NOTES FOR SUMMER STATISTICS INSTITUTE COURSE

**COMMON MISTAKES IN STATISTICS –
SPOTTING THEM AND AVOIDING THEM**

Day 4: Common Mistakes Based on Common Misunderstandings about Statistical Inference

MAY 22 – 25, 2016

Instructor: Mary Parker

Slides and Appendix for Day 4 as prepared by Martha K. Smith
for course in 2016

CONTENTS OF DAY 4

II. MULTIPLE INFERENCE

"Recognize that any frequentist statistical test has a random chance of indicating significance when it is not really present. Running multiple tests on the same data set at the same stage of an analysis increases the chance of obtaining at least one invalid result. Selecting the one "significant" result from a multiplicity of parallel tests poses a grave risk of an incorrect conclusion. Failure to disclose the full extent of tests and their results in such a case would be highly misleading."

Professionalism Guideline 8, *Ethical Guidelines for Statistical Practice*, American Statistical Association, 1997

Performing more than one statistical inference procedure on the same data set is called ***multiple inference***, or ***joint inference***, or ***simultaneous inference***, or ***multiple testing***, or ***multiple comparisons***, or ***the problem of multiplicity***.

Performing multiple inference using frequentist methods *without considering the implications for Type I error* is a **common error** in research using statistics.

- For example, A. M. Strasak et al (2007) examined all papers from 2004 issues of the *New England Journal of Medicine* and *Nature Medicine* and found that 32.3% of those from *NEJM* and 27.3% from *Nature Medicine* were "Missing discussion of the problem of multiple significance testing if occurred."
- These two journals are considered the top journals (according to impact figure) in clinical science and in research and experimental medicine, respectively.

The Problem

Recall: If you perform a hypothesis test using a certain significance level (we'll use 0.05 for illustration), and if you obtain a p-value less than 0.05, then there are *three possibilities*:

1. The model assumptions for the hypothesis test are not satisfied in the context of your data.
2. The null hypothesis is false.
3. Your sample happens to be one of the 5% of samples satisfying the appropriate model conditions for which the hypothesis test gives you a Type I error – i.e., you falsely reject the null hypothesis.

Now suppose you're performing *two* hypothesis tests, *using the same data* for both.

- Suppose that in fact all model assumptions are satisfied and *both* null hypotheses are true.
- *There is in general no reason to believe that the samples giving a Type I error for one test will also give a Type I error for the other test.*
- See Jerry Dallal's Simulation (http://www.jerrydallal.com/LHSP/multtest.htm; linked with instructions on course home pate)
- This motivates considering the *joint Type I error rate*

Joint Type I error rate: This is the probability that a randomly chosen sample (of the given size, satisfying the appropriate model assumptions) will give a Type I error for *at least one* of the hypothesis tests performed.

The joint Type I error rate is also known as the ***overall Type I error rate***, or ***joint significance level***, or the ***simultaneous Type I error rate***, or the ***family-wise error rate*** (***FWER***), or the ***experiment-wise error rate***, etc.

- The acronym FWER is becoming more and more common, so will be used in the sequel, often along with another name for the concept as well.

Examples of common mistakes involving multiple inference:

1. An *especially serious* form of neglect of the problem of multiple inference is the one alluded to in the quote from the ASA ethics page:
   - Trying several tests and reporting just one significant test, without disclosing how many tests were performed or correcting the significance level to take into account the multiple inference.
   - *Don't do it!*
   - To help you remember: Think Jelly Beans, http://xkcd.com/882/
   - To help drive home the message, see more of Jerry Dallal's simulations:
     - http://www.jerrydallal.com/LHSP/jellybean.htm
     - http://www.jerrydallal.com/LHSP/cellphone.htm
     - http://www.jerrydallal.com/LHSP/coffee.htm

2. Some textbooks and software packages advise using a hypothesis test for equal variance before using a hypothesis test that has equal variance as a model assumption (e.g., equal variance two-sample t-test; standard ANOVA test).
   - This can produce misleading results two ways
     - First, either test could produce Type I errors.
     - But the sequential use of the tests may lead to more misleading results than just the use of two tests.
   - Zimmerman (2004) discusses this in more detail.

Multiple inference with confidence intervals

The problem of multiple inference also occurs for confidence intervals.

- In this case, we need to focus on the *confidence level*.
- *Recall*: A 95% confidence interval is an interval obtained by using a procedure that, for 95% of all suitably random samples, of the given size, from the random variable and population of interest, produces an interval containing the parameter we are estimating (assuming the model assumptions are satisfied).
- In other words, the procedure does what we want (i.e. gives an interval containing the true value of the parameter) for 95% of suitable samples.
- *If we're using confidence intervals to estimate two parameters, there's no reason to believe that the 95% of samples for which the procedure "works" for one parameter* (i.e. gives an interval containing the true value of the parameter) *will be the same as the 95% of samples for which the procedure "works" for the other parameter*.
- If we're calculating confidence intervals for more than one parameter, we can talk about the ***joint (***or ***overall*** *or* ***simultaneous*** *or* ***family-wise or experiment-wise) confidence level***.
- For example, a group of confidence intervals (for different parameters) has an ***overall 95% confidence level*** (or ***95% family-wise confidence level***, etc.) if the intervals are calculated using a procedure which, for 95% of all suitably random samples, of the given size from the population of interest, produces for *each* parameter in the group an interval containing that parameter (assuming the model assumptions are satisfied).

What to do about multiple inference

Unfortunately, *there is not (and can't be) a simple formula to cover all cases:*

- Depending on the context, the samples giving Type I errors for two tests might be the same, they might have no overlap, or they could be somewhere in between – and we can't know which might be the case.

- Various techniques for bounding the FWER (joint Type I error rate) have been devised for various special circumstances.
  - Some will be discussed below.

- There are also alternatives to considering FWER.
  - Some of these will be discussed below.

- For more information on other methods for specialized situations, see, e.g., Hochberg and Tamhane (1987) and Miller (1981)

- See Efron (2010) for both an account of the history (Chapter 3) of the subject and discussion of some somewhat more recent developments in dealing with multiple inference, especially in large data sets.

Bonferroni method:

Fairly basic probability calculations show that *if the sum of the individual Type I error rates for different tests is ≤ α, then the overall ("family-wise") Type I error rate (FWER) for the combined tests will be ≤ α.*

- For example, if you're performing five hypothesis tests and would like an FWER (overall significance level) of at most 0.05, then using significance level 0.01 for *each* test will give an FWER (overall significance level) of at most 0.05.
- Similar calculations will show that if you're finding confidence intervals for five parameters and want an overall confidence level of 95%, using the 99% confidence level for each confidence interval will give you overall confidence level at least 95%. (Think of confidence level as 1 - *α*.)

The Bonferroni method can be a used as a fallback method when no other method is known to apply.

- However, if a method that applies to the specific situation is available, it will often be better (less conservative; have higher power) than the Bonferroni method, so calculate by both methods and compare.
- Holm's procedure (which depends on the Bonferroni idea, but in a more sophisticated way) is a relatively easy (e.g., on a spreadsheet) method that gives higher power than the basic Bonferroni method. It's described various places on the web – e.g., http://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method

The basic Bonferroni method is also useful for dividing up the overall Type I error between different types of inference.

- *Example*: If three confidence intervals and two hypothesis tests are planned, and an overall Type I error rate of .05 is desired, then using 99% confidence intervals and individual significance rates .01 for the hypothesis tests will achieve this.
- This method can also be used to apportion Type I error rate between *pre-planned inference* and *post-hoc inference*
    - *pre-planned inference*: the inferences planned as part of the design of the study
    - *post-hoc inference:* the inferences based on looking at the data and noticing other things of interest.
        - These are also called "data-snooping" – more on this tomorrow.
    - Example: If you plan 3 hypothesis tests, but might decide later to do more, you could plan to do the three "preplanned" hypothesis tests each at significance level .01, leaving .02 to divide between the data-snooping hypothesis tests
- However, *this apportioning should be done before analyzing the data*.

Whichever method is used, *it's important to make the calculations based on the number of tests that have been done, not just the number that are reported.*

- Remember Jelly Beans!

False discovery rate:

An alternative to bounding Type I error was introduced by Benjamini and Hochberg (1995): bounding the *False Discovery Rate*.

> The ***False Discovery Rate*** (FDR) of a group of tests is the *expected value of the ratio of falsely rejected hypotheses to all rejected hypotheses*.

("Expected value" refers to the mean of a distribution. Here, the distribution is the sampling distribution of the ratio of falsely rejected hypotheses to all rejected hypotheses tested.)

Note:

- The family-wise error rate (FWER) focuses on the possibility of making *any* Type I error among all the inferences performed.
- The false discovery rate (FDR) tells you what *proportion* of the *rejected* null hypotheses are, *on average*, really false.
- Bounding the FDR rather than the FWER may be a more reasonable choice when many inferences are performed, especially if there is little expectation of harm from falsely rejecting a null hypothesis.
- Thus it's increasingly being adopted in areas such as micro-array gene expression experiments or neuro-imaging.
- However, these may involve variations rather than the original definition given above; see Efron (2010) for more details.

As with the FWER, there are various methods of actually bounding the false discovery rate.

- For the original false discovery rate, see Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), and Benjamini and Yekutieli (2005)
- For variations of false discovery rate, see Efron (2010).

Higher Criticism (HC)

This is another alternative to bounding Type I error that's sometimes used in situations such as genome-wide testing.

- See Donoho and Jin (2015) for a review article on HC.
- See Klaus and Strimmer (2013) for a discussion of connections between HC and FDR.

Random Field Theory (RFT)

- This method is used in functional imaging data to try to account for spatial correlation when performing multiple hypothesis tests.
- See http://biostatistics.oxfordjournals.org/content/14/1/129

Randomization methods

- Lee and Rubin (2016) report a randomization-based method that can be helpful in some post-hoc subgroup analyses.

Multilevel Modeling

Gelman et al (2012) have proposed that in some cases multilevel modeling is a better way to address multiple inference than frequentist methods

- They point out that methods such as Bonferroni corrections have unfortunate side effects:
  - They give large interval estimates for effects.
  - Since they require smaller cutoffs for significance, they are likely to produce Type M errors (because of The Winner's Curse).
- Multilevel modeling uses a different approach to inference that typically produces both smaller interval estimates, and more moderate point estimates of effects than standard frequentist methods, so may be a better way to approach multiple inference.

Bayesian Methods

Bayesian methods of statistical inference can be used to obtain a "joint posterior distribution" for several parameters of interest.

- Bayesian methods are not frequentist, so the problem with Type I errors does not apply (although Type M and S errors are still relevant.)
- They still require careful modeling, both of the *prior* and of the *likelihood model*.

Subtleties and controversies

Bounding the overall Type I error rate (FWER) will reduce the power of the tests, compared to using individual Type I error rates.

- Some researchers use this as an argument against multiple inference procedures.
- The counterargument is the argument for multiple inference procedures to begin with: Neglecting them will produce excessive numbers of false findings, *so that the "power" as calculated from single tests is misleading*.
  - See Maxwell and Kelley (2011) and Maxwell (2004) for more details.
- Bounding the False Discovery Rate (FDR) will usually give higher power than bounding the overall Type I error rate (FWER).

Consequently, it's important to consider the particular circumstances, as in considering both Type I and Type II errors in deciding significance levels.

- *In particular, it's important to consider the consequences of each type of error in the context of the particular research.*

Examples:

1. A research lab is using hypothesis tests to screen genes for possible candidates that may contribute to certain diseases.
    - Each gene identified as a possible candidate will undergo further testing. The results of the initial screening are not to be published except in conjunction with the results of the secondary testing,
        - Case I: If the secondary screening is inexpensive enough that many second level tests can be run, then the researchers could reasonably decide to ignore overall Type I error in the initial screening tests, since there would be no harm or excessive expense in having a high Type I error rate.
        - Case II: If the secondary tests were expensive, the researchers would reasonably decide to bound either family-wise Type I error rate or False Discovery Rate.

2. Consider a variation of the situation in Example 1:
    - The researchers are using hypothesis tests to screen genes as in Example 1, but plan to publish the results of the screening *without* doing secondary testing of the candidates identified.
    - In this situation, ethical considerations warrant bounding either the FWER or the FDR -- *and* taking pains to emphasize in the published report that these results are just of a preliminary screening for possible candidates, and that these preliminary findings need to be confirmed by further testing.

The Bottom Line: No method of accounting for multiple inference is perfect, which is one more reason why ***replication of studies is important!***

Note: For more discussion of multiple inference in exploratory research, see Goeman and Solari plus discussion (2011).

III. DATA SNOOPING

Remember Jelly Beans**:** http://xkcd.com/882/

Data snooping refers to statistical inference that the researcher decides to perform *after* looking at the data

- Also known as *post protocol analysis* or *post hoc analysis*
- Contrast with *pre-planned* inference ("*per protocol analysis*"), which the researcher plan has planned *before* looking at the data.

Data snooping can be done:

- professionally and ethically, or
- misleadingly and unethically, or
- misleadingly out of ignorance.

Misleading data snooping out of ignorance is a **common mistake** in using statistics.

The problems with data snooping are essentially the problems of multiple inference.

- So if you're likely to engage in data snooping frequentist inference, plan to allocate some part of the overall Type I error rate to pre-planned inference and some part to data snooping.
  - For example, if you plan to have overall Type I error rate (FWER) 0.05, you might decide to use FWER 0.04 for pre-planned inference, and FWER 0.01 for data snooping.
- One way in which researchers unintentionally obtain misleading results by data snooping is in *failing to account for all of the data snooping they engage in.*
  - In particular, *in accounting for Type I error when data snooping, you need to count not just the actual hypothesis tests performed, but also all comparisons looked at when deciding which post hoc (i.e., not pre-planned) hypothesis tests to try.*

For lots of amusing examples, see Tyler Vigen's website http://tylervigen.com/.

- The original version of the site allowed you to choose two variables from a very large list and find their correlation.
- I got tired counting at several hundred, but I'd guess that he had over 1000 variables listed.
- That makes around 1,000,000 pairs of variables.
- If you did significance tests (at a .05 individual significance rate) for correlation for all those pairs, you would expect about 50,000 to be significant – so it shouldn't be surprising if many of these 50,000 pairs are indeed highly correlated.

A More Serious Example: A group of researchers plans to compare three dosages of a drug in a clinical trial.

- There's no pre-planned intent to compare effects broken down by sex, but the sex of the subjects is recorded.
- The researchers have decided to have an overall Type I error rate of 0.05, allowing 0.03 for the pre-planned inferences and 0.02 for any data snooping they might decide to do.
- The pre-planned comparisons show no statistically significant difference between the three dosages when the data are not broken down by sex.
- However, since the sex of the patients is known, the researchers decide to look at the outcomes broken down by combination of sex and dosage.
  - They notice that the results for women in the high-dosage group look much better than the results for the men in the low dosage group, and decide to perform a hypothesis test to check that out.
- *In accounting for Type I error, the researchers need to take the number of data-snooping inferences performed as 15, not one.*
  - The reason : *They've looked at fifteen comparisons* -- there are 3×2 = 6 dosage×sex combinations, and hence (6×5)/2 = 15 pairs of dosage×sex combinations.
  - Thus the significance level for the post hoc test should not be 0.02, but (if the Bonferroni method is used) 0.02/15.

See the Appendix for more detailed suggestions on data snooping professionally and ethically.

IV: P-HACKING, THE REPLICABILITY CRISIS, P-CURVING, AND "THE GARDEN OF FORKING PATHS"

P-hacking and the replicability crisis:

Simonsohn et al (2013) introduced the term ***p-hacking*** to refer to a common practice that involves data snooping, outcome switching, and aspects of the file-drawer problem:

> *Performing many hypothesis tests in analyzing the data for a study, but when publishing the results of the study, omitting mention of those tests that were not statistically significant.*

- So in p-hacking, researchers don't relegate entire studies to "the file-drawer" -- just parts of studies.

P-hacking (like may other common mistakes discussed here) contributes to what has become known as the ***replicability crisis***:

> *The large number of published "findings" that have never been confirmed by a follow-up study.*

- Many such results might indeed be "irreproducible results."
- Ioannidis' paper, "Why Most Published Research Findings Are False," (Ioannidis 2005) brought widespread attention to the replicability crisis.
- Although there was initial skepticism and criticism of Ioannidis' claims, scientists have increasingly been recognizing the lack of replications, and the practices contributing to this, as a serious problem.
    - See, e.g., Pashler and Harris (2012), and the examples given in Day 3 Notes under Underpowered Studies, Type M and S Errors, Using the same sample size for a replication, and The File Drawer Problem

There are many ways to p-hack. Some ways fall under the category of data snooping. These include:

- Collecting data until a statistically significant result is obtained.
    - Why is this a problem?

- Deciding to exclude outliers on the basis of whether or not doing so will give a statistically significant result.
    - Why is this a problem?

- Trying out more than one measure of a quantity of interest, then selecting one that gives statistical significance when others do not.
    - Why is this a problem?

- First trying an analysis without breaking down into subgroups, then if results are not statistically significant, analyzing the data broken down into subgroups (e.g., gender), but reporting only the statistically significant results.
    - Why is this a problem?

- Trying various methods of "binning" (discussed below) until getting one that gives a statistically significant result.

Like data-snooping, *p-hacking is often done out of ignorance that it gives deceptive results.*

- There's also a gray area/slippery slope where researchers feel impelled to "make the most" of their data.
  - This can also lead to "spinning," which might also include describing results that are not statistically significant as "promising," or results that are questionably practically significant as "strong" rather than "modest."
- For a real example of p-hacking in cancer research, plus discussion of spinning and the file drawer problem, see Couzin-Frankel (2013)

Contrived example: The course description for this SSI course included the sentence,

> "In 2011, psychologists Simmons, Nelson and Simonsohn brought further attention to this topic by using methods common in their field to "show" that people were almost 1.5 years younger after listening to one piece of music than after listening to another."

Some of the things these authors did to produce this nonsensical conclusion:

- Lots of data snooping.
  - In particular, they gathered information on several covariates, but adjusted for only one (father's age), in the "report".
- Lack of transparency in reporting results.
  - In particular, not mentioning that they had gathered the information on other covariates and cherry-picked the that gave the result they wanted.
- The sample size was not set in advance.
  - There was no consideration of power in deciding on sample size.
  - Instead, the researchers checked every few observations and stopped when the results reached a preset significance level.
  - The sample size was too small to give reasonable power.
- There was no adjusting for multiple testing despite all the multiple inference involved in data snooping and in deciding when to stop sampling.

***Caution**:* Although the Simmons et al paper did a good job of making the point that common but questionable practices can lead to absurd results, the author's recommendations for better practices fall short of what is needed. (See http://www.ma.utexas.edu/blogs/mks/2013/01/09/a-mixed-bag/ for more discussion.)

P-curving

Simonsohn et al (2013) have proposed a method, called p-curving, to help detect the presence of p-hacking.

- The purpose of p-curving is [to try] "to rule out selective reporting as a likely explanation for a set of statistically significant findings." (p. 5) – just as the purpose of significance testing is [to try] "to rule out chance as a likely explanation for an observed effect" (p. 5)

- A *p-curve* is "the distribution of statistically significant p-values for a set of independent findings" (p. 3)

- The utility of p-curves depends on results in mathematical statistics saying that a p-curve will have a different shape when the null hypothesis is false than when the null hypothesis is true, and that the shape will also depend on effect size and sample size.

    - The net result is that p-hacking will produce alterations in the shape of the p-curve.

- The authors have also produced an online app and user's guide at http://www.p-curve.com/

- The technique appears to have prompted a fair amount of discussion and self-questioning among psychologists.

    - You might want to do your own web search on the topic

- However, use of the technique has been overzealous in some cases, accusing people of deliberate malfeasance when they were acting out of ignorance.

The garden of forking paths

Gelman and Loken (2013 and 2014b) introduced the metaphor "Garden of Forking Paths" to refer to the many branching choices researchers can make when analyzing their data.

- They point out that there are many possible choices that researchers can make in analyzing data *that seem reasonable yet might be influenced by the data*.
- Thus different data sets analyzed for the same question might reasonably lead to different choices.
- Thus the metaphor "garden of forking paths" (from the short story by Jorge Luis Borges)
- These choices are *often not made deliberately to "game the system*.
- Thus terms such as "fishing" or "p-hacking," which suggest deliberate acts, are often falsely accusatory.
- Nonetheless, the fact that decisions are contingent on the data means that calculated p-values are not meaningful.
- At the same time, studying the data to find out patterns and make tentative analysis decisions can be of value in understanding the problems being studied.
- Machine learning methods are usually highly data-dependent.
- Bayesian methods also encounter the garden of forking paths.

- Meta-analysis attempts to use results from several studies to obtain better understanding.
    - However, meta-analysis is subject to the same problems as any statistical analysis, and can be biased by not accounting for bias in studies included.
    - Example: Jauhar et al (2014) point out problems with previous meta-analyses of studies of cognitive behavior therapy for schizophrenia, and provide a new one taking into account sources of bias in the studies included – with differing results.
- The metaphor of the garden of forking paths can also apply to analysis decisions that are not entirely data dependent.

  Example: Silberzahn and Uhlman (2015) asked 29 research teams to answer the same research question ("are football (soccer) referees more likely to give red cards to players with dark skin than to players with light skin?") with the same data set. The results obtained by the different teams are summarized in the graph at http://www.nature.com/news/crowdsourced-research-many-hands-make-tight-work-1.18508#/b3
- Bernau et al (2014) introduces a possibly promising method of dealing with some of the problems of data-dependence, different methods of study, and weaknesses of meta-analysis. They produced a combined data set of gene expression and ovarian cancer survival based on 10 different data sets. They also identified from the literature 14 models for predicting patient survival from gene expression, and did cross-validation of models across data-sets. The effort required was huge. [See Donoho (2015) p. 30 for a more detailed summary.]

- Ioannidis (2008) introduced the term “vibration of effects” to refer to the variability of effect sizes that can result from the different analysis choices for the same data set (i.e., from the garden of forking paths).
    - He shows results of a simulation where some analysis choices produced effect sizes twice as large as others.
    - This provides another reason (in addition to The Winner’s Curse) why calculated effect sizes (as well as p-values) can be misleading

The bottom line: Replication is really important!

Gelman and Loken (2013 and 2014b) also discuss possible (at least partial) solutions:

- Preregistration works for some fields.
    - Caution: Some preregistration sites defeat their purpose by changing analysis details after the fact; see http://andrewgelman.com/2015/10/10/doomed-to-fail-a-pre-registration-site-for-parapsychology/#comment-246794 for some elaboration.

- *Real Example:*
  Psychologists Brian Nosek and Matt Motyl obtained a statistically significant (p = 0.01) result with sample size N = 1,979.
    - However, before publishing their findings, they decided to do a replication study.
    - They did a power analysis and determined that a sample size of 1300 would give power .995 to detect the effect size found in the original study at significance level .05.
    - The replication study gave p = .59.
    - See Nosek et al (2012) for details.

- An exploratory study followed by pre-publication replication can work well in some situations (e.g., the Nosek et al study mentioned yesterday.)

- In areas where most data is observational, Gelman and Loken recommend full study of the data despite the problem of multiplicities.
    - In some cases, multilevel modeling can help.

- Researchers need to distinguish carefully between exploratory and confirmatory data analysis and be aware of the value and limitations of each.

- More research is needed into how to handle the problem of multiple comparisons, particularly in light of the garden of forking paths and vibration of effects.

    - One proposal to help is to do a "mock" analysis and report, using simulated data, before initiating collecting or analyzing real data. See Humphreys et at (2013) and Gelman

Impediments to quality replication

Ioannidis (2012) points out several problems that may occur with replications:

- In some fields, current customs may make it difficult to get funding for replications or to get replications published.
- In some cases, replications done by the same research team as the original study may be influenced by the researchers' belief in the findings of the original study, or may have the same weak spots as the original study.
- Even replications done by different research teams may be influenced by the original research team. For example,
    - The replication team may believe that the purpose of replication is to confirm the original results ("obedient replication"); the garden of forking paths often makes this possible.
    - The original research team may have the influence to determine which replications do or do not get published, their interpretation, etc. ("obliged replication")

Ioannidis offers possible means to help reduce these problems:

- Replications involving several teams and researchers, preferably including those endorsing different theories.
- Using pre-specified protocols and analyses
- Fully documenting the reasoning and analysis
- Making raw data and analysis code available
- Post-publication review.

V: USING AN INAPPROPRIATE METHOD OF ANALYSIS

"Assumptions behind models are rarely articulated, let alone defended. The problem is exacerbated because journals tend to favor a mild degree of novelty in statistical procedures. Modeling, the search for significance, the preference for novelty, and the lack of interest in assumptions -- these norms are likely to generate a flood of nonreproducible results."

David Freedman, *Chance* 2008, v. 21 No 1, p. 60

Recall: Each frequentist inference technique (hypothesis test or confidence interval) involves *model assumptions.*

- Different techniques have different model assumptions.

- *The validity of the technique depends* (to varying extents) on whether or not the model assumptions are true for the context of the data being analyzed.

- Many techniques are *robust* to departures from at least some model assumptions.
  - This means that if the particular assumption is not too far from true, then the technique is still approximately valid.
  - Illustration: Rice Virtual Lab in Statistics Robustness Simulation

Thus, when using a statistical technique, it's important to ask:

- What are the model assumptions for that technique?
- Is the technique robust to some departures from the model assumptions?
- What reason is there to believe that the model assumptions (or something close enough, if the technique is robust) are true for the situation being studied?

Neglecting these questions is a* very common mistake *in using statistics.

- Sometimes researchers check only some of the assumptions, perhaps missing some of the most important ones.

Unfortunately, the model assumptions vary from technique to technique, so there are few if any general rules. One general rule of thumb, however is:

Techniques are least likely to be robust to departures from assumptions of independence.

- *Recall*: Assumptions of independence are often phrased in terms of "random sample" or "random assignment", so these are very important.
- One exception is that, for large enough populations, sampling *without* replacement is good enough, even though "independent" technically means sampling *with* replacement.
- Variance estimates depend strongly on the assumption of independence, so results can be very misleading when observations are not independent.

Note: Many techniques are most robust to violations of normality assumptions, at least if the sample size is large and the distribution is not strongly skewed or multimodal.

- This is because test statistics are often sums or linear combinations, which by “the” Central Limit Theorem are often approximately normally distributed. (See Appendix re Checking Model Assumptions)

General advice and cautions:

- You may need to look hard to find model assumptions and information about robustness!
  - For basic statistical techniques, DeVeaux, Velleman and Bock, *Statistics, Data and Models* is quite good on model assumptions and robustness.
  - For other techniques, try searching for review articles in journals such as *Statistical Science*, *The American Statistician,* or *Journal of the American Statistical Society*.
- Sometimes simulations (*if* well done) can help. For example:
  - Simulations might help decide how plausible it is that your data come from a certain distribution.
  - Simulations can sometimes help get a feel for how robust a procedure is to departures from model assumptions.
- Do *not* automatically use default settings in software.

How do I know whether or not model assumptions are satisfied?

Unfortunately, there are no one-size-fits-all methods, but here are some rough guidelines:

1. When selecting samples or dividing into treatment groups, be very careful in randomizing *according to the requirements of the method of analysis to be used*.

- Remember that “random” is not the same as “haphazard”!
- Be careful to check the precise randomizing assumptions of the study design/method of analysis you plan to use.
    - For example, there are many types of ANOVA analyses, each with its own requirements for study design, including randomization.

2. Sometimes (*but not very often!*) model assumptions can be justified plausibly by well-established facts, mathematical theorems, or theory that’s well supported by sound empirical evidence.

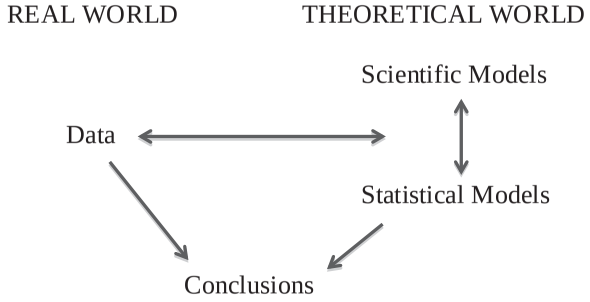
- Here, "well established" means *well established by sound empirical evidence and/or sound mathematical reasoning*.
- This is *not* the same as "well accepted," since sometimes things may be well accepted without sound evidence or reasoning.
- More in Appendix

3. Sometimes a rough idea of whether or not model assumptions might fit can be obtained by plotting the data or residuals obtained from a tentative use of the model.

- Unfortunately, these methods are typically better at telling you when the model assumption does *not* fit than when it does.
- Some examples, guidelines, and cautions are in the Appendix.
- But always remember “The Big Picture”:

Robert Kass' Big Picture of Statistical Inference

In Kass (2011, p. 6, Figure 1), Robert Kass has proposed the following diagram to depict the "big picture" in using statistics:



Points this picture is intended to show include:

- Both statistical and scientific models are abstractions, living in the "theoretical" world, as distinguished from the "real" world where data lie.
- Conclusions straddle these two worlds: *conclusions about the real world typically are indirect, via the scientific models*.
- "When we use a statistical model to make a statistical inference we implicitly assert that the variation exhibited by data is captured reasonably well by the statistical model, so that the theoretical world corresponds reasonably well to the real world." (p. 5)
- Thus "careful consideration of the connection between models and data is a core component of … the art of statistical practice…" (p. 6)

For a recent accessible discussion of problems with model assumptions in a topic of current wide interest (value-added models in education), see Wainer (2011).

VI. METHODS (AND THEIR LIMITATIONS) FOR CHECKING MODEL ASSUMPTIONS

See Appendix for some suggestions.

But bear in mind:

- These typically can help *sometimes* to see that a model is wrong, *but can't tell you if a model is right*.
- "Reality resists imitation through a model, " physicist Erwin Schroedinger, https://www.tuhh.de/rzt/rzt/it/QM/cat.html#sect6
- "All models are wrong but some are useful," statistician George Box https://en.wikipedia.org/wiki/All_models_are_wrong

VII. SOME SPECIFIC SITUATIONS WHERE MISTAKES INVOLVING MODEL ASSUMPTIONS ARE COMMON

A. Comparing groups in studies with drop-outs (Intent-to-treat analysis)

B. Using a two-sample test comparing means when cases are paired (and generalizations)

C. Not distinguishing between fixed and random factors in ANOVA

D. Analyzing data without regard to how they were collected

E. Pseudoreplication

F. Mistakes in regression

For more discussion of some inappropriate methods of analysis, see:

- References in the Appendix
- Harris et al (2009)
- The Common Mistakes in Using Statistics website at http://www.ma.utexas.edu/users/mks/statmistakes/TOC.html

A. Intent to Treat Analysis: Comparing groups when there are Dropouts

The Problem: In many forms of comparison of two treatments involving human subjects (or animals or plants), there are subjects who do not complete the treatment.

- They may die, move away, encounter life circumstances that take priority, or just decide for whatever reason to drop out of the study or not do all that they are asked.
- It's tempting to just analyze the data for those completing the protocol, essentially ignoring the dropouts. ***This is usually a serious mistake***, for two reasons:

1. In a good study, subjects should be randomized to treatment.
    - Analyzing the data for only those who complete the protocol *destroys the randomization, so that model assumptions are not satisfied.*
    - To preserve the randomization, outcomes for *all* subjects *assigned* to each group (whether or not they stick with the treatment) need to be compared. This is called **intent-to-treat** (or intention-to-treat, or ITT) analysis.

2. Intent-to-treat analysis is usually more informative for consumers of the research.

- For example, in studying two drug treatments, dropouts for reasons not related to the treatment can be expected to be, on average, roughly the same for both groups.
- But if one drug has serious side-effects that prompt patients to discontinue use, that would show up in the drop-out rate, and be important information in deciding which drug to use or recommend.

Reason 1 (and sometimes also reason 2) also applies when treatments are applied to animals, plants, or even objects.

Unfortunately, when subjects drop out of an experiment, data collection for them is incomplete.

- Thus, analysis often requires figuring out how best to deal with missing data.

For more information on intent-to-treat analysis, see Freedman (2005, pp. 5, 15), Freedman (2006), van Belle (2008, pp. 156 – 157), and Moher et al (2010)

B. Using a Two-Sample Test Comparing Means when Cases Are Paired (and similar problems)

One of the model assumptions of the two-sample t-tests for means is that the observations *between groups*, as well as within groups, are independent.

- Thus if samples are chosen so that there is some natural pairing, then the members of pairs are not independent, so the two-sample t-test is *not* appropriate.

Example 1: A random sample of heterosexual married couples is chosen. Each spouse of each pair takes a survey on marital happiness. The intent is to compare husbands' and wives' scores.

- The two-sample t-test would compare the *average* of the husband's scores with the *average* of the wives' scores.
- However, it's *not* reasonable to assume that the samples of husbands and wives are independent -- some factors influencing a particular husband's score are likely to influence his wife's score, and vice versa.
- Thus the independence assumption *between* groups for a two-sample t-test is violated.
- In this example, we can instead consider the individual differences in scores for each couple: (husband's score) - (wife's score). If the questions of interest can be expressed in terms of these differences, then we can consider using the one-sample t-test (or perhaps a non-parametric test if the model assumptions of that test are not met).

Example 2: A test is given to each subject before and after a certain treatment. (For example, a blood test before and after receiving a medical treatment; or a subject matter test before and after a lesson on that subject)

- This poses the same problem as Example 1: The "before" test results and the "after" test results for each subject are *not independent*, because they come from the same subject.
- The solution is the same: analyze the *difference in scores*.
- Example 2 is a special case of what is called *repeated measures*: some measurement is taken more than once on the same unit.
    - Because repeated measures on the same unit are not independent, the analysis of such data needs a method that takes this lack of independence into account.
    - There are various ways to do this; just which one is best depends on the particular situation.

Similar Problem: *Hierarchical (multilevel) situations may violate model assumptions of independence*

Example: Researchers are studying how well scores on a standardized eighth grade math exam predict performance on an Algebra I end-of-course exam for ninth-grade students.

- They have data from an entire school district.
- They propose to analyze it by simple linear regression.
- However, standard regression methods of inference assume that observations are uncorrelated, whereas observations from students in the same school can be expected to be correlated.
- Instead, the researchers need to use a multilevel (also called hierarchical) model that takes into account that observations from the same school may be correlated.

C. Inappropriately Designating an Effect as Fixed, Variable, or Random

In Analysis of Variance and Multilevel Modeling, there are two types of factors: *fixed effect* and *random effect*.

- *Fixed effect factors and random effect factors are analyzed differently, so it's important to classify a factor correctly.*
- *Confusing the matter further, different definitions of "fixed" and "random" effects are used by different people.*

Correct classification of a factor as fixed or random depends on

- the context of the problem,
- the questions of interest, and
- how the data are gathered, and
- the method of analysis

1. Fixed and random effects ***for Analysis of Variance***:

***Fixed effect factor* in Analysis of Variance**: Data has been gathered from *all the levels of the factor that are of interest.*

Example: The purpose of an experiment is to compare the effects of three specific dosages of a drug on the response.

- "Dosage" is the factor.
- The three specific dosages in the experiment are the levels.
- There is no intent to say anything about other dosages.
- Therefore this is a fixed factor.
- The analysis will estimate the effect of each of the three dosages.

***Random effect factor* for Analysis of Variance**:

- The factor has *many possible levels*.
- *All* possible levels are of interest.
- Only a *random sample of levels* is included in the data.
- The analysis will estimate the *variability* of effects of the factor as levels vary, but not effects of specific levels.

Example: A large manufacturer is interested in studying the effect of machine operator on the quality of the final product. The researcher selects a random sample of operators from the large number of operators at the manufacturer's factories and collects data on just these operators.

- The factor is "operator."
- Each operator is a level of the factor.
- Since interest is not just in the operators for whom data is gathered, this is a random factor.
- The analysis will *not* estimate the effect of each of the operators in the sample, but *will instead estimate the variability attributable to the factor "operator"*.

(See Appendix for more discussion)

The appropriate statistical analysis depends on whether the factor is treated as fixed or as random. That is, fixed and random effects require different models

- Consequently, inferences may be incorrect if the factor is classified inappropriately.
- Mistakes in classification are most likely to occur when more than one factor is considered in the study.

Example: Two surgical procedures are being compared.

- Patients are randomized to treatment.
- Five different surgical teams are used.
- To prevent possible confounding of treatment and surgical team, each team is trained in both procedures, and each team performs equal numbers of surgery of each of the two types.
- Since the purpose of the experiment is to compare the *procedures*, the intent is to generalize to other surgical teams.
- Thus *surgical team* should be considered as a *random factor*, not a fixed factor.

Comments:

- This example can help understand why inferences might be different for the two classifications of the factor: Asserting that there is a difference in the results of the two procedures *regardless of the surgical team* is a stronger statement than saying that there is a difference in the results of the two procedures *just for the teams in the experiment*.
- Technically, the levels of the random factor (in this case, the five surgical teams) used in the experiment should be a random sample of all possible levels.
    - In practice, this is usually impossible, so the random factor analysis is usually used if there is reason to believe that the teams used in the experiment could reasonably be a random sample of all surgical teams who might perform the procedures.
    - However, this assumption needs careful thought to avoid possible bias.
    - For example, the conclusion would be sounder if it were limited to surgical teams that were trained in both procedures in the same manner and to the same extent, and who had the same surgical experiences, as the five teams actually studied.

2. Fixed and random effects ***for Multilevel (Hierarchical) Modeling:***

In this context, definitions vary, but one common one is that a ***fixed effect*** is one that is the same for all units within the same grouping, whereas a ***random effect*** is one that is allowed to vary between units of the same grouping.

Simple example: Suppose we're using a linear model for the heights of a group of children.

- Since some children are inherently taller than others, it may be appropriate to allow different intercepts for different children.
- This would give a model

$$h_{ij} = \alpha_j + \beta A_i + \varepsilon_{ij},$$

where h_{ij} is the height of child j at age A_i.

- In this example, α is called a *random effect* and β is called a *fixed effect*.

Note:
i. In this context, *the α_j's are estimated, and we're not interested in levels other than the ones corresponding to the children in the study. This contrasts with the use of "random effect" in ANOVA.*

ii. Some people use the terminology *variable effect* or *varying effect* rather than *random effect* in this context. That helps avoid the confusion with the use of "random effect" in ANOVA.

iii. See http://andrewgelman.com/2005/01/25/why_i_dont_use/ for more detail on the various ways the terms "fixed" and "random" are used.

D. Analyzing Data without Regard to How They Were Collected

Using a two-sample t-test when observations are paired (see above) is one example of this. Here's another:

Example: [See Potcner and Kowalski (2004) for data and details.] An experiment was conducted to study the effect of two factors (pretreatment and stain) on the water resistance of wood.

- Two types of pretreatment and four types of stain were considered.
- For reasons of practicality and economy, the experiment was conducted with a *split-plot design* as follows:
  - Six entire boards were the *whole plots*.
  - One pretreatment was applied to each board, with the two pretreatments randomly assigned to the six boards (three boards per pretreatment).
  - Then each pre-treated board was cut into four smaller pieces of equal size (these were the *split-plots*).
  - The four pieces from each entire board were randomly assigned to the four stains.
  - The water resistance of each of the 24 smaller pieces was measured; this was the response variable.
- The following chart shows the p-values of the three significance tests involved if the correct split-plot analysis is used, and also if an incorrect analysis (assuming a crossed design, with the 6 treatment combinations randomly assigned to the 24 smaller pieces of wood, with 4 small pieces per treatment combination) is used.
- Note that the conclusions from the two analyses would be quite different!

p-values	**Correct (Split Plot) Analysis**	**Incorrect (Crossed Design) Analysis**
Interaction	0.231	0.782
Pretreatment	0.115	0.002
Stain	0.006	0.245

Additional lessons to learn from this example:

- If you're using data collected by someone else, *be sure to find out how it was collected; that might affect how you need to analyze it.*
- If you're making data available to others, *be sure to include a description of how the data was obtained.*

Some of the many considerations to take into account in deciding on an appropriate method of analysis include:

- The sampling or randomization method
- Whether or not there was blocking in an experimental design
- Whether factors are nested or crossed
- Whether factors are fixed or random
- Pseudoreplication (See below)

E. PSEUDOREPLICATION

The term *pseudoreplication* was coined by Hurlbert (1984,p. 187) to refer to

> "the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent."

His paper concerned ecological field experiments, but pseudoreplication can occur in other fields as well.

In this context, *replication* refers to having more than one experimental (or observational) unit with the same treatment. Each unit with the same treatment is called a *replicate*.

> *Note*: There are other uses of the word replication -- for example, repeating an entire experiment is also called replication; each repetition of the experiment is called a replicate. This meaning is related to the one given above: If each treatment in an experiment has the same number r of replicates (in the sense given above), then the experiment can be considered as r replicates (in the second sense) of an experiment where each treatment is applied to only one experimental unit.

Heffner et al (1996, p. 2558) distinguish a pseudoreplicate from a *true replicate*, which they characterize as

> "the smallest experimental unit to which a treatment is independently applied."

Most models for statistical inference require *true* replication.

- *True* replication permits the estimation of *variability within a treatment*.
- Without estimating variability within treatments, it is impossible to do statistical inference.

Illustration: Consider comparing two drugs by trying drug A on person 1 and drug B on person 2.

- Drugs typically have different effects in different people.
- So this simple experiment will give us *no* information about generalizing to people other than the two involved.
- But if we try each drug on several people, then we can obtain some information about the *variability* of each drug, and use statistical inference to gain some information on whether or not one drug might be more effective than the other on average.

True replicates are often confused with repeated measurements or with pseudoreplicates. The following illustrate some of the ways this can occur.

Examples:

1. Suppose a blood-pressure lowering drug is administered to a patient, and then the patient's blood pressure is measured twice.
   - This is a *repeated measurement*, not a replication.
   - It can give information about the *uncertainty in the measurement process*, but *not* about the *variability in the effect of the drug*.
   - On the other hand, if the drug were administered to two patients, and each patient's blood pressure was measured once, we can say *the treatment has been replicated*, and the replication might give some information about the variability in the effect of the drug.

2. A researcher is studying the effect on plant growth of different concentrations of CO_2 in the air.
   - He needs to grow the plants in a growth chamber so that the CO_2 concentration can be set.
   - He has access to only two growth chambers, but each one will hold five plants.
   - However, since the five plants in each chamber share whatever conditions are in that chamber besides the CO_2 concentration (and in fact may also influence each other), the individual plants do *not* constitute independent replicates – they're pseudoreplicates.
   - The growth chambers are the experimental units: the treatments (CO_2 concentrations) are applied to the growth chambers, not to the plants independently.

3. Two fifth-grade math curricula are being studied.
   - Two schools have agreed to participate in the study.
   - One is randomly assigned to use curriculum A, the other to use curriculum B.
   - At the end of the school year, the fifth-grade students in each school are tested and the results are used to do a statistical analysis comparing the two curricula.
   - There is *no true replication* in this study; *the students are pseudo-replicates*.
   - The schools are the experimental units; they, not the students, are randomly assigned to treatment.
   - Within each school, the test results (and the learning) of the students in the experiment are not independent; they're influenced by the teacher and by other school-specific factors (e.g., previous teachers and learning, socioeconomic background of the school, etc.).

Consequences of doing statistical inference using pseudoreplicates rather than true replicates

*Variability will probably be under**estimated***. This will result in:

- Confidence intervals that are too small.
- An inflated probability of a Type I error (falsely rejecting a true null hypothesis).

Comments

- Note that in Example 2, there's no way to distinguish between effect of treatment and effect of growth chamber; thus the two factors (treatment and growth chamber) are *confounded*. Similarly, in Example 3, treatment and school are confounded.
- Example 3 may also be seen as applying the two treatments to two different *populations* (students in one school and students in the other school)
- Observational studies are particularly prone to pseudoreplication.
- Regression can sometimes partially account for lack of replication, provided data are close enough to each other.
    - The rough idea is that the responses for nearby values of the explanatory variables can give some estimate of the variability.
    - However, having replicates is better.

(See Appendix for suggestions on dealing with pseudoreplication.)

F. MISTAKES IN REGRESSION

There are many common mistakes involved in regression!

Only one will be discussed here; some others will be listed at the end of these notes, with a web reference to more discussion.

Overfitting

> *With four parameters I can fit an elephant and with five I can make him wiggle his trunk.*
> John von Neumann

If we have n distinct x values and corresponding y values for each, it is possible to find a curve going exactly through all n resulting points (x, y); this can be done by setting up a system of equations and solving simultaneously.

- But this is *not* what regression methods typically are designed to do.
- Most regression methods (e.g., least squares) estimate *conditional means* of the response variable given the explanatory variables.
- They're *not* expected to go through all the data points.

For example, with one explanatory variable X (e.g., height) and response variable Y (e.g., weight), if we fix a value x of X, we have a *conditional distribution of Y given X = x* (e.g., the conditional distribution of weight for people with height x).

- This conditional distribution has an expected value (population mean), which we will denote E(Y|X = x) (e.g., the mean weight of people with height x).
- This is the *conditional mean of Y given X = x*. It depends on x -- in other words, E(Y|X = x) is a mathematical function of x.

In least squares regression (and most other kinds of regression), *one of the model assumptions is that the conditional mean function has a specified form.*

- Then we use the data to find a function of x *that approximates the conditional mean function E(Y|X = x)*.
- This is different from, and subtler (and harder) than, finding a curve that goes through all the data points.

Example: To illustrate, I've used simulated data:

- Five points were sampled from a joint distribution where the conditional mean E(Y|X = x) is known to be x^2, and where each conditional distribution Y|(X = x) is normal with standard deviation 1.
- I used least squares regression to estimate the conditional means by a quadratic curve $y = a + bx + cx^2$. That is, I used least squares regression, with

  $$E(Y|X=x) = \alpha + \beta x + \gamma x^2$$

  as one of the model assumptions, to obtain estimates a, b, and c of α, β, and γ (respectively), based on the data.
  - There are other ways of expressing this model assumption, for example,

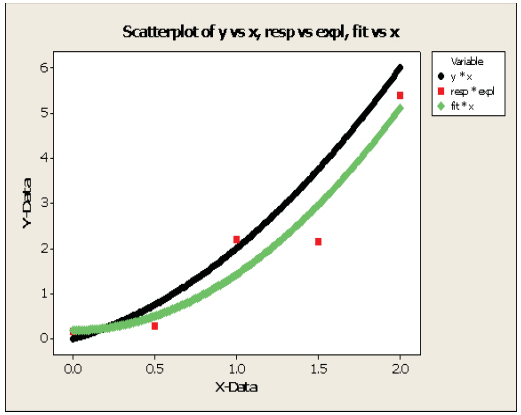
    $$y = \alpha + \beta x + \gamma x^2 + \varepsilon,$$

    or

    $$y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i$$

The graph below shows:

- The five data points in *red* (one at the left is mostly hidden by the green curve)
- The curve $y = x^2$ of true conditional means (*black*)
- The graph of the calculated regression equation (in *green*).



Note that:

- The points sampled from the distribution do *not* lie on the curve of means (black).
- The green curve is not exactly the same as the black curve, but is close.
- In this example, the sampled points were mostly below the curve of means.
- Since the regression curve (green) was calculated using just the five sampled points (red), the red points are more evenly distributed above and below it (green curve) than they are in relation to the real curve of means (black).

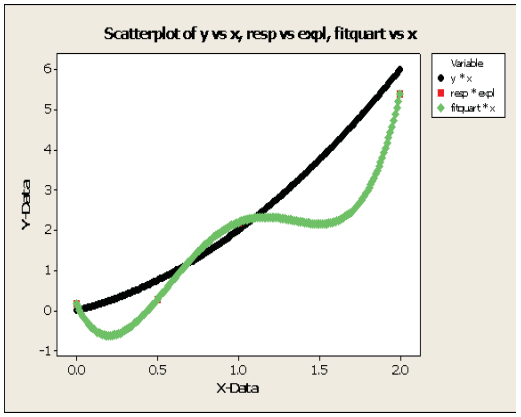
Note: In a real world example, we would *not* know the conditional mean function (black curve) -- and in most problems, would not even know in advance whether it is linear, quadratic, or something else.

- Thus, *part of the problem of finding an appropriate regression curve is figuring out what kind of function it should be.*

Continuing with this example, if we (naively) try to get a "good fit" by trying a quartic (fourth degree) regression curve -- that is, using a model assumption of the form

$$E(Y|X=x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4,$$

we get the following picture:



You can barely see any of the red points in this picture.

- That's because they're all on the calculated regression curve (green).
- We've found a regression curve that fits all the data!
- But it's *not* a good regression curve -- because what we're really trying to estimate by regression is the *black curve* (curve of conditional means).
- We've done a rotten job of that; we've made the mistake of *over-fitting*. We've fit an elephant, so to speak.

If we had instead tried to fit a cubic (third degree) regression curve -- that is, using a model assumption of the form

$$E(Y|X=x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3,$$

we'd get something more wiggly than the quadratic fit and less wiggly than the quartic fit.

- However, it would still be over-fitting, since (by construction) the correct model assumption for these data would be a quadratic mean function.

See the Appendix for suggestions on trying to avoid overfitting.

Other Common Mistakes in Using Regression

For further discussion of these mistakes, see links from **http://www.ma.utexas.edu/users/mks/statmistakes/regression.html**

- Using Confidence Intervals when Prediction Intervals Are Needed.
- Over-interpreting High R^2
- Mistakes in Interpretation of Coefficients
  - Interpreting a coefficient as a rate of change in Y instead of as a rate of change in the conditional mean of Y.
  - Not taking confidence intervals for coefficients (i.e., uncertainty of estimation of coefficients) into account
  - Interpreting a coefficient that's not statistically significant
  - Interpreting coefficients in multiple regression with the same language used for a slope in simple linear regression.
  - Neglecting the issue of multiple inference when dealing with more than one coefficient in the same data set.
- Mistakes in Selecting Terms
- Assuming linearity is preserved when variables are dropped. (*See also Appendix.)*
- Problems involving stepwise model selection procedures.

See also *http://www.jerrydallal.com/LHSP/important.htm* for another common mistake in using regression.

If you have further questions, feel free to:

Consult my website Common Mistakes in Using Statistics (table of contents at http://www.ma.utexas.edu/users/mks/statmistakes/TOC.html)

Email me at mks@math.utexas.edu (or through this class's Canvas site)

Leave a comment on my blog, Musings on Using and Misusing Statistics, http://www.ma.utexas.edu/blogs/mks/