Benchmarks Online

Skip Navigation Links

Page One

Campus Computing News

SkillPort Training Site Update

Summer Hours

Fry Street Fire Leaves Campus Powerless, Disconnected

InHouse Series on Information Security

Today's Cartoon

RSS Matters

The Network Connection

Link of the Month

Helpdesk FYI

Short Courses

IRC News

Staff Activities

Subscribe to Benchmarks *Online*

Research and Statistical Support
University of North Texas

RSS Matters

You can link to the last RSS article here: R Techniques: Summarizing Data By Grouping Variables - Ed.

How Long Should My Data Analysis Take?

By Dr. Rich Herrington, Academic Computing and User Services, CITC

These thoughts **(**disclaimer at the bottom of this column*) are motivated by the "quick-fix", "take the shortcut" mentality that I am seemingly surrounded with on a day-to-day basis....this was a real question posed to me:

The comment:

I was told by 'so-and-so' that it should take no more than two hours to clean, and 'run' my data. What do you think?

My long reply to the student:

Yes, two hours is a reasonable estimate of how long it might take to finish your data modeling/analysis project IF ONE WERE TO:

1) IGNORE checking assumptions of the parametric model(s) being generated, and ignore any steps necessary to "correct" for those problems found (e.g. normality of residuals, issues of heteroscedasticity, lack of independence in observations - most people completely ignore this last item (independence)...that is looking for the presence of clustering (e.g. presence of a "significant" intra-class correlation)....in other words, effects due to undetected or unrecognized clustering (lack of random samples). While some may feel this is un-necessary, it is completely clear that violations of this assumption are devastating for model validity (one might even think of this lack of independence as a model mis-specification).

2) Along the lines of 1), IGNORE any issues with bias generated due to missing values and the pattern of missingness that might be present. Many people incorrectly model missing values because they assume (in an unthinking way) "missing completely at random" (MCAR). MCAR is not usually a reasonable condition that can be met with confidence. Even so, one would like to know how that missing-ness is presenting itself in the observed

data (e.g. numeric or graphical displays that depict patterns of missing-ness are very helpful here).

3) NOT validate any of the fitted models after estimation (e.g. cross-validation; bootstrap-validation, etc).

4) NOT produce a calibrated model (with calibrated beta coefficients) after validation, that takes into account the optimism (bias) of the originally fitted model (e.g. optimism in R^2)...additionally, do not estimate the predictive validity of the model after calibration for bias.

5) NOT produce revised p-values or CI's for the significance tests of model fit that take into account the potential bias in R squared (numerical index of model fit that, in small sample sizes, is a biased estimate of the population estimate of R squared).

6) NOT generate confidence intervals for effect sizes and not generate graphical based displays of those intervals for communicating succinctly the main information concerning the parameter estimates (e.g. CI's for R-squared or Cohen's f effect size).

7) IGNORE any issues concerning uncertainty in the "model selection" stage - e.g. using variable subsets from the original set of variables. That is, ignoring any adjustment methods that would take into account over-fitting and account for inflation of error rates when searching through potential models - either type I, II, or some false detection rate (FDR) error rate.

8) IGNORE the power (sensitivity) of the statistical test. If one's data is a LARGE data set, then it is likely that the sensitivity for statistical tests (power) will not be an issue for effects sizes that one cares about. But, if it were, how would we deal with the fact that your data is observational? - in these situations it is not uncommon to have low a-posteriori power (not that this is a meaningful concept anyway) given a small sample size. This issue is problematic with a lot of folks. Many people just don't accept that "power" (within the Neyman-Pearson framework) was and has always been a design parameter that predetermines the *expected* operating characteristics of the hypothesis(s) test; and that this a-priori probability (if design procedures are followed appropriately) is more meaningful and useful if predetermined before sampling occurs (along with sample size, test threshold "alpha", and the expected difference - the "effect size"). It is clear that after data have been collected (with a predetermined critical threshold, and sample size), that the observed p-value for the data and the power are completely co-determined. For the observed effects size and observed sample size, the power will be small when the p-value is large, and vice-versa. Power is for the most part a useless concept after data collection (within the inferential framework of Neyman-Pearson). So, again, for a fixed sample size and low power with observational data - what valid conclusions can one draw? How can one use all of the available evidence, in hand or otherwise, to maximize the utility of the study?

9) NOT appreciate that classical inference based on thresholds (or critical values) and error rates (Type I and Type II) was not designed to provide "evidence" in a single study toward the hypotheses under consideration. Classical inference as taught in introductory statistics courses has been considered by some to be a contentious synthesis of two (arguably irreconcilable) inferential paradigms:

a) Fisherian p-values under the NULL hypothesis designed to provide evidence of the discrepancy of the current data from the null hypothesis (assumed in the current single study),
and the;

b) Neyman-Pearson behavioral approach which is based on the minimization of decision errors
across the domain of all such equivalent tests. That is, the rejection of an observed p-value in
comparison to a threshold p-value provides information about the collection of all potential,
similar tests. It was not intended to provide information about the single, current test, that is
under scrutiny.

This is all to say that drawing conclusions about how the data "in hand" informs the hypotheses under consideration is tricky business at best and is certainly NOT automatic - this process takes time and careful thinking! An approach that some folks advocate is to take the observed p-values under the null sampling distribution (and under appropriate conditions), generate "bayes factors" to supplement the information that is obtained using the hybrid logic of the Fisher and Neyman-Pearson framework. ***Note*** that there are many readable accounts that inform one of these methodological considerations...it seems that most folks just don't want to take the time.

What Researchers Could Be About

Probably the most important task for the data modeler is to make sense out of what the fitted model(s) communicates, in light of the semantic, theoretical framework that one has provisionally adopted prior to the model development stage. With an eye toward our best inferential model, we should be attempting to reduce bias, optimize predictive validity, and (when realistically possible) increase the interpretability of the fitted model - the "bottom line" so to speak. No personal offense is intended toward anyone in these next statements: it is clear (to me at least) that it doesn't matter:

1) How long one has been teaching or applying disciplinary specific methodologies to model data;

2) It doesn't matter how much credentialing one has behind their name;

3) And it doesn't matter how many other esteemed people are willing to line up and tell you how
gloriously gifted and intelligent you are as a data modeler - if one ignores current methodological practice. *This seems clear to me because:*

Data Modeling as an Evolving Body of Practices

Data modeling (as a science or an art) is an evolving body of practices - much critical debate gives rise to new practices; that all conscientious researchers (modelers) contribute to by thinking thoughtfully about their data; and hopefully, subsequently share those thoughts with the WIDER community of practitioners and theoreticians. Hopefully, a truly WIDER community: ecology, epidemiology, biology, psychology, education, sociology, political science, economics, business, medical informatics, etc. To be out of touch with that changing body of practices is to be going against the grain of the current learned experiences of that wider consensus. While this is NOT NECESSARILY a bad practice, I would think that ignoring consensus should NOT be done lightly; it should NOT be done without awareness or without a worthy purpose in mind. When ignoring the experience of others, it probably goes without saying that it should not be done out of "laziness". *Here are what I* *think are some good indicators of how one might compare in relation to that WIDER community:*

How A Researcher Might Compare to the Wider Methodological Community

IF, one is "of the mind" or "practicing" the following:

1) REFLEXIVELY utilizing standard fare null hypothesis significance tests as presented by the bulk of introductory applied statistics textbooks. That is, focusing on classical-frequentist observed p-values under assumed, random influence, hypotheses (i.e. null hypothesis), as the main evidence in drawing conclusions about the data;

2) Believe that using data imputation methods for missing data is somehow "cheating";

3) NEVER use non-parametric, semi-parametric, and robustly estimated models;

4) Stick RIGIDLY to confirmatory practices while ignoring the importance of "exploratory practices" in the initial stages of model development (and I mean exploratory in: after data has been collected);

5) Think that "Data Mining", "Knowledge Discovery", etc, is somehow "beneath" serious data modelers;

6) NOT APPRECIATE how re-sampling and simulation based methods have revolutionized the practice of statistics (e.g. applications of the Bootstrap and Monte Carlo Markov Chain estimated modeling);

7) NOT APPRECIATE that a multivariate (or multi-variable) approach should be a "first choice" modeling framework that is utilized (that is only to say that it should be adopted more often) - not a univariate framework; And that a univariate framework should be the exception to the practice. Statistical models in non-experimental settings (and arguably in experimental settings as well) are only going to have external or ecological validity to the extent that complexity in the "real world" (as reflected in the data relations) is realistically taken into account. Singular T-tests and ANOVA's used in non-experimental settings, are in various ways, deficient. In other words, using univariate, mean-difference testing approaches on observational data, is a good recipe to MISS consistent, valuable patterns in one's data.

8) OVER UTILIZE (OR ONLY utilize) Classical frequentist approaches in model estimation, model comparison, and model validation (e.g. relying on BLUE theory that uses MLE estimation for models). NOT appreciate that in evaluating statistical models, that estimated "believability values" (I stop short of calling them "truth-values", can be usefully assigned to models or parameters (e.g. using probability or information-theoretic based measures to rank order or average models or model parameters parameters - e.g. Bayesian Model Averaging). From one view, one can permute the data space (create a sampling distribution), but from another view it is also useful to look at permutations of the parameter space as well - in other words, one may NOT be close to the actual "best" model, and in assuming the wrong model there can be quite a cost associated with using BLUE theory and MLE estimation to arrive at one's predictive model (bias and lack of efficiency).

9) NOT APPRECIATE the importance of Bayesian inferential logic (and other alternatives) as complimentary to, or as a replacement for classical frequentist inferential logic (e.g. using "Bayes Factors" in lieu of, or as a compliment to observed p-values under and assumed

sampling distribution; and/or using Bayesian "credible intervals" from a posterior distribution, rather than confidence intervals based on NULL sampling distributions, whenever the statistical models are based on medium to small sample sizes, and/or the possibility of choosing reasonable priors for parameters exist.

THEN:

I would suggest that one is out of touch with emerging methodological trends that are becoming evident in a number of disciplines. Methodological wisdom evolves, so must the basic pedagogical practices that communicate those evolving methods.

A Common Sentiment

Examples of a common sentiment that reflect this lack of evolution in thinking, in my experience (more often than NOT), are demonstrated by *variations on the following statement:*

"I just want to make sure that students can interpret a t-test, a correlation and a probability value, and get the interpretation of the null hypothesis correct...to be able to use confidence intervals and effect sizes correctly..."

A seemingly well informed position to have - at least an optimistic position. However, from one perspective, this position is short-sighted when judged from an awareness of the history of science, education, public policy, and the relationship amongst them. These methods are but one small part of a number of limited tools, in a larger set of decision science tools that contribute to lowering decision uncertainty, for potential actions of individuals in both a private and public arena (e.g. "Do I use drug XYZ for myself or for my family? Is genetic engineering safe - what do we mean by safe?, and safe for whom?, How can we model and predict the next pandemic outbreak?, Is global warming a real phenomenon?, how do we take measures to reverse the potentially ongoing negative impact that humans have on worldwide climatological and ecological changes?").

Our problems are complex; Our interactions with ourselves and our world are complex, so why should the decision tools that we use to deal with this complexity be neatly and narrowly circumscribed? Now for the global, cynical generality - Seems to me that for the most part, introductory statistics courses, for your generic institution, do students a disservice - we train students to expect "neatness", and "tidiness" for the sake of pedagogical closure. Student's come out of these methods courses looking for the correct formula to "turn the crank on"; look for that software button to push to provide the expected answer. We inspire algorithmic thinking in the pursuit of credentialing...so that nowadays, it seems that *critical thinking* is one of those obvious decision science tools that has NOT been taught and is in sparing use.

An Alternative Sentiment

Consider the following statement as a potential alternative sentiment:

"I want students to be able to think critically, creatively, and substantially about data in a way where their understanding is not led astray by the singular inferential framework and methodology that happened to be adopted. To understand that in the end, what is wanted by most, are helpful suggestions as to which optimal decisions can be made about important, uncertain, future, events that occur in each of our lives. That, at the end of the data modeling process, the specifics of certain, select statistical models, are mostly beside the point. Whereas, the generalities of the statistical models, taken as a whole, can and often do

provide a larger range of useful solutions for resolving decision uncertainty. Furthermore, I want students to appreciate that a pluralistic approach to inference is a real strength, bordering on mandatory, and that picking only one inferential framework as a "lens" to the data is an impoverished strategy (possible lenses: Classical Frequentist based inference, Information Theoretic and Likelihood based inference; Bayesian inference; Algorithmic and Set-theoretic based approaches - e.g. Data Mining, Machine Learning and AI approaches). In other words, I want students to recognize the potential danger in allowing the modeling technique, by its very epistemological nature, to create a narrow (possibly biased) view of the data. Similarly, I want students to understand that it is important to NOT pick the question just so as to allow for the convenience of using, in an unthinking way, a singular, default inferential framework - I suppose one could put this more colorfully as: 'There is a real danger in letting the tail wag the dog' ".

Side note: I offer the following, much seen example, as evidence of the "tail wagging the dog" phenomenon: using the median to create groups from continuous data whereby mean differences are statistically tested using hypothesis tests using the classical frequentist logic - forcing what is regression with continuous data to be data that is convenient for an ANOVA framework.

In the End, There Are Just More Questions

"Lastly, I want students appreciate that truth lies in paradox, and that one way to get to the heart of paradox is to critically examine assumptions - one doesn't do this by avoiding questioning for the sake of neatness - for the sake of pedagogy - for the sake of progress. In the end, we (researchers, citizens of our respective countries, one species among many on planet Earth) have NOT fulfilled our better 'nature', if we are not left with a sense of awe, mystery and curiosity - if we are not left with more questions."

My Short Reply To The Student

All in all, my short reply to the student's question was:

"No, two hours is not enough time to finish your data modeling/analysis project. How about a day?" (note that I am being somewhat sarcastic here....I really believe it takes much longer; a day is really rushing the process, in my opinion :-)

I would love to hear other views on these research and statistical matters. This current column is a "cleaned-up" or revised version (hopefully for the better!), of a previously published entry in the web blog for the RSS group. Comments on this current column can be posted at:

https://web2survey.unt.edu/RSS-Blogs/7#comments

* Please note that the **opinions and information** expressed **herein** do not necessarily reflect those of UNT or my colleagues within the RSS group!

References

***Note:* I do not consider this reference list necessarily representative or complete; this list is composed of references that I found motivating, enlightening, informative, or just plain entertaining to read. I have made no attempt to organize this list thematically or by importance. I provide this list so that readers have access to some of the influences on my thinking.**

Some Statistical Heresies, J. K. Lindsey, The Statistician, Vol. 48, No. 1 (1999), pp. 1-40.

Avoiding Statistical Pitfalls, Christopher Chatfield, Statistical Science, Vol. 6, No. 3 (Aug., 1991), pp. 240-252.

Data Mining: Statistics and More?, David J. Hand, The American Statistician, Vol. 52, No. 2. (May, 1998), pp. 112-118.

Statistical Modeling: The Two Cultures, Leo Breiman, Statistical Science, Vol. 16, No. 3. (Aug., 2001), pp. 199-215.

Confessions of a Pragmatic Statistician, Chris Chatfield, The Statistician, Vol. 51, No. 1. (2002), pp. 1-20.

Model Uncertainty, Data Mining and Statistical Inference, Chris Chatfield, Journal of the Royal Statistical Society. Series A (Statistics in Society), Vol. 158, No. 3 (1995), pp. 419-466.

Teaching a Course in Applied Statistics, C. Chatfield, Applied Statistics, Vol. 31, No. 3. (1982), pp. 272-289.

Some General Aspects of the Theory of Statistics, D. R. Cox, International Statistical Review , Vol. 54, No. 2. (Aug., 1986), pp. 117-126.

The Current Position of Statistics: A Personal View, D. R. Cox, International Statistical Review, Vol. 65, No. 3 (Dec., 1997), pp. 261-276.

Role of Models in Statistical Analysis, D. R. Cox, Statistical Science, Vol. 5, No. 2. (May, 1990), pp. 169-174.

Second Thoughts on the Bootstrap, Bradley Efron, Statistical Science, Vol. 18, No. 2, Silver Anniversary of the Bootstrap. (May, 2003), pp. 135-140.

The Impact of the Bootstrap - Bradley Efron: A Conversation with Good Friends Susan Holmes; Carl Morris; Rob Tibshirani; Bradley Efron, Statistical Science, Vol. 18, No. 2, Silver Anniversary of the Bootstrap. (May, 2003), pp. 268-281.

Bayesian Measures of Model Complexity and Fit, David J. Spiegelhalter; Nicola G. Best; Bradley P. Carlin; Angelika van der Linde, Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 64, No. 4. (2002), pp. 583-639.

Bayesian Approaches to Randomized Trials, David J. Spiegelhalter; Laurence S. Freedman; Mahesh K. B. Parmar, Journal of the Royal Statistical Society. Series A (Statistics in Society), Vol. 157, No. 3. (1994), pp. 357-416.

Relationships among Sample Size, Model Selection and Likelihood Regions, and Scientifically Important Differences, J. K. Lindsey, The Statistician, Vol. 48, No. 3 (1999), pp. 401-411.

Markov Chain Monte Carlo: 10 Years and Still Running! Olivier Cappe; Christian P. Robert, Journal of the American Statistical Association, Vol. 95, No. 452. (Dec., 2000), pp. 1282-1286.

Bayesian Analysis: A Look at Today and Thoughts of Tomorrow, James O. Berger, Journal of the American Statistical Association, Vol. 95, No. 452. (Dec., 2000), pp. 1269-1276.

Could Fisher, Jeffreys and Neyman Have Agreed on Testing?, James O. Berger, Statistical Science, Vol. 18, No. 1. (Feb., 2003), pp. 1-12.

An Introduction to Empirical Bayes Data Analysis, George Casella, The American Statistician, Vol. 39, No. 2 (May, 1985), pp. 83-87.

Explaining the Gibbs Sampler, George Casella; Edward I. George, The American Statistician, Vol. 46, No. 3. (Aug., 1992), pp. 167-174.

Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence, James O. Berger; Thomas Sellke, Journal of the American Statistical Association, Vol. 82, No. 397. (Mar., 1987), pp. 112-122.

Frequentist Post-Data Inference, Constantinos Goutis; George Casella, International Statistical Review, Vol. 63, No. 3. (Dec., 1995), pp. 325-344.

A Bayesian Perspective on the Bonferroni Adjustment, Peter H. Westfall; Wesley O. Johnson; Jessica M. Utts, Biometrika, Vol. 84, No. 2. (Jun., 1997), pp. 419-427.

Calibration of p Values for Testing Precise Null Hypotheses, Thomas Sellke; M. J. Bayarri; James O. Berger, The American Statistician, Vol. 55, No. 1. (Feb., 2001), pp. 62-71.

P Values: What They Are and What They Are Not, Mark J. Schervish, The American Statistician, Vol. 50, No. 3. (Aug., 1996), pp. 203-206.

The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis, John M. Hoenig; Dennis M. Heisey, The American Statistician, Vol. 55, No. 1. (Feb., 2001), pp. 19-24.

Statistical Significance Tests: Equivalence and Reverse Tests Should Reduce Misinterpretation, David F. Parkhurst, BioScience, Vol. 51, No. 12. (Dec., 2001), pp. 1051-1057.

Bayesian Inference for Comparative Research, Bruce Western; Simon Jackman, The American Political Science Review, Vol. 88, No. 2 (Jun., 1994), pp. 412-423.

Statistical Inference for Apparent Populations, Richard A. Berk; Bruce Western; Robert E. Weiss, Sociological Methodology, Vol. 25 (1995), pp. 421-458.

Statistical Thinking in Empirical Enquiry, C. J. Wild; M. Pfannkuch, International Statistical Review, Vol. 67, No. 3. (Dec., 1999), pp. 223-248.

The Use of Resampling Methods to Simplify Regression Models in Medical Statistics, Willi Sauerbrei, Applied Statistics, Vol. 48, No. 3. (1999), pp. 313-329.

Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors, Frank E. Harrell Jr., Kerry L. Lee, Daniel B. Mark, Statistics in Medicine, Volume 15, Issue 4, February 1996, pp 361-387.

Why Isn't Everyone a Bayesian?, B. Efron, The American Statistician, Vol. 40, No. 1 (Feb., 1986), pp. 1-5.

The Bootstrap and Modern Statistics, Bradley Efron, Journal of the American Statistical Association, Vol. 95, No. 452. (Dec., 2000), pp. 1293-1296.

The Variable Selection Problem, Edward I. George. Journal of the American Statistical Association ,Vol. 95, No. 452 (Dec., 2000), pp. 1304-1308.

Statistics and Mathematics - Trouble at the Interface? P. Sprent, Journal of the Royal Statistical Society: Series D (The Statistician) 47 (2), 239?244.

The Unity and Diversity of Probability, Glenn Shafer, Statistical Science, Vol. 5, No. 4. (Nov., 1990), pp. 435-444.

R. A. Fisher in the 21st Century, Bradley Efron, Statistical Science, Vol. 13, No. 2 (May, 1998), pp. 95-114.

Controversies in the Foundations of Statistics, Bradley Efron, The American Mathematical Monthly, Vol. 85, No. 4 (Apr., 1978), pp. 231-246

Markov Chain Monte Carlo Method and Its Application, Stephen P. Brooks, The Statistician, Vol. 47, No. 1 (1998), pp. 69-100.

Bayesian Computation: A Statistical Revolution, Stephen P. Brooks, Philosophical Transactions: Mathematical, Physical and Engineering Sciences, Vol. 361, No. 1813, Mathematics, Physics and Engineering (Dec., 2003), pp. 2681-2697.

Originally published, July 2007 -- Please note that information published in *Benchmarks Online* is likely to degrade over time, especially links to various Websites. To make sure you have the most current information on a specific topic, it may be best to search the UNT Website - http://www.unt.edu . You can also search ***Benchmarks Online*** - http://www.unt.edu/benchmarks/archives/back.htm as well as consult the UNT Helpdesk - http://www.unt.edu/helpdesk/
Questions and comments should be directed to benchmarks@unt.edu

Return to top