Project Instructions

You will conduct a statistical study on a topic of your choice. This task will require you to write a project proposal, acquire and analyze relevant data, present your results orally to the class, and hand in a written report describing your study and its findings. Your project must involve fitting a *regression* model. The project is an opportunity to show off what you've learned about data analysis, visualization, and statistical inference. It is a major component of the class, and successful completion is required to pass.

* Group Formation

You will work in a group of three, of your choosing. Apart from unusual circumstances, all three of you must be in the same section of MTH 291. Each group will be assigned a letter, and you will use `Group-X` with your particular letter in all communications from that point forward.

* Assignment

You should pose a problem that you find interesting and which may be addressed (at least in part) through the analysis of data. Many interesting quantitative questions (and perhaps even more uninteresting ones) involve the relationships among several variables. Recent projects have considered the following questions:

- What is the role of executive compensation in determining company performance?
- What is the relationship between the value of a share of stock and financial characteristics of a company?
- How is the state's murder rate affected by its demographics and social characteristics?
- How is the percentage of Massachusetts high school seniors going on to four-year colleges influenced by town and school characteristics?
- What factors influence the incidence of tuberculosis in the U.S.?
- How can we predict real estate prices in Northampton?

You should pose the problem that you want to solve as precisely as possible at the outset. Next, identify the population you want to describe, and think about how you will obtain relevant data. What kind of model might be appropriate in this context? You should also make a hypothesis, *a priori* (before you analyze the data), about the results you expect to see.

Most of you will pose your own question and acquire data from the Internet, some may wish to analyze data that someone else (e.g., a professor or office at Smith, published data in a magazine or newspaper) has collected for another purpose. Please consult us if you plan to do this.

General Rules

You all *must* speak during the oral presentation. You may discuss your project with other students, but each of you will have a different topic, so there is a limit to how much you can help each other. You may consult other sources for information about the non-statistical, substantive issues in your problem, but you should credit these sources in your report. Feel free to consult with us about statistical questions.

* Submission

All deliverables described above must be delivered electronically via Moodle by 11:55pm (five minutes before midnight) on the dates in the Syllabus. Only one person from the group should submit the group's product for each checkpoint (with the exception of the Group Dynamic, which is individual).

Components

Roster

Form a group of 3 students from your section. Have one person send a group roster electronically by the date listed above, with appropriate cc's, using the message subject header `MTH 291 Group Roster`. Take the initiative to ask around and find a group to work with.

Proposals

Count on brainstorming at least half a dozen serious ideas before you can groom one of them into a mature proposal.

For the most part, the choice of topic is left up to you. Try to pick something that's interesting yet substantial and worth studying, and aim for a topic that you think nobody has tried before; remember that part of your overall grade will be based on originality.

Places to Find Data

Finding the right data to answer your particular question is part of your responsibility for this assignment. Public data sets are available from hundreds of different websites, on virtually any topic. You might not be able to find the exact data that you want, but you should be able to find data that is relevant to your topic. You may also want to refine your research question so that it can be more clearly addressed by the data that you found. But be creative! Go find the data that you want!

Below is a list of places to get started, but this list should be considered grossly non-exhaustive:

- Finding Data on the Internet (http://www.inside-r.org/howto/finding-data-internet)
- Gapminder (www.gapminder.org)
- Data.gov (explore.data.gov)
- StatLib at Carnegie Mellon (http://lib.stat.cmu.edu/)
- U.S. Bureau of Labor Statistics (www.bls.gov)
- U.S. Census Bureau (www.census.gov)

Keep the following in mind as you select your topic and dataset:

- You need to have enough data to make meaningful inferences. There is no magic number of individuals required for all projects. But aim for at least 200 individuals and make sure there are at least 20 individuals in each category of each of your categorical variables (if you have any).
- Most projects will measure a quantitative outcome, with at least two other variables included in the dataset (ideally at least one of which is quantitative). Most of you will use multiple linear regression for your primary analyses.

Once we respond to your initial proposal, you will revise it (perhaps starting with a different dataset), then submit a new proposal that addresses our feedback. Supply essentially the same information required for the initial proposal, but give a bit more detail.

Content

Your initial and revised proposals should contain the following content:

1. Group Members: List the members of your group
2. Title: Your title
3. Purpose: Describe the general topic/phenomenon you want to study, as well some focused questions that you hope to answer and specific hypotheses that you intend to assess.
4. Data: Describe the data that you plan to use, with specifications of where it can be found (URL) and a short description. Eventually, you will probably want to combine data from multiple sources into one file. We will discuss data management techniques in the coming weeks, but for now you should simply list multiple sources if you have them.
5. Population: Specify what the observational units are (i.e. the rows of the data frame), describe the larger population/phenomenon to which you'll try to generalize, and (if appropriate) estimate roughly how many such individuals there are in the population.
6. Response Variable: What the response variable? What are its units? Estimate the range of possible values that it may take on.
7. Explanatory Variables: Describe the variables that you'll examine for each observational unit (i.e. the columns of the data frame). Carefully define each variable and describe how each was measured. For categorical variables, list the possible categories; for quantitative variables, specify the units of measurement. You may want to add more variables later on, but you should have at least 5 variables already.

Data

You must finalize and submit your data file to us. Your data file should also be placed in your Dropbox folder. Your technical report should import this data into RStudio using the `read.csv()` command.

- The data must be in CSV format (`.csv`). This means that the first row should be a comma-separated list of variable names, and the rest should be rows of data.
- Your data file should be named `group-X-data.csv`.
- Name all variables helpfully and contextually, e.g., use `Airport` and `WaterTemp`, not `Individuals` and `Treatments`, and certainly not `A` and `B`. Similarly, for the category names, use whole words and phrases, not cryptic codes, e.g., use `Male` and `Female`, not `1` and `2`. A binary variable `isFemale` can be coded 0 for male, and 1 for female (and then is self-documenting). A variable `sex` coded 1 and 2 is just asking for trouble.
- That said, try to limit your variable and category names to about a dozen characters. This may take some abbreviation.
- Check for data integrity! Manual inspection is OK, but it's tedious and it's easy to overlook misspellings. Running some simple analyses can more quickly make most data entry errors obvious.

Technical Report

Your technical report will be an annotated R Markdown file (`.Rmd`) that contains your R code, interspersed with explanations of what the code is doing, and what it tells you about the problem.

Content

You should **not** need to present *all* of the R code that you wrote throughout the process of working on this project. Rather, the technical report should contain the *minimal* set of R code that is necessary to understand your results and findings in full. If you make a claim, it *must* be justified by explicit calculation. A knowledgeable reviewer should be able to compile your `.Rmd` file without modification, and verify every statement that you have made. All of the R code necessary to produce your figures and tables *must* appear in the technical report. In short, the technical report should enable a reviewer to reproduce your work in full.

Tone

This document should be written for peer reviewers, who comprehend statistics at least as well as you do. You should aim for a level of complexity that is more statistically sophisticated than an article in the Science section of *The New York Times*, but less sophisticated than an academic journal. [*Chance* magazine might provide a good example.] For example, you may use terms that that you will likely never see in the *Times* (e.g. bootstrap), but should not dwell on technical points with no obvious ramifications for the reader (e.g. reporting F-statistics). Your goal for this paper is to convince a statistically-minded reader (e.g. a student in this class, or a student from another school who has taken an introductory statistics class) that you have addressed an interesting research question in a meaningful way. Even a reader with no background in statistics should be able to read your paper and get the gist of it.

Format

Your technical report should follow this basic format:

1. Abstract: a short, one paragraph explanation of your project. The abstract should not consist of more than 5 or 6 sentences, but should relate what you studied and what you found. It need only convey a general sense of what you actually did. The purpose of the abstract is to give a prospective reader enough information to decide if they want to read the full paper.

2. Introduction: an overview of your project. In a few paragraphs, you should explain *clearly* and *precisely* what your research question is, why it is interesting, and what contribution you have made towards answering that question. You should give an overview of the specifics of your model, but not the full details. Most readers never make it past the introduction, so this is your chance to hook the reader, and is in many ways the most important part of the paper!

3. Data: a brief description of your data set. What variables are included? Where did they come from? What are units of measurement? What is the population that was sampled? How was the sample collected? You should also include some basic univariate analysis.

4. Results: an explanation of what your model tells us about the research question. You should interpret coefficients in context and explain their relevance. What does your model tell us that we didn't already know before? You may want to include negative results, but be careful about how you interpret them. For example, you may want to say something along the lines of: "we found no evidence that explanatory variable x is associated with response variable y," or "explanatory variable x did not provide any additional explanatory power above what was already conveyed by explanatory variable z." On other hand, you probably shouldn't claim: "there is no relationship between x and y."

5. Conclusion: a summary of your findings and a discussion of their limitations. First, remind the reader of the question that you originally set out to answer, and summarize your findings. Second, discuss the limitations of your model, and what could be done to improve it. You might also want to do the same for your data. This is your last opportunity to clarify the scope of your findings before a journalist misinterprets them and makes wild extrapolations! Protect yourself by being clear about what is *not* implied by your research.

Additional Thoughts

The technical report is *not* simply a dump of all the R code you wrote during this project. Rather, it is a narrative, with technical details, that describes how you addressed your research question. You should *not* present tables or figures without a written explanation of the information that is supposed to be conveyed

by that table or figure. Keep in mind the distinction between *data* and *information*. Data is just numbers, whereas information is the result of analyzing that data and digesting it into meaningful ideas that human beings can understand. Your technical report should allow a reviewer to follow your steps from converting data into information. There is no limit to the length of the technical report, but it should not be longer than it needs to be. You will not receive extra credit for simply describing your data *ad infinitum*. For example, simply displaying a table with the means and standard deviations of your variables is not meaningful. Writing a sentence that reiterates the content of the table (e.g. "the mean of variable x was 34.5 and the standard deviation was 2.8. . . ") is equally meaningless. What you should strive to do is interpret these values in context (e.g. "although variables x_1 and x_2 have similar means, the spread of x_1 is much larger, suggesting. . . ").

You should present figures and tables in your technical report in context. These items should be understandable on their own – in the sense that they have understandable titles, axis labels, legends, and captions. Someone glancing through your technical report should be able to make sense of your figures and tables without having to read the entire report. That said, you should also include a discussion of what you want the reader to learn from your figures and tables.

Your report should be submitted via email as an R Markdown (`.Rmd`) file and the corresponding rendered output (`.html`) file.

Presentation

An effective oral presentation is an integral part of this project. One of the objectives of this class is to give you experience conveying the results of a technical investigation to a non-technical audience in a way that they can understand. Whether you choose to stay in academia or pursue a career in industry, the ability to communicate clearly is of paramount importance. As a data analyst, the burden of proof is on you to convince your audience that what you are saying is true. If your audience (who may very well be less knowledgeable about statistics than you are) cannot understand your results or their interpretations, then the technical merit of your project is irrelevant.

You will make a 10-minute oral presentation to the class. You should make (good) slides. Your goal should be to convey to your audience a clear understanding of your research question, along with a basic understanding of your model, and how well it addresses the research question you posed. You should **not** tell us everything that you did, or show a bunch of models that didn't work well. After hearing your talk, each student in the class should be able to answer:

1. What was your project about?
2. What kind of model did you build?
3. How well did it work?

You should prepare electronic slides for your talk. PowerPoint is fine, but you might also want to consider Google Presentation, Beamer (LaTeX), or alternative, non-linear presentation software like Prezi. Use your creativity! One thing you should *not* do is walk us through your calculations in RStudio. There will be an opportunity to rehearse your presentation with one of us a few days before your talk.

You will need to submit your slides before your presentation, but you should also bring the slides on a flash drive as a backup. You will not be able to hook up your laptop to the computer in SR 301.

Advice

There are many sources of advice for how to make a good presentation, but an excellent place to start is Technically Speaking. Watch the videos on this site to identify some common mistakes. You should also read Joe Gallian's article on how to make a good presentation.

Here are is some general advice:

- Budget your time. You only have 10 minutes, and we will be running a very tight schedule. If your talk runs too short or too long, it makes you seem unprepared. Rehearse your talk ahead of time (with your group) several times in order to get a better feel for your timing. Note also that you may have a tendency to talk faster during your actual talk than you will during your rehearsal. Talking faster in order to speed up is not a good strategy – you are much better off simply cutting material ahead of time. You will probably have a hard time getting through 10 slides in 10 minutes.
- Don't write too much on each slide. You don't want people to have to read your slides, because if the audience is reading your slides, then they aren't listening to you. You want your slides to provide visual cues to the points that you are making – not substitute for your spoken words. Concentrate on graphical displays and bullet-pointed lists of ideas.
- Put your problem in context. Remember that most of your audience will have little or no knowledge of your subject matter. The easiest way to lose people is to dive right into technical details that require prior 'domain knowledge." Spend a few minutes at the beginning of your talk introducing your audience to the most basic aspects of your topic and present some motivation for what you are studying.
- Speak loudly and clearly. Remember that you know more about your topic that anyone else in the room, so speak and act with confidence!
- Tell a story – not necessarily the whole story. It is unrealistic to expect that you can tell your audience everything that you know about your topic in 10 minutes. You should strive to convey the big ideas in a clear fashion, but not dwell on the details. Your talk will be successful if your audience is able to walk away with an understanding of what your research question was, how you addressed it, and what the implications of your findings are.

Group Dynamic Report

Ideally, all group members would be equally involved and able and committed to the project. In reality, it doesn't always work that way. We'd like to reward people fairly for their efforts in this group endeavor, because it's inevitable that there will be variation in how high a priority people put on this class and how much effort they put into this project.

To this end, we'd like each of you (individually) to describe how well (or how poorly!) your project group worked together and shared the load. Also give some specific comments describing each member's overall effort. Were there certain group members who really put out exceptional effort and deserve special recognition? Conversely, were there group members who really weren't carrying their own weight? And then, at the end of your assessment, estimate the percentage of the total amount of work/effort done by each member. (Be sure your percentages sum to 100%!)

For example, suppose you have 3 group members: X, Y and Z. In the (unlikely) event that each member contributed equally, you could assign:

- 33.3% for member X, 33.3% for member Y, and 33.3% for member Z

Or in case person Z did twice as much work as each other member, you could assign:

- 25% for member X, 25% for member Y, and 50% for member Z

Or if member Y didn't really do squat, you could assign:

- 45% for member X, 10% for member Y, and 45% for member Z

We'll find a fair way to synthesize the (possibly conflicting) assessments within each group. And eventually we'll find a way to fairly incorporate this assessment of effort and cooperation in each individual's overall grade. Don't pressure one another to give everyone glowing reports unless it's warranted, and don't feel pressured to share your reports with one another. Just be fair to yourselves and to one another. Let us know if you have any questions or if you run into any problems.

Assessment Criteria

Your project will be evaluated based on the following criteria:

- General: Is the topic original, interesting, and substantial – or is it trite, pedantic, and trivial? How much creativity, initiative, and ambition did the group demonstrate? Is the basic question driving the project worth investigating, or is it obviously answerable without a data-based study?
- Design: Are the variables chosen appropriately and defined clearly, and is it clear how they were measured/observed? Can the effects of lurking variables be controlled for? Is there sufficient data to make meaningful conclusions?
- Analysis: Are the chosen analyses appropriate for the variables/relationships under investigation, and are the assumptions underlying these analyses met? Do the analyses involve fitting and interpreting a multiple regression model? Are the analyses carried out correctly? Is there an effective mix of graphical, numerical, and inferential analyses? Did the group make appropriate conclusions from the analyses, and are these conclusions justified?
- Technical Report: How effectively does the written report communicate the goals, procedures, and results of the study? Are the claims adequately supported? How well is the report structured and organized? Does the writing style enhance what the group is trying to communicate? How well is the report edited? Are the statistical claims justified? Are text and analyses effectively interwoven in the technical report? Clear writing, correct spelling, and good grammar are important.
- Oral Presentation: How effectively does the oral presentation communicate the goals, procedures, and results of the study? Do the slides help to illustrate the points being made by the speaker without distracting the audience? Do the presenters seem to be well-rehearsed? Did they properly budget their time? Does she appear to be confident in what she is saying? Are her arguments persuasive?