellent: some are humorous; some show keen insight; some are innovative and much impressed our faculty. This outline is *not* a constraint on your write-up.

Examples of past student projects for regression analysis and time series are posted on the discussion forum. Review past projects that interest you to see what other candidates have done. Many projects are well written, with clear exposition of hypotheses, statistical tests, and conclusions. Projects differ in their conclusions; your project need not confirm the hypotheses.

The student project has two parts:

- A statistical workbook, such as Excel spread-sheets, SAS, MINITAB, R
- A written document, such as Word, WordPerfect, PDF, or a text file.

Many candidates use Excel for the workbook and Word for the text document. You may use any software; you are not restricted to Excel and Word.

~ You can use any statistical software package or spread-sheet. Copy the primary output to your Word, WordPerfect, or PDF file and attach the workbook to your email. The course textbooks emphasize data visualization by graphics, such as correlograms, scatterplots, and qq plots. If you form graphics in Excel, copy them to your written document.
~ You can use any text document: Word, WordPerfect, PDF or a TXT file. We grade the content of the file, not its format.

Your documentation is the text file (Word, PDF, WordPerfect). Copy graphs and exhibits that support your conclusions to the document, and include all statistical results, such as the parameters of your regression equation or ARIMA model, goodness-of-fit tests (*F* test, *p* value, R^2, $\overline{R}^2$, Durbin-Watson statistic, Box-Pierce Q statistic, Bartlett's test), and comparison of forecasts to actual values. Refer to other exhibits or regression output in the workbook, and specify where they are found (work-sheet name, tab).

The written document is essential. Many candidates can run the *REGRESSION* add-in and graph the data even without taking a statistics course. The student project shows that you understand what you are doing. Explain what the workbook contains, the logic from the initial hypothesis to the conclusions, why you chose specific statistical tests, significance levels, and procedures, and the implications of your project.

Our faculty does not judge if your conclusions agree with their own views. They check if you use the statistical techniques appropriately and if you understand the material.

The course instructor reads the written document, which should include all results of your project, such as

- "The fitted regression line is …; the R^2 is Z%, indicating that …; the *p*-values for the explanatory variables are … indicating that …"
- "The fitted ARIMA process is …; the Box-Pierce Q statistic is …; with T observations in the time series and K lags, the χ-squared value for a 10% significance level is …"

If your word processor allows, copy relevant graphs to the document. Form the residual plot or correlogram in Excel (or other software) and copy the graph to Word (or other word processor). Explain in the document what the graph shows, such as

- "The residual plot, showing the mean residual at each calendar year, is shaped as a V, indicating that …"
- "The correlogram slopes downward slowly. The sample autocorrelation function is positive for the first 20 lags and negative for the next 20 lags, indicating a non-stationary time series with a changing mean …"

The written document is a report, not just documentation of your workbook. It should include the topics listed in this posting.

- Do not say: "The course instructor can read the Excel workbook to see what I did."
- The course instructor examines if you understand the statistical analysis. The project templates, the data on the NEAS web site, and Excel built-in functions enable you to put together an Excel file. The student project shows if you understand the analysis.

The workbook (or other file) has the supporting spread-sheets and exhibits. The written document may say

- The data are from the web site www.datasource.com.
- They are shown as a matrix of dates and values in Sheet 1 in the Excel workbook.
- The regression of (Y variable) on (X variables) uses the *REGRESSION* add-in.
- Residual plots on all the explanatory variables are shown in Charts 1, 2, and 3.
- Correlograms of the time series and its first and second differences are in Charts …

Statistical analyses that you do **not** need for the final report should be referenced in the text document, with the exhibits in the workbook.

Illustration: "I examined monthly birth rates for seasonality several ways. Sheet S1 shows average birth rates by month, and Sheet S2 shows the average month-to-month change. No month has significantly different rates. Sheet S3 shows the sample autocorrelations for lags 1 through 24. Lags 12 and 24 are not materially higher than other lags. Sheet S4 shows an ARMA model with an AR(12) coefficient. The *p*-value for this coefficient is 55%, indicating that it is not significant."

If you use the statistical analysis, copy the graphs into your text document.

Illustration: "I examined monthly marriage rates for seasonality several ways. Sheet S1 shows average marriage rates by month, and Sheet S2 shows the average month-to-month change. The graphs are reproduced here as Tables T1 and T2. June has higher marriage rates than other months. Sheet S3 shows the correlogram for lags 1 through 24, reproduced below as Figure F1. Lags 12 and 24 have high autocorrelations. Sheet S4 shows an ARMA model with an AR(12) coefficient. The R^2, t values, and *p*-values are reproduced in Table T4. The *p*-value for the AR(12) coefficient is 2.5%, indicating that it is significant."

State your hypotheses clearly, and explain how you test them.

Illustration: "Marriages are presumed to be most frequent in June and September, and least frequent in February and March. The monthly averages show … The June, September, and February averages lie outside a 95% confidence interval …"

Explain if any results reflect random fluctuation or spurious correlations.

Illustration: "Marriage rates have a high three month serial correlation. I believe this reflects the high rates in June, September, and December, not a true autocorrelation with a three month lag. I modeled marriage rates with a 12 month seasonal autocorrelation, not a quarterly correlation."

Some candidates have writers' block. They do the Excel analyses well, but they can't write up the results. Don't worry about the quality of your writing.

- We do *not* grade you on writing style, exposition, or grammar.
- For many candidates, English is a second language; we don't expect flawless style.

You write memos at work. Your supervisor may ask you to clarify the style, rewrite the memo, or shorten the text. The student project is not graded on writing style. We check if

- you understand how to apply the statistical techniques to real data.
- you explain what you have done and how it supports the conclusions.

Many projects written by overseas candidates for whom English is not a first language may not have a smooth style but show good grasp of the statistical concepts.

THE SUGGESTED OUTLINE BELOW IS A MINIMUM; DO NOT FEEL CONSTRAINED BY IT.

We give a *minimum* outline for the written document (write-up). The elements below are recommendations, not requirements. You may write the project in any style you like.

- Don't restrict yourself to this outline. Every student project differs, and your document may have a different structure. Don't force your writing style into the format here.
- Read this outline to see what is expected, and write your student project as you write other memos. If you have trouble with the style, say to yourself: "They are not grading me on the writing style." Explain what you have done and submit the project.

The minimum write-up has the following sections.

Introduction: State the problem and how you approach it. Use statistical terms, but explain clearly your work.

Illustration: For a project on non-constant regression coefficients, you might write:

> Classical regression analysis assumes regression coefficients are constant. If the regression coefficients change over the range of explanatory variables, the R^2, *t* statistic, *p*-value, and *F* statistic do not show this. I use residual plots to identify changes in the regression coefficients. To adjust for these changes, I use dummy variables (or the square of an explanatory variable).

For project on modeling interest rates with ARIMA processes, you might write:

> Financial economists assume that nominal interest rates move with inflation rates, but real interest rates are more stable. I regressed the three month Treasury bill rate on the CPI and examined how to model the residuals. I used first differences and a seasonal adjustment to create a stationary time series, which I modeled with AR(1), AR(2), and MA(1) processes. I used Bartlett's test, the Box-Pierce Q statistic, and the principle of parsimony to select the AR(1) model.

These are sample paragraphs. Say what your student project does, so that the faculty member reviewing it knows what to expect. Feel free to use any format.

- One candidate used a Jacob/Rachel dialogue – except that Rachel asked questions and Jacob answered.
- Many candidates use personal introductions, explaining their interest in the topic, such as "My parents were both divorced from previous marriages, but their present marriage works well. I have often wondered whether divorce rates differ for first vs second marriages or by age at marriage."

Data: You gain most from the student project if you use data that interest you.

- The NEAS web site has many data sets that you can use for the project.
- The internet has thousands of web sites with good data. Choose a topic and search the web for data.

If you use data not from the NEAS web site, describe the data. If the student project is a time series, define the series, the period, and the intervals. You might say:

- The time series is the French unemployment rate from 1980 through 2003 at quarterly intervals from the *XYZ.gov* web site.
- The time series is daily temperature in Puerto Rico for 2000-2005, from the *XYZ.com* web site. I adjusted for seasonality by dividing by the average daily temperatures from 1950 through 2005.
- The data are (i) the monthly crime rate in Los Angeles, (ii) immigration rates and gang activity in Los Angeles, and (iii) unemployment rates in Los Angeles. These figures are published on *XYZ.com*.
- I used female mortality rates for ages 35 through 65 based on the *XYZ* mortality table.

Almost any data set can be used for the student project. Do not worry that your data will not show an ARIMA process or will not support your statistical hypothesis. But:

- Use enough data points to perform the statistical techniques. If you have only ten points, you can not fit an ARIMA process and your regression will not be significant.
- Do not choose data that are white noise or simple random walks.

Illustration: A time series of earthquake frequency may be white noise. If you expect no autocorrelations and the correlogram shows no significant autocorrelations, choose a different data set.

We can not judge *a priori* if you have enough data points. If a topic interests you and has only 50 data points, use it. If your topic is good, we do not fault you for having a short data set.

Use appropriate periods. For a time series project on temperature, use daily readings (or hourly readings), not monthly averages. The monthly averages lose the day to day sample autocorrelation. Hourly readings make an excellent project: the time series has a daily cycle overlaid on an annual cycle.

Company data: If you use your own company's data, disguise the data by scaling the figures. Multiply the data by a factor or add a constant. The adjustment may cause different regression coefficients or ARIMA parameters; that is fine. You might write:

- I used a paid loss triangle for 12 years. I multiplied the figures by a factor, subtracted a constant, and added a random draw from a normal distribution with a mean of zero.

You can do a student project on the same topic as your company work. You have a data sample or a time series that you know well and which you have already examined for data errors and outliers. You may have done all the analysis for your company work and written a memo with your results. For your student project, re-write your memo to show how you applied the statistical techniques taught in the on-line courses.

Illustration: Many past candidates have written student projects on loss cost trends, loss development factors, or loss completion factors. The student project provides a different perspective on the actuarial work. Some candidates have done generalized linear modeling on general insurance class relativities. This analysis is excellent for a regression analysis student project.

If you use data from a project template on the NEAS web site, state which data you chose. You might write:

- I used baseball won-loss records for the National League for 1911 through 1960.
- I simulated paid loss triangles for a 15 by 15 array.
- I used overnight LIBOR rates from 1945 through 1975.

If you use outside data, explain its characteristics.

Illustration: If you regress mortality rates on age and sex, say what table the data are from, whether they are population data or insured data, and what years they are from.

If you simulate data, describe the simulation parameters (the α, the ß's, and the σ). You can simulate data of various types for a student project. The simulated paid loss triangles for the project templates on regression analysis for loss reserving are an example. Mahler's *Guide to Regression* has many simulations for actuarial subjects.

Topic: Good student projects have limited focus. Choose a topic, form questions, apply statistical techniques covered in the course, and answer the questions.

Illustration: For the project template on regression analysis of loss reserving, you forecast future loss payments. The student project has a specific topic, such as:

~ The inflation rate or the development trend may not be constant.
~ The variance of the error term may not be constant, but may vary with the calendar year or the development period.

Simulate data with a discrete change in the inflation rate or payment pattern, or a continuous change in the inflation rate or payment pattern, or a variance of the error term that depends on the calendar year or development period.

Show the R^2, *t* statistic, *p*-value, and *F* statistic assuming a constant inflation rate, payment pattern, or variance. Explain why these values indicate the regression is significant even though the assumption is not correct. The explanation need not be long. You might write:

> These statistical tests help decide whether to reject a null hypothesis if the assumptions of classical regression analysis are true. They do not test whether the assumptions are true. If the inflation rate changes from 5% in one part of the sample to 15% in another part, and the null hypothesis is that inflation is zero, the estimated inflation rate may be 10%. The statistical tests reject the null hypothesis that inflation is a constant 0%. We should assume inflation is 10% only if it is constant over the entire sample.

For the time series student project, state the topic clearly. Do not just say "the project deals with interest rates." Your project might

- Fit an ARIMA process to a given time series to see if an autoregressive or moving average process explains the observed values.
- Compare two periods to see if the same ARIMA process is appropriate before and after a critical event.
- Fit an ARIMA process to the residuals of a regression analysis.

You might write:

- I fit ARIMA processes to time series of nominal interest rates and real interest rates, or nominal interest rates divided by the expected inflation rate. Both nominal and real interest rates can be fit with ARMA(1,1) processes, but the fit works well only for real interest rates.
- I compare the interest rate time series for two periods to see if the same ARIMA process is appropriate. The rising interest rates in the stagflation era of the 1970's require logarithms and first differences to form a stationary time series; the stable interest rates of the 2000's can be fit with an MA(1) process.

Statistical techniques: State the statistical techniques you use and explain how they relate to the problem in your project. Explain the results of your analysis and the implications.

Illustration: For an *F* test, explain what the unrestricted and restricted equations are, so readers understand the project.

- Write the equations for the unrestricted and restricted regression lines.
- State the null hypothesis, relating it to the restricted regression equation.
- State the degrees of freedom for the *F* test.
- State the result of the *F* test, the critical values, and the significance level.

You might write: "I use an F test to see if the two Leagues have the same relation of past experience to future won-loss records. The null hypothesis is that the different relations in the previous section of this project reflect fluctuations, not true differences. The result is significant at 10% level, for which the critical value is …, but not at the 5% level, for which the critical value is ….

Do not just say: "I used an *F* test to see if two samples are from the same distribution." The course instructor judges if you understand what an F-test shows. Demonstrate that you understand the statistical procedures.

Take heed: Explain your charts. Label the indices, and explain the index values.

- If you analyze monthly interest rates over 30 years, the horizontal axis may be a month from 1 to 360. If the interest rates peak at month 125, identify the date of this month.
- For residual plots, specify the axes. The horizontal axis is an independent variables or the dependent variable; the vertical axis is the residual.

Say what you look at in a graph. For a residual plot, you might write:

~ To test if the slope coefficient is constant, I examine the slope of the line joining the average residuals.
~ For conditional heteroscedasticity, I examine the spread of the residuals.

Copy Excel graphs to your write-up and label the axes, so the course instructor follows your arguments. If you leave the graphs and charts in the Excel workbook, it is difficult to follow your reasoning. A course instructor who does not know what graph or chart you refer to or who does not know what the axes represent may ask for better documentation, and it may take loner to get your VEE credit approved.

Explain what the statistical test implies.

~ If the residual plot looks like an upside-down V, say what this implies. If you get stuck writing an abstract explanation, give an example.
~ If the residuals are more spread out for high values of X than for low values, explain what this means.

Illustration: The residual plot appears like an upside-down V. If the actual inflation rate is declining / rising… (explain the effects). For example, if the inflation rate is Z in the first 5 years and Z′ in the next five years …

Corrections and Adjustments: Explain how you correct problems in your data. For a project on non-constant regression coefficients, explain how dummy variables, squares of explanatory variables, or other methods correct for the problem. Show the results of the revised regression equation and explain why it is superior.

Not all student projects have corrections. Your project may use an *F* test to compare two sports teams, or your project may fit an AR(2) process to a time series and conclude that it fits well. But many projects have a series of statistical tests or a series of tests and adjustments. Fitting ARIMA models uses a sequence of procedures. You graph the time series, looking at means, trends, drifts, variances, cycles, seasonality. The first adjustments are

- Breaking the time series into periods
- Using seasonally adjusted data
- Using structural models, such as real interest rates instead of nominal interest rates

Your student project may be a series of adjustments. You

- Form a correlogram and check if the time series is stationary.
- Compute moving averages to identify trends.
- Adjust by taking differences or logarithms and differences.
- Compute monthly averages or 12 month autocorrelations to identify seasonality.
- Adjust the data to offset seasonality.
- Fit ARIMA models, based on the sample autocorrelation function.

For each step, you may form charts or graphs. Copy the charts or graphs into your Word document and explain how you formed them and what they imply. Don't explain the Excel code, but say what you did.

Illustration: I checked for seasonality two ways:

- I formed monthly averages in three steps:
  - Compute the average monthly sales in each year
  - Divide each month's sales by the average monthly sales to get monthly relativities
  - Take the straight average of the monthly relativities over 20 years

    - The column chart by month (Figure 1) shows high sales in May, June, and July and low sales in December, January, and February.
- I examined ths sample autocorrelations for 6, 12, and 24 months:
    - The 12 and 24 months sample autocorrelations are high
    - The 6 month sample autocorrelation is negative
    - The correlogram is Figure 2, with markers at 6, 12, and 24 months

Do not write: "Everything is in the Excel work-sheets." Figuring out what you did from an Excel worksheet is hard. If the work-sheet is not well documented, it may be impossible.

Copy graph, charts, and conclusions table from the work-sheet into your document.

- Copy correlograms and plots into your document and put them next to your comments.
- If you have trouble with cut and paste from Excel into Word, or if you use a text file that does not accept pictures, refer to the Excel chart or graphic.

Don't copy long Excel tables into the written document.

- Refer to the work-sheet or region in your written document.
- Cite the important results, such as an *F* statistic or a Box-Pierce Q statistic.

STOCHASTICITY AND SIMULATION

Some student projects simulate data. If you simulate data, turn off the stochasticity on your first run. You might write:

- I ran the simulation first with σ = zero to verify that I get the expected results. The regression equation is …, with an R^2 of 100%.
- I then used σ = 0.01 to verify that I am adding the error term correctly. The regression equation is …, with an R^2 of 95%.
- I then used σ = 0.2 for the student project.

PROJECT LENGTH (TIME AND PAGES)

The student project is a serious task. To cover the topics in this posting, you need several pages of text plus charts, graphs, and tables. The project has no required length, but do not leave out critical sections.

If you are unsure whether to include a section, put it in. Instead of saying: "I did not find any seasonality in the time series," write: "I computed monthly averages, as shown in Chart 1, which do not show seasonality." The data are yields on 20 year Treasury bonds, so I did not expect seasonality.

The time needed for the student project varies with your grasp of the statistical concepts, your motivation, and past work with these topics.

- If you understand the concepts, the student project does not take long to complete.
- If you don't understand the concepts, you might spend days wondering what to do.

TOPICS

If you have never before written statistical reports and you are not familiar with internet search engines, you may find the project difficult at first. If you feel lost, use one of the project templates and follow the instructions. But the student project is far more enjoyable if you choose a topic that interests you. Choosing an interesting topic is not hard, given the enormous variety of data on the world wide web.

Illustration: You want to do a student project on crime rates: perhaps a time series analysis of murder rates over the past 20 years or a regression analysis project relating crime rates to city governance or city size. Use

internet search engines (Google, Yahoo, MSN) to find data on crime rates. A few minutes of looking at sites leads you to the FBI home page. Use the site map to find a section called *statistics*, which brings up hundreds of files of crime statistics. You can download many of these files in Excel format. You may spend an hour looking at the different files and picking data you want to analyze.

If you use statistical analysis at work, you can mold your work into a student project.

Illustration: You fit exponential curves to average claim severities to project loss cost trends. Use the same data for a time series student project, with several additions:

- Take logarithms and first differences to make the series stationary.
- Fit an ARIMA process, not an exponential curve.
- Examine seasonality. Average claim severities vary by quarter in most lines.
- Use a structural model. Deflate average claim severities and apply the ARIMA process to the claim severities in real dollars.

The student project will not take long, since you already have the data in Excel files, you know the data characteristics, and you have already worked with these figures. Most insurers do not adjust trend data for seasonality and use very simple exponential trends. Your analysis should improve the trend projections. You receive VEE credit for the on-line course and the time series analysis may be used in your pricing studies.

If you feel lost, review the project templates and the past student projects on the discussion forum. Project templates for sports scores, interest rates, daily temperature, and other topics are on the discussion forum. The NEAS web site has hundreds of data sets in Excel format, with full explanations of the types of analyses you can do. The postings for the project templates have several forms:

- Step-by-step guides for the more common analyses.
- Jacob / Rachel dialogues for the hypotheses and analyses you can use.
- Discussions of topics used for past student projects. The discussions are broad, with many suggestions. You are not expected to do everything; pick a topic to focus on.

Look at the past student projects on the discussion forum. The discussion forum has the write-ups, not the work-sheets. The past projects give you ideas, which you implement in our own work-sheets. Your student project need not be new. If a past student project interests you, choose other data and do a similar project.

Illustration: Many candidates analyze won-loss records of their home teams. The material on the web site gives dozens of ways to write a student project on sports figures. If you want to do a sports project, read the project templates, see what other candidates have done, select data from the NEAS web site or from other internet sites, and do the analysis. You can't repeat another candidate's project, but we can do a similar sports project on a different team or a different player, using data from a different web site and different hypotheses.

DISCUSSION FORUMS

The student projects are independent, but you get ideas by discussion with others. Feel free to discuss the student project on the discussion forum. Some items are confusing at first, but they become clear after a bit of discussion.

The Excel built-in functions are hard if you have never used them. Discuss how to use the Excel built-in functions like *RAND* and *NORMSINV*, Excel names, the *REGRESSION* add-in, and the *SOLVER* add-in. The student project does not test Excel; feel free to copy material from any Excel file on the web site.

Use your own data for the student project. Data from another web site are the best. If you use data from the NEAS web site, use a different time period for a time series or different parameters for a simulation.

If you are completely lost, begin by reproducing the analysis in a project template. Once you understand what is expected, choose a different set of data and do a similar project.

JOINT PROJECTS

Two or more candidates from one firm may be taking the same on-line course. You can discuss the project with another candidate at your firm. You might both do ARIMA modeling of interest rates or analysis of sports won-loss records. But your projects and your write-ups must be separate.

- You should not have a single project with two authors.
- You should not have the same project with different parameters.

The SOA / CAS wants each candidate to write a student project.

DUE DATES

Actuarial candidates are busy with work, courses, student projects, and study for actuarial exams. Some candidates have young children and many responsibilities. A candidate might take a course in January and February, study for an actuarial exam in March and April, and do the student project in May and June.

We don't set due dates. You are responsible for effective use of your time. But be mindful of your future work schedule. Some candidates presume they have more time after their exam. But after your exam, you will be thinking about the next exam. Many actuaries' work schedules keep getting busier until they retire.

We recommend: Right after the final exam, do the student project. You will finish it efficiently when the course material is fresh. Getting VEE credit gives enormous motivation for exam study. A project that takes a few days when you know the material well may drag on for several weeks if you keep postponing it.

QUESTIONS

Questions about the student projects are of several types. For quicker response, send questions to the proper person(s). Send administrative questions (by email) to the NEAS office: to whom should you send the project, when is the project due, and so forth. Send the project to the NEAS office; you don't receive credit for the course until the project is received and graded.

Questions about statistical techniques and Excel built-in functions should be posted on the discussion forum. The illustrative work-sheets and step-by-step guides show to use many techniques and built-in functions needed for the student projects. Additional questions may be posted on the discussion forums, such as "Does Excel have a built-in function for generalized linear modeling? What is a reasonable confidence interval for Bartlett's test?"

The NEAS faculty compiles questions into dialogues and general postings. The common questions about student projects are discussed on the discussion forum.

Some candidates have questions about their project results.

- The correlogram for a time series declines slowly; is the time series stationary?
- The *p* value for an explanatory variable is 11%; can I use it in the regression analysis?

The student project shows how you deal with these questions. Statistical analyses are subjective because the data are stochastic. Answer each question, and explain why you chose a specific answer, and how else you might answer the question. The student project shows if you can deal with real statistical issues. We grade the project by your reasoning, not by whether the answer is right or wrong.

Illustration: A correlogram declines slowly. You might write:

- The correlogram declines to about zero after 8 lags, suggesting an autoregressive process with several parameters.

- The correlogram does not decline rapidly enough for a stationary process. The correlogram of the first differences declines to zero by the second lag.

Some items that occur frequently in actuarial and financial time series are not covered well in the textbook. Candidates repeatedly find correlograms that decline slowly, are positive for the first K lags, and then negative for the next K lags. This correlogram indicates two phases of a time series with different means. We explain the implications of this correlogram in a separate discussion forum posting. Read the postings on the discussion forums; they answer many of the questions on student projects.