SESUG Paper 250-2018

Comparison of SAS®, SAS University Edition®, and Microsoft Excel® in Collaborative Data Analysis of Physical Properties by Scientist and Statistical Programmer

John Schreiber, University of South Florida; Elizabeth Schreiber, IRS

ABSTRACT

Calibration data for novel instrumentation and laboratory data were analyzed using common regression techniques presented in physics literature and a pooled regression respecting the underlying physical constraints of mechanical analysis. This paper compares and contrasts implementation of these techniques in SAS®, SAS University Edition®, and Microsoft Excel®.

INTRODUCTION

The researcher developed an apparatus to measure physical properties of samples. Analysis of this data using standard regression models was unsatisfactory. A more appropriate parameterization of a new pooled regression model is presented. The researcher and analyst compare the capabilities and convenience of three software products for fitting many linear regressions and model fitting diagnostic plots.

RESEARCH GOAL

The researcher (a physicist) developed a novel apparatus to measure physical properties of a protein-based polymer. The custom-built apparatus measures vertical displacement as a function of force applied to the sample. In Phase 1 of the analysis, the response variables were Young's Modulus and the spring constant. Calculation of both response variables includes estimating the slope of the regression line for displacement as a function of force. In later phases the research goal is to will develop response surface plots of Young's Modulus and the spring constant.

To validate the apparatus (Figure 2), the researcher measured springs with known spring constants manufactured to standard dimensions. Each spring was subjected to 5 trials and in each trial the spring was loaded and unloaded through a series of 11 masses. The springs and masses[1] were chosen in ranges where the response would be a linear function of the independent variable force.

Unlike the response of a spring to increasing force, polymers do not exhibit completely linear behavior. When applying a range of increasing force to non-rigid samples, we expect a contact phase, a linear phase, and a failure phase (Figure 2). As expected, samples of differing composition progress through phases of the typical response curve at different force levels. While using the apparatus to measure polymer samples, physical constraints caused challenges. There was uncontrolled variation in inserting "squishy" samples into the apparatus. Some samples broke. Sample sizes differ between trials and some observations within each trial had to be excluded.

---

[1] When weights are stacked on the sample, $F_{applied} = m\ g$. For small deformations that remain within the region of linear response, the sample or spring responds as $F = k\ \Delta x$ where Δx is the change in height under load. When the function $\Delta x(F)$ is plotted the spring constant, k, may be found as $k = \frac{1}{slope}$.

When studying deformable bodies, it is customary to use the contact area the force is applied to and the initial height of the samples as scaling factors to redefine these variables as stress, $\sigma = \frac{F}{A}$, and strain, $\epsilon = \frac{|\Delta x|}{x_0}$. Now—again for small deformations that remain within the region of linear response—the sample or spring responds as $\sigma = Y\ \epsilon$. Y is known as "Young's Modulus" or "elastic modulus" and is typically reported in units of Pa, Pascals. If the spring constant has already been found, Young's Modulus may be solved for as $Y = k\ \frac{x_0}{A}$.

Figure 1. The Apparatus measures displacement for a known application of force.

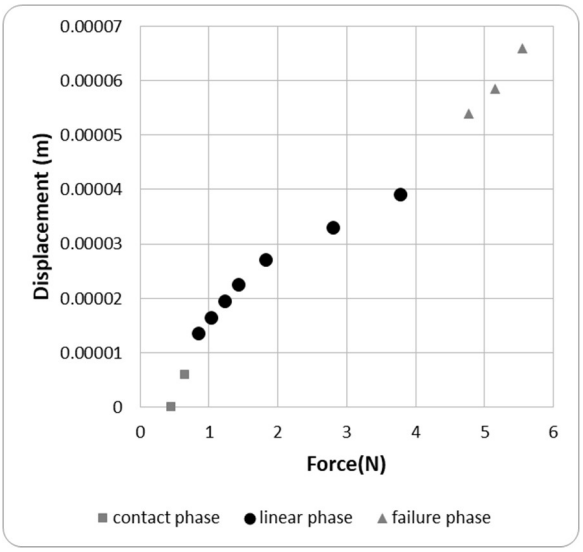


Figure 2. Typical displacement as a function of force curve showing initial conditions/contact phase, linear phase, and failure phase.

LINEAR MODELS

In physics classroom manuals, the typical calibration or linear regression exercise fits each trial separately and averages the resulting slope estimates. Assumptions about error distributions are not rigorously defined.

Step 1: For each trial, estimate b_i, the intercept, and m_i, the slope, due to x for the model

$$y_{ij} = m_i x + b_i + \varepsilon_{ij}$$ for trials $i=1, 2, \ldots, t$ and observations $j=1, 2, \ldots, n_t$.

Step 2: The overall slope estimate is the average over all trials of the slope estimates

$$m = m_1 + m_2 + \ldots + m_t$$

For the controlled equipment calibration using springs with known spring constants, manufactured to standard dimensions, and with all trials having the same number of paired x,y observations; this is simplistic but effective. The individual b_i estimates represent uncontrollable initial conditions and the overall m estimates the effect the independent variable has on the dependent variable (the effect of force on displacement). For t trials with unequal n_t, the degrees of freedom for the two-step slope estimate becomes $df = (\sum_{i=1}^{t}(n_i - 2)) - 1 = (\sum_{i=1}^{t}(n_i)) - 1 - 2t$. This is the overall number of observations minus 1 minus the number of parameters estimated in Step 1.

When the apparatus was used for polymer samples to explore their unknown physical characteristics, the number of observations possible in each trial varied. Weighting each trial equally in Step 2 would artificially inflate the influence of samples that failed in testing and undervalue samples with many observations in the linear phase.

To objectively reweight the samples, a new three-step approach was used.

Step 1: For each trial, estimate b_i, the intercept, and m_i, the slope, due to x for the model

$$y_{ij} = m_i x + b_i + \varepsilon_{ij}$$ for trials $i=1, 2, \ldots, t$ and observations $j=1, 2, \ldots, n_t$.

Step 2: For each trial, subtract the intercept to correct for uncontrolled variation in initial conditions

$$y_{ij}^* = y_{ij} - b_i$$

Step 3: For all observations pooled together, estimate m, the pooled slope, for a no-intercept model

$$y_{ij}^* = m x_{ij} + \varepsilon_{ij}$$ for trials $i=1, 2, \ldots, t$ and observations $j=1, 2, \ldots, n_t$.

This new pooled slope estimate both avoids inflating the influence of sparse trials and improves the degrees of freedom available for the slope estimate, the overall number of observations minus 1 minus t (the number of $y_{ij}^* = y_{ij} - b_i$ correcting equations in step 2).

The new pooled approach can be re-written as a linear function of a regressor variable x and a series of indicator variables $I_1, I_2, ..., I_t$ representing initial condition and contact phase of each trial. This allows estimation of β_1, the slope due to x; and $\gamma_1, \gamma_2, ..., \gamma_t$, the correction coefficients for each of the t trials, in

$$y_j = \beta_1 x_{ij} + \gamma_1 I_1 + \gamma_2 I_2 + ... + \gamma_t I_t + \varepsilon_{ij}$$

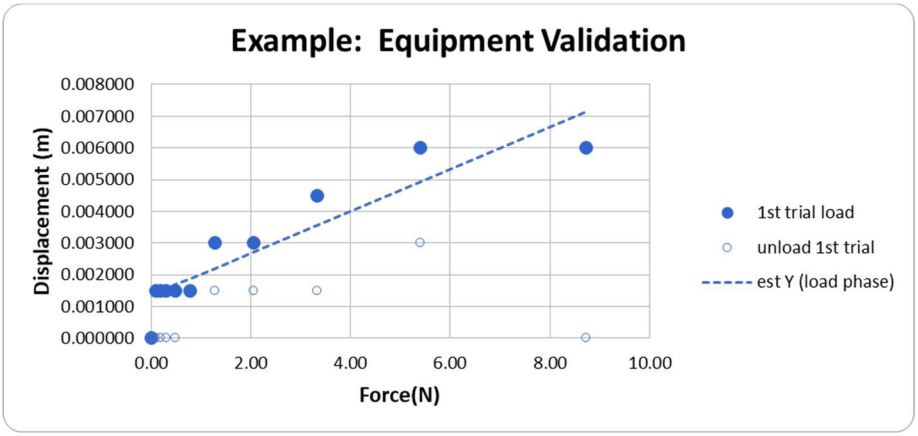
where the indicator variable $I_j = 1$ for trial j and 0 otherwise

for the trials $i = 1, 2, ..., t$ and observations $j = 1, 2, ..., n_t$.

The usual intercept parameter β_0 is eliminated to avoid over-specifying the model. This new parameterization essentially partitions the traditional β_0 intercept into custom intercepts for each trial representing initial sample condition and contact phase. This new parameterization uses fewer degrees of freedom $df = (\sum_{i=1}^{n_t}(n_i)) - 1 - t$. This is the overall number of observations minus t+1 parameters. This protects t more degrees of freedom degrees of freedom for more powerful hypothesis tests.

EXAMPLE ANALYSIS RESULTS

For each trial, a regression was fit and graphed as in **Error! Reference source not found.**. For illustration purposes, the unpublished data was multiplied by an obfuscating constant.

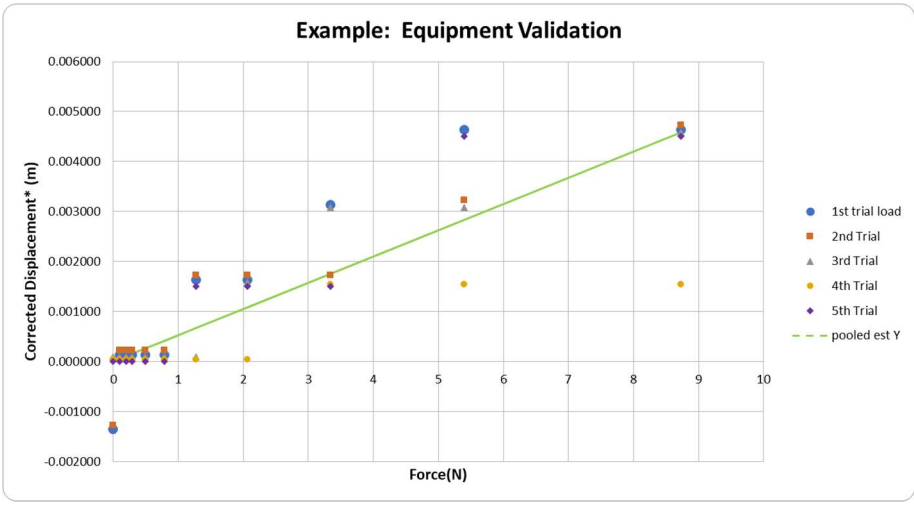


linest	Linear Estimation y=mx+b
49.9848	ANOVA F
9	degrees of freedom
0.00066	m=slope estimate
0.00136	b=intercept estimate
0.84742	r^2=coeff of determination
0.92055	correlation
0.00009	standard error of slope est
0.00031	standard error of intercept est
0.00082	standard error of y estimate
0.00003	sum of squares regression
0.00001	sum of squares residual

Figure 3. Load and unload phases from one trial applying known force to a spring. Prediction line for the model displacement=m*force+b.

Regression models were fitted for each trial, displacement adjusted for the initial condition estimate was calculated, and a pooled no-intercept regression was fitted to the data pooled across all trials as in Figure 4. This was done manually in Excel and programmed in SAS with a macro called once for each trial. The adjusted displacement data was pooled and run through a final no-intercept regression to get the needed pooled slope estimate. The alternative parameterization using indicator variables can also be fitted in SAS to get the pooled slope estimate. Once a shared folder for SAS University Edition is set up and data stored in that folder, updating only the fully qualified path in libname and filename statements enables the same SAS programs to run in the virtual box using the SAS Studio interface.

The researcher's data includes trials of a range of springs to validate the novel apparatus and trials to measure experimental samples. Samples cover a range of values for two factors. The experimental design space will allow 3-D response surface mapping once Young's modulus and the spring constant are calculated after obtaining the pooled regression estimates from repeated trials at each specific composition. Thus, scalability of the regression analysis process is critical.



Source of Estimate	Slope Estimate	df
1st trial	0.000662	9
2nd Trial	0.000572	9
3rd Trial	0.000572	9
4th Trial	0.000221	9
5th Trial	0.0005971	9
pooled est Y	**0.000525**	49
average slope	0.000525	44

Figure 4. Load phase data from five trials applying known force to a spring with prediction line from pooled no-intercept regression. Corrected displacement is displacement corrected for initial conditions for each individual trial.

SOFTWARE COMPARISON

Many factors should be considered in choosing software for collaborative analysis. Availability of the needed analytical and graphing capabilities is highest priority. Scalability of the analysis is critical. Among other critical features are cost and convenience.

All three software options have most of the analytical capabilities needed (Table 1) to produce linear model results, related goodness-of-fit plots and diagnostics, and summary result graphs and tables. Scaling to efficiently process many pooled regressions is easier in PC SAS® or SAS® University Edition. Excel cannot produce the 3-D response surface graphs required in later phases of the researcher's work.

Analytical and Graphing Capabilities	Microsoft Excel	PC SAS® 9.3 with Display Manager Interface	SAS® University Edition
Fit simple linear regression y=mx+b	{=LINEST(Known_ys, Known_xs, const, stats)} (Stat Add-In)	SAS/STAT® PROC REG (and many other procedures)	Point and click or use the SAS program
Fit no-intercept linear regression y=mx	LINEST function with const=FALSE	PROC REG with a NOINT option	Point and click or use the SAS program
Estimate model parameters	YES	YES	YES
Estimate standard error of parameters	YES	YES	YES
ANOVA fit test	YES	YES	YES
Calculate predicted values and residuals	Incorporate LINEST results in formulas to calculate predicted and residuals.	PROC REG with an OUTPUT statement with PREDICTED= and RESIDUAL= options	Point and click or use the SAS program
model fit diagnostic plots	Requires template development. Manual intervention is required to customize the range of observations.	YES. SAS/STAT® PROC REG plots automatically. SAS/GRAPH® can plot PROC REG output.	YES

customize plots (titles, axes, symbols, data ranges, …)	YES, manual intervention is required.	SAS/GRAPH® PROC PLOT using PROC REG output	YES
Researcher can independently edit labels and titles in graphs	YES	NO	Possibly
Create graph files with descriptive filenames	manually	Programmatically with SAS/GRAPH® and PROC PLOT	YES
Scalable to multiple independent trials	Templates allow limited manual scaling.	SAS allows manually programmed scaling & macro facility enables efficient scaling.	Point and click scalability is limited. Programs allow efficient scaling.
3-D plots (required for analysis of later research phases)	NO	SAS/GRAPH® PROC G3D Plots with animated (gif) rotation can be programmed.	YES

Table 1. Comparison of Analytical and Graphing Capabilities among Microsoft Excel®, SAS® and SAS University Edition®

While analytical capabilities are most important, other factors should be considered when choosing a software package for remote collaboration. Knowledge brought to the collaboration by both researcher and analyst, cost, convenience, and compatibility across operating systems are also important. A full comparison is presented in Table 2. Excel is free and well known; however, filesharing challenges and occasional poor translation of advanced features and graph colors across operating systems limit its usefulness. Cost and limited filesharing make collaboration using traditional SAS® programs less than ideal. Features to simplify remote collaboration and a moderate learning curve make SAS® University Edition a good option for collaborators with differing levels of analytical programming experience.

Cost and Convenience Factors	**Microsoft Excel**		**SAS® Software**		**SAS® University Edition**	
	Researcher	**Analyst**	**Researcher**	**Analyst**	**Researcher**	**Analyst**
Cost	already installed		Need to buy a license and use a Windows emulator	Currently licensed	Free for noncommercial use Student: PhD research	Independent learner: practice new skills
Software available to researcher and analyst	Office 365 for macOS X	Office 365 for Windows 10	Not currently licensed	PC SAS® 9.3 with Display Manager Interface	• VMware Fusion for OS or Oracle VirtualBox • Google Chrome • Mac OS X	• Oracle VM Virtual Box • Microsoft Explorer 11 • Windows 10
Prior knowledge	YES	YES	No	YES	No	minimal
Level of effort to learn the software	none	none	high	none	moderate	moderate
input data updates directly with easy version control	YES	YES	No. Researcher e-mails files to analyst	YES, or read files e-mailed by analyst	Either YES or e-mail to analyst	YES

edit labels and titles in graphs	YES	YES	No	YES	YES	YES
Time to run the analysis and proofread results	• Discussion: Unknown hours • Template development: 20 hours • Execution: 1 hour per set of trials • Edits: 1 hour per set		• Discussion: Unknown hours • Program development: 8 hours • Execution: <10 min. per set of trials • Edits: <10 minutes per set after minimal programming time		• Discussion and Setup: Unknown hours to enable remote sharing • Used PC SAS program • <10 minutes per set of trials • Edits: <10 minutes per set after minimal programming time	
Remote file sharing with version control	Google Docs (Graphs and formulas don't always translate.)		No. PC SAS is on a PC with strong firewalls.		YES	
Edit graph text after inclusion in a document or slides	YES, if the graph is pasted with an embedded workbook		No. (Technically, *.emf can be edited but moving files from PC to Mac can have unexpected results.)		No	

Table 2. Comparison of Cost and Convenience Factors among Microsoft Excel®, SAS® and SAS University Edition®

CONCLUSION

While all three software solutions can be used for the illustrated analysis, future collaboration between the researcher and analyst will use the SAS University edition. Both SAS® and SAS® University Edition have response surface graphing capability that will be used in later phases of the research. Excel has scalability limitations. While SAS® is scalable, expecting the researcher to learn the display manager interface used by the analyst is not reasonable. For projects meeting the noncommercial requirement, SAS® University Edition's cloud solution mitigates many of the logistic challenges of remote collaboration with only moderate effort to learn new programming and without sacrificing scalability.

ACKNOWLEDGMENTS

The authors wish to thank current and former University of South Florida professors Donald Haynie, Garret Matthews, and Alex Volinsky.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

John Schreiber
University of South Florida
jhschrei@mail.usf.edu

Elizabeth Schreiber
elizabeth.schreiber@irs.gov