431 Project Instructions

Thomas E. Love

Version: 2018-09-11 08:25:52

Contents

Overview

This website contains the Fall 2018 project information for PQHS / CRSP / MPHP 431: Statistical Methods in Biological & Medical Sciences, Section 1.

- All materials related to the project (including these instructions) are maintained and linked at https://github.com/THOMASELOVE/431-2018-project.
- The direct link to this document is https://thomaselove.github.io/431-2018-project.

Your Project includes Two Studies

Your final project for this course will result in a portfolio of work related to two studies.

Study 1 - Class Survey. In the first study, you (sometimes working individually, sometimes in a group) will design, administer, analyze and present the results of a survey designed to compare two or three groups of subjects on some *categorical* and *quantitative* outcomes we will develop from your initial ideas.

Study 2 - Your Data. In the second study, you (working individually) will propose a research question and relevant data of interest to you, and then complete all elements of a data science project designed to create a statistical model for a *quantitative* outcome, then use it for prediction and assess the quality of those predictions.

You have Nine Tasks to Complete this Semester

The project involves two analyses (one for the class survey and one for your personal study), and a total of 9 tasks (deliverables.) Each task is to be completed by **12 NOON** on the specified date.

- Task A (The Proposal) is due at noon on 2018-10-12
- Task B (Presentation Sign-Up) is **also** due at noon on 2018-10-12
- Task C (Survey Editing) involves group work and is due at noon on 2018-10-23
- Task D (Survey Comparison Plan) is **also** due at noon on 2018-10-23
- Task E (Taking the Survey) is due at noon on 2018-10-31
- Task F (Sharing Study 2 Data) is due at noon on 2018-11-14
- Task G (The Update) is due at noon on 2018-11-28
- Task H (The Portfolio) is due at noon on 2018-12-13
- Task I (Your Presentation) will be held on 2018-12-10, 2018-12-11 or 2018-12-13

The bulk of this document contains specific instructions for each of these tasks.

Working with This Document

1. This document is broken down into multiple sections. Use the table of contents at left to navigate.
2. At the top of the document, you'll see icons which you can click to

- search the document,
- change the size, font or color scheme of the page, and
- download a PDF or EPUB (Kindle-readable) version of the entire document.

3. The document is a work in progress, and will be updated occasionally through the semester. Check the Version information above to verify the last update time[1].

Need Help?

Questions about the project or the course can be directed to **431-help at case dot edu** or to Dr. Love directly at `thomas dot love at case dot edu`.

- The course home page is at https://github.com/THOMASELOVE/431-2018

[1] Note that the ePub and PDF versions will show slightly different times (but on the same day) as the HTML version.

Chapter 1

Project Objectives

It is hard to learn statistics (or anything else) passively; concurrent theory and application are essential[1]

1.1 Study 1 is about making comparisons and visualizing groups of data.

Study 1 involves data from a **class survey**, to be conducted in October. We will design, administer, analyze and present survey results designed to compare two or three groups of subjects from the class on some *categorical* and *quantitative* outcomes. In the analysis stage, everyone will be working with different parts of the same data set.

> Think of a graph as a comparison. All graphs are comparisons (indeed, all statistical analyses are comparisons). If you already have the graph in mind, think of what comparisons it's enabling. Or if you haven't settled on the graph yet, think of what comparisons you'd like to make. Andrew Gelman

In your eventual analysis of Study 1, you will be comparing both quantitative and categorical outcomes across 2-3 groups. All tools necessary for Study 1 are in Parts A and B of the course, and include the following…

- Descriptive and exploratory summaries of the data across the groups for each of your chosen outcomes, including, of course, attractive and well-constructed visualizations, graphs and tables.
- Comparisons of the population mean difference for at least one quantitative outcome across a set of two (or three) groups, including appropriate demonstrations of the reasons behind the choices you made between parametric, non-parametric and bootstrap procedures.
- Comparisons of the population proportions for at least one categorical outcome across your set of two (or three) groups, including appropriately interpreted point estimates and confidence intervals.

Note well that Study 1 is **not** about building sophisticated statistical models, and using them to make predictions. That's Study 2.

1.2 Study 2 is about building a model, and making predictions.

Study 2 involves data about a **research question that you will propose**, involving data of interest to you. Thus, everyone will be working with a different data set. You will complete all elements of a data

[1] Though by no means an original idea, this particular phrasing is stolen from Harry Roberts.

science project designed to create a statistical model for a *quantitative* outcome, then use it for prediction, and assess the quality of those predictions.

> All models are wrong but some are useful. George E. P. Box

In Study 2, you will be building a multiple linear regression model, and using it to predict a quantitative outcome of interest. The tools necessary for Study 2 appear in each Part of the course, especially Part C, and include the following…

- Describing the experimental or observational study design used to gather the data, as well as the complete data collection process.
- Sharing the complete raw data in an appropriate way with a statistician (Dr. Love). This means that, in general, data including protected health information are *not* appropriate for this project.
- Developing appropriate research questions that lead to the identification of smart measures for predictors and outcomes, and then the development of a prediction model using multiple linear regression.
- Using a training sample to develop a model, and present the process that leads to a final set of 2-3 candidate models in the training sample.
- Using a test sample to evaluate the quality of predictions from each of the candidate models, and making a final selection.
- Evaluating the adherence of the data you've collected to the assumptions of multiple linear regression, and iterating through the model-building process as necessary until the final model shows no strong violations of those assumptions.

1.3 Why Two Studies?

The main reason is that I can't figure out a way to get you to think about all of the things I hope you'll learn from this project in a single study.

1. I set different tasks for Study 1 and for Study 2, allowing us to touch on a wider fraction of the things I hope you'll learn in 431.
2. I want some of the work to be done as a class, some in groups, some as individuals.
3. Some of you have easy access to great data you want to study in this class, and in fact, that's a primary motivation for taking the class. But not all of you.
4. I have to evaluate each of your projects, and there are many students in the class. Knowing at least one of the data sets you'll be working with helps me manage this.
5. Having a broad range of activities to evaluate helps reduce the cost of a mistake on any one of them, so that we can build on what you do well.
6. All of Study 1 can be done by the middle of November, leaving the last few weeks of the semester for you to focus on Study 2.

1.4 Educational Objectives

> "Statistics has no reason for existence except as the catalyst for investigation and discovery." George E. P. Box

I am primarily interested in your learning something interesting, useful and even valuable from your project. An effective project will demonstrate:

1. The ability to create and formulate research questions that are statistically and scientifically appropriate.
2. The ability to turn research questions into measures of interest.
3. The ability to pull and merge and clean and tidy data, then present the data set following Jeff Leek's guide to sharing data with a statistician.
4. The ability to identify appropriate estimation / testing procedures for the class survey using both continuous and categorical outcomes.

5. The ability to build a reasonable model, including interactions and transformations to deal with non-linearity, assess the quality of the model and residual plots, then use the model to make predictions.
6. The ability to build a Table 1 to showcase potential differences between variables.
7. The ability to identify and (with help) solve problems that crop up
8. The ability to comment on your work within code, and in written and oral presentation.
9. The ability to build a Markdown-based report and a Markdown-based set of slides for presentation.

Chapter 2

Task A (The Proposal) Instructions

Task A requires you to:

1. develop and propose two research questions for Study 1. Here, your research questions should clearly identify meaningful statistical comparisons.
2. propose 5 "homemade" survey questions for Study 1. Students in MS or PhD programs in the Department of Population and Quantitative Health Sciences (PQHS) have a modest additional requirement here.
3. develop and propose a meaningful research question for Study 2. This question needs to clearly relate to modeling and prediction of a quantitative outcome on the basis of a set of predictor variables.
4. identify and present a data set that is likely to lead to an answer to the research question proposed for Study 2

The rest of this section contains guidance as to what sort of questions you need to propose for the survey, and as to what sorts of data sets and research questions are appropriate for your proposal. You will submit your proposal via Blackboard.

2.1 Study 1 work for Task A

In Study 1, you will survey your fellow students through an instrument we will develop in September and October and administer in November. The course survey will be done online, and will include responses (de-identified, of course) from all students in the current 431 class, plus the teaching assistants, and perhaps some volunteers from previous iterations of the course, in an effort to get a reasonable sample of graduate students at CWRU. The final survey will be much longer than would be ideal, and will include questions from each student in the class.

In Task A, all students will specify first specify two *research questions*, and then specify 5 *survey questions*. Students in the MS/PhD programs in PQHS will also specify a published scale (with proper citations) that is available for our use.

2.1.1 Study 1 is about Making Comparisons

In your eventual analysis of Study 1, you will be responsible for comparing both quantitative and categorical outcomes across two (or three) groups. Some tools you will use in completing Study 1 include:

- Descriptive and exploratory summaries of the data across the groups for each of your chosen outcomes, including, of course, attractive and well-constructed visualizations, graphs and tables.

- Comparisons of the population mean difference for at least one quantitative outcome across a set of two (or three) groups, including appropriate demonstrations of the reasons behind the choices you made between parametric, non-parametric and bootstrap procedures.
- Comparisons of the population proportions for at least one categorical outcome across your set of two (or three) groups, including appropriately interpreted point estimates and confidence intervals.

Note well that Study 1 is **not** about building sophisticated statistical models, and using them to make predictions. That's Study 2.

2.1.2 Specifying Research Questions

- The research questions you write for Study 1 state the main objective of the study in terms that allow us to apply statistical methods to test data to obtain an answer.
- The research question should be written in the form of a comparison of two exposures or groups in terms of first your quantitative outcome, and then, in a second question, your categorical outcome.
- A research question is the fundamental core of a research project, study, or review of literature. It focuses the study, determines the methodology, and guides all stages of inquiry, analysis, and reporting. Source

Quoting Roger Peng, from Exploratory Data Analysis with R:

> Formulating a question can be a useful way to guide the exploratory data analysis process and to limit the exponential number of paths that can be taken with any sizeable dataset. In particular, a sharp question or hypothesis can serve as a dimension reduction tool that can eliminate variables that are not immediately relevant to the question. For example, (suppose that we are interested in) looking at an air pollution dataset from the U.S. Environmental Protection Agency (EPA).
>
> A general question one could ask is "Are air pollution levels higher on the east coast than on the west coast?" But a more specific question might be "Are hourly ozone levels on average higher in New York City than they are in Los Angeles?"
>
> Note that both questions may be of interest, and neither is right or wrong. But the first question requires looking at all pollutants across the entire east and west coasts, while the second question only requires looking at single pollutant in two cities. It's usually a good idea to spend a few minutes to figure out what is the question you're really interested in, and narrow it down to be as specific as possible (without becoming uninteresting).

2.1.3 Checklist for Research Questions

1. Is my research question (RQ) something that I am curious about and that others might care about?
2. Does my RQ present an issue on which I can justify a stand prior to data collection about what I think will happen?
3. Is my RQ too broad, too narrow, or OK, given the time frame and restrictions of this assignment?
4. Is my RQ measurable? What type of information do I need? Can I find a way to ask a limited number of survey questions in such a way to allow me to (after the data are collected) either support or contradict a position on my RQ?

- Adapted from this online tutorial from Empire State College

2.1.4 Tips on Writing Good Research Questions

- Duke has a nice overview online of key issues.
- Vanderbilt has some nice materials, built from the tutorial at Empire State College quoted earlier
- Jeff Leek provides several relevant tips in *The Elements of Data Analytic Style*
- https://researchrundowns.com/intro/writing-research-questions/ has some excellent tips on wording

2.1.5 Specifying Survey Questions

A piece of Task A, then is to specify the exact wording for your survey questions (and potential answers for any categorical responses.) This will likely require some editing and rework, once we have the complete set of proposed questions from all students. Be prepared to revise and resubmit, quickly, so that all items can be resolved in time for publication of the draft survey.

- Question 1 should ask the people taking the survey to provide you with a number that expresses a quantitative outcome of interest to you, and this outcome should directly relate to your research question(s).
- Question 2 should do the same thing as Question 1: ask for a quantitative measure of interest that will serve as your backup outcome in case the first question does not work well. A sensible item will show some variation across individuals, and also across groups of interest.
- Question 3 should ask the people taking the survey to provide you a response that expresses a categorical outcome of interest to you, and this outcome should directly relate to your research question(s). You will need to specify each of the available responses that you wish to use in the survey. No more than five options for your categorical outcome, please.
- Question 4 should do the same thing as Question 3 and will again serve as a backup. Again, all responses must be pre-specified.
- Question 5 should ask the people taking the survey to categorize themselves into one of two (or three) groups. You will need to have at least 10 students in each group, and there will be somewhere between 52 and 56 people taking the survey, in total.

For each survey question, you will need to specify the *type* of question you are asking, and the type of outcome (quantitative, or categorical.)

2.1.6 The 15 Items Dr. Love will automatically include

The following items will be included in the survey to identify "groups" of students in a reasonable way. As a result, you will not want to ask these questions, although you should consider these groupings as candidates for application in your research questions, in addition to the grouping you specify in your Question 5.

Together, after the survey is complete, we will identify "cutpoints" for these first eight items to identify groups of meaningful size.

1. In what year were you born?
2. How would you rate your current health overall (Excellent, Very Good, Good, Fair, Poor)
3. For how long, in months, have you lived in Northeast Ohio?
4. What is your height in inches? (If you are five feet, eight inches tall, please write 68 inches. To convert from centimeters to inches, multiply your height in centimeters by 0.3937, and then round the result to the nearest inch.)
5. What is your weight in pounds? (To convert from kilograms to pounds, multiply your weight in kilograms by 2.2046, and then round the result to the nearest pound.)
6. What is your pulse rate, in beats per minute? (Please either use a tracking device, or count your pulse for 15 seconds then multiply by 4)
7. Last week, on how many days did you exercise? (0 - 7)
8. Last night, how many hours of sleep did you get?

The following 7 items will have yes/no responses, and thus produce binary groups for analysis.

1. Were you born in the United States?
2. Is English the language you speak better than any other?
3. Do you identify as female?
4. Do you wear prescription glasses or contact lenses?
5. Before taking 431, had you ever used R before?
6. Are you currently married or in a stable domestic relationship?
7. Have you smoked 100 cigarettes or more in your entire life?

2.1.6.1 Permitted Types of Items

The survey will be conducted using a Google Form, rather than Survey Monkey or some other tool. Thus, we have a somewhat restricted set of item types.

For **quantitative measures**, Google Forms permit the use of

1. a *short answer* item without any restrictions on the response (except a character limit)
2. a *short answer* item where respondents are forced to insert a number within a given range through a validation process that only accepts the response if it falls within the specified limits.
3. *linear scale* items for ordered categorical ratings (but only on a scale of up to ten points - i.e. 1 to 10)

For **categorical measures**, Google Forms permit the use of

1. *multiple choice* items for endorsing a single choice from a group of 2-10 alternatives.
2. *checkbox* items for the endorsement of one or more choices from a group of 2-10 alternatives.
3. *linear scale* items for ordered categorical ratings (often on a 1-X scale, where X is between 2 and 10)
4. *dropdown* items for selections of one option from a group of 2-3 choices.

2.1.6.2 A Few Notes

- If you are asking people to respond to a prompt using a scale, that scale should be expressed on a wide scale. We will accept 1-10 so long as careful descriptions are provided for the meaning of each endpoint.
- Our preference is a 0-100 scale for quantitative items that involve a rating, where 0 represents the most negative reaction and 100 the most positive reaction to the item. One common choice is 0 = Strongly Disagree to 100 = Strongly Agree. The reason we prefer a 0-100 scale is to increase variation in our responses. When answering questions, please use the whole scale.
- In the survey, I will ask students to provide their age, select their sex, and whether they were born in the United States. I will use this to provide three additional groupings (lower half of ages vs. higher, males vs. females, and US natives vs. non-natives) that you will be able to use instead of your Question 5 grouping if that grouping doesn't work out for some reason. So, don't include age, gender or country of origin in your survey questions.
- Old surveys (with more than 100 items each) are available for you to view. If you are curious about questions that have been used in the past, here are links to the 2014 survey, the 2015 survey, the 2016 survey and the 2017 survey.

2.1.7 Extra task for PQHS MS/PhD students

Students in MS or PhD programs in the Department of Population and Quantitative Health Sciences will also need to specify a published scale (available for public use) to generate an outcome or grouping(s) of interest from those completing the survey. A scale is a published list of items, usually accompanied by a scoring rubric. Examples of scales we have used in the past include:

-

Those outside the EPBI MS/PhD track are permitted to submit a scale as well, if they would prefer to use such a scale instead of some part of their homemade group of 5 questions. Please indicate this preference clearly in your proposal.

2.2 Study 2 Work for Task A

You will present a proposal **summary** (< 300 words) and a brief **data description** for Study 2 in Task A.

You will be building a multiple regression model, and using it to predict your outcome of interest.

- We prefer data sets for this work containing 250 to 250,000 observations, including at least one quantitative outcome, and at least four predictor variables (one of which may be identified as the "key" predictor of interest.)
- Predictors may be quantitative or categorical.
- If you would like to use a data set which does not meet these specifications, contact Dr. Love via email well in advance of the Task A deadline (2016-10-11) to explain your reasons so that he can either approve your choice of data set, or not, in time for you to find a new data set.

2.2.1 The Proposal Summary

Take the time to come up with a good, interesting title. You are going to work hard on this thing; please resist the temptation to kill my interest at the start by calling it "EPBI 431 Statistics Project."

Provide me a very brief summary of what you're trying to accomplish - specifically, what your research question is, and what you hypothesize will happen.

- The summary is the heart of the proposal, and requires some care. You will need to convince me that your topic is interesting, your data are relevant, and building a model and making predictions of a quantitative outcome using the predictors available to you will be worthwhile.
- The summary ends with a statement of the research question or questions (you may have one, or possibly two.) An excellent question conveys the main objective of the study in terms that allow us to apply statistical models to describe an association between one or more predictors and a quantitative outcome.
- It should be possible for me to explain your study accurately just by reading this summary. If it's not possible, it will come back to you for speedy rework.
- This summary should be less than 300 words.
- Use complete English sentences. Write in plain language. Use words we all know. Avoid jargon.

2.2.2 The Data Description

Your data description can be as long as it needs to be. It should include:

1. Your data source, which can be an online source (in which case include a working link), a published paper or journal article (in which case I need a link and a PDF copy of the paper), or unpublished data (in which case I need the details of how the data were gathered)
2. A thorough description of the data collection process, with complete details as to the nature of the variables, the setting for data collection, and complete details of any apparatus you used which may affect results.
3. Specification of the people and methods involved.
    - Who are the subjects under study?
    - When were the data gathered? By whom?
    - How many subjects are included?
    - What caused subjects to be included or excluded from the study?
4. Your planned quantitative outcome, which must relate directly to the research question you specified above. Provide a complete definition, including specifying the exact wording of the question or details of the measurement procedure used to obtain the outcome. If available, you can also include descriptios of secondary quantitative outcomes.
5. Your predictors of interest, which should also relate to the research question in an obvious way. Again, define the variables carefully, as you did with the outcome.
6. If you already have the data, tell me that. If you don't, specify any steps you must still take in order to get the data, and specify the date by which you will have your data (must be no later than 2016-11-01.)

2.2.3 Some Potentially Useful Data Sources

The ideal choice of data source for this project is a public-use version of a meaningful data set without access restrictions. With so many students in the class, I cannot be responsible for supervising your work with restricted data personally. Some appealing sources to explore include:

- https://www.data.gov/ The home of the U.S. Government's open data
- http://www.census.gov/data.html The U.S. Census Bureau has many interesting data sets, including the Current Population Survey
- http://www.healthdata.gov/ 125 years of U.S. Health Care Data
- http://www.cdc.gov/nchs/nhanes/index.htm National Health and Nutrition Examination Survey.
  - You may want to look at the nhanesA package in R
- http://dashboard.healthit.gov/datadashboard/data.php Office of the National Coordinator for Health IT's dashboard
- http://www.icpsr.umich.edu/icpsrweb/ ICSPR (Inter-university Consortium for Political and Social Research) is a source for many public-use data sets
  - This includes the Health and Medical Care data archive of the Robert Wood Johnson Foundation
- http://gss.norc.org/ The General Social Survey
- http://www.bls.gov/data/ Bureau of Labor Statistics
- http://nces.ed.gov/surveys/ National Center for Education Statistics
- http://www.odh.ohio.gov/healthstats/dataandstats.aspx Ohio Department of Health
- http://open.canada.ca/en Canada Open Data
- http://digital.nhs.uk/home Health data sets from the UK National Health Service.
- http://www.who.int/en/ World Health Organization
- http://www.unicef.org/statistics/ UNICEF has some available data on women and children
- http://www.pewinternet.org/datasets/ Pew Research Center's Internet Project
- http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi Broad Institute's Cancer Program
- http://www.kdnuggets.com/datasets/index.html is a big index of lots of available data repositories
- https://www.kaggle.com/ Kaggle competition data sets are an interesting possibility

I cannot guarantee the quality of any of the data sets available at these sites, but I've spent at least a little time at most of them in recent months.

2.2.4 Some Restrictions

Note that it is especially appealing, in Study 2, to make use of data that you are studying in your own field that fit the criteria I describe and which can be made available to me. Ideally, therefore, you would be working with data that are available to the public.

1. You need to be able to share your data with a statistician (Dr. Love) following Jeff Leek's guide to sharing data with a statistician. This means you need to have access to the data in the raw, and it means that I have to be able to have access to it in the raw, as well. So, studies involving protected health information are **not** appropriate.
2. A full citation for any and all data elements, including a complete codebook, must be provided as part of your proposal.
3. There is more to a statistical application than the analysis of a canned data set, even a good canned data set. Googling "data sets for regression projects" or something similar is not a good strategy. I am not interested in you using pre-cleaned data from an educational repository, such as:
   - this one at the Cleveland Clinic, or this one at Vanderbilt University, or this one at UCLA, or this one at the University of Florida, or this one at Florida State University, or
   - StatLib at Carnegie-Mellon University, or the Journal of Statistics Education Data Archive, or
   - StatSci.org's repository of textbook examples and ready for teaching data, or
   - any of the many textbook-linked repositories of data sets, like this one for Statistics: Unlocking the Power of Data, or

- any similar repository Professor Love deems to be inappropriate

4. While there are some great resources available to some people in this class by virtue of their affiliation with one of the health systems in town, I can do nothing to get you access to health system specific data as part of your project for this class or for 432, and in general, data from those sources are not especially appropriate because of issues with protected health information.

2.3 Evaluating Task A

Dr. Love will evaluate all proposals (Task A) personally, in the order in which they are received. Proposals will receive one of two grades: OK or REDO. REDO will be accompanied with specific requests that should be accomplished within a short time window (approximately 24 hours). If you materially deviate from these specifications, he will return your proposal without comment other than to re-specify what needs to be fixed before he responds.

Chapter 3

Task B (Presentation Sign-Up) Instructions

Chapter 4

Task C (Survey Editing) Instructions

or Task B, you need to submit via Blackboard, a single Word document (maximum one page, 12 point font, with your name and Project Task B on the top of the Word document) containing these two things:

1. Your list of typographical errors, clarifications or other edits to the 96 items currently included in the Survey Draft, available at https://sites.google.com/a/case.edu/love-431/home/projects/class-survey. If you found no errors or items in need of clarification, write a sentence saying that. If you did find an issue, please be sure to specify the item number (1-96) where you feel a revision is needed.
    - If you see any items in these 96 that you, personally, are not comfortable answering, **please indicate that to us** in this list, and we will consider revisions appropriately.
2. Your list of 0-3 new items[1] that you would like to add to the survey.
    - Note that your new items *can* be but do not *need* to be anything you've previously suggested.
    - Please begin with the following sentence: `I would like to submit # new items for consideration.`
    - If your number of new items to suggest is zero, then you need not write anything else here.
    - Should you wish to have us include 1-3 additional items, please remember that nothing about sex, drugs, or performance in 431 can be asked, and:
        a. list each new item, being sure to specify the type (for instance, short answer, multiple choice, or checkbox) and the set of possible responses, as you did in the proposal.
        b. describe (in 2-3 complete sentences per new item) your reasons to include the item.
            – Good reasons would begin with a statement of what you intend to do. As an example of such a statement, consider `I wish to study the result of this new item as a quantitative outcome across groups established by current item #*** from the survey.` Or, perhaps, something like: `I wish to use this new item as a grouping variable to study current item #***.`
            – In either case, follow your statement with a short explanation as to why your new item's result is of interest, and is not already captured by the existing survey.

[1] We will not consider more than 3 new items from anyone, and are eager to hold the total set of new items to 25 or less, across all 66 students. I would argue that data related to each of the 66 accepted project proposals may be found in the existing set of 96 items.

Chapter 5

Task D (Survey Comparison Plan) Instructions

1. **A Word document submitted via Blackboard**, specifying the list of items from the survey that you want to be able to use in your analysis, using the **Template for Task C** available at https://sites.google.com/a/case.edu/love-431/home/projects/class-survey.
    - You need not do any analyses connected to the items you originally suggested, nor do you need to do analyses that mirror your original research questions.
    - The Template asks you to specify (by item number and name) the items you wish to you use in your analyses, for each of the six analyses you will complete for Study 1.
    - Task E has details. You need to complete either Analysis 1a or 1b, and then Analyses 2-6.
    - In addition to the items you select related to each Analysis, you will also select two backup quantitative variables, and two backup factors, as described in the Template.
    - Items with at least 10 possible responses will be treated as quantitative. Other items will be treated as categorical (factors.) For ordered categories, you can consider assigning a score to each response, then treating that score as quantitative.
        - You are permitted to categorize any quantitative item you choose.
        - You are permitted to collapse any categories in an item with more than 2 categories, as you choose.
        - Some items are part of multiple-item scales. If you want to use a scale, specify each item that would go into that scale in Task C.

Analysis	Variables needed
[1a] 2 means via paired samples	Two quantitative (outcomes)
[1b] 2 means via indep. samples	One quantitative (outcome) and one categorical (2 levels)
[2] ANOVA with Tukey	One quantitative (outcome) and one categorical (3-6 levels)
[3] Regression Model	Same as either [1b] or [2], plus one quantitative (covariate)
[4] 2x2 Table	Two categorical (2 level) variables
[5] JxK Table	Two categorical variables, one with 2-6, other with 3-6 levels
[6] 2x2xJ Table	Same as [4], plus one categorical with 3-6 levels

Chapter 6

Task E (Taking the Survey) Instructions

2. **Completion (Google form) of the final course survey**, available on November 7 by 5 PM.
    - The final item asks for your name, and the system is collecting your email address (you must be logged into Google via CWRU). These will be pruned from the survey before data sets are created.
    - You should answer all of the items. Please don't skip any items you can answer. Your colleagues need data.
    - If you want to save your work and return later, note that only the *first* item in each section of the survey must be completed for Google to let you submit your work. Once you've submitted a partially completed survey, you can return as often as you like before the deadline to finish up.

6.1 Receiving Your Study 1 Data (November 18)

- We will post **two** data files for you, each containing some of the variables you need.
- You will need to download both files, and then *combine* and tidy to suit your needs.
- The two files will be linked by the subject `id` number.
- We discuss combining two data sets, using `dplyr`, as part of the Data Management Tips section.

Chapter 7

Task F (Sharing Study 2 Data) Instructions

Task F requires you to share your data for Study 2. The model for this Task is Jeff Leek's Guide to Data Sharing. Specifically, Task D requires that you submit the following to Dr. Love via email by the deadline. Please make your email's subject: `431 TASK D for YOUR NAME`

1. a direct link to the raw data set (without any need for me to sign up for anything) or a .csv copy of the raw data set called `yourname-raw.csv`
2. a single .csv file with a name of your choice containing a clean, tidy data set for Study 2, along with
3. a Word or PDF file containing both
    a. a **codebook** section which describes every variable (column) and its values in your .csv file,
    b. a **study design** section which reminds (and updates) us about the source of the data and your research question.

7.1 The Raw Data Set

You need to show me the raw, de-identified data. The data are raw if you:

- Ran no software on the data and Did not manipulate any of the numbers in the data
- You did not remove any data from the data set other than to de-identify it and eliminate protected information and anything else that you cannot share
- You did not summarize the data in any way

A direct link (without me having to sign up for anything) is preferred. If this is not possible, send a .csv file of the raw data set, called `yourname-raw.csv`. Note that you should not send me any variables you have no chance of using in your analyses, but may include some variables you haven't made a final decision on.

7.2 The Tidy Data Set

Your .csv file should include only those variables you will actually use in your analysis of Study 2. Your .csv file should include one row per subject in your data, and one column for each variable you will use. Your data are tidy if each variable you measure is in its own column, and each different observation of that variable is in its own row, identifed by the subject `id`.

You need to provide:

1. a header row (row 1 in the spreadsheet) that contains full row names. So if you measured age at diagnosis for patients, you would head that column with the name `AgeAtDiagnosis` or `Age.at.Diagnosis`

instead of something like `ADx` or another abbreviation that may be hard for another person (or you, two years from now) to understand.

2. a study identification number (I would call this variable `id` and use consecutive integers to represent the rows in your data set) which should be the left-most variable in your tidy data.
3. a quantitative outcome with a meaningful name using no special characters other than a period (`.`), hyphen(`-`) or underscore (`_`) used to separate words, which should be the second variable in your data.
    - If you have any missing **outcome** values, **delete those rows** entirely from your tidy data set.
4. at least four predictor variables, each with a meaningful name using no special characters other than `.` or `_` to separate words, and the predictors should be shown in columns to the right of the outcome.
    - *Continuous* variables are anything measured on a quantitative scale that could be any fractional number.
    - *Ordinal categorical* data are data that have a fixed, small (< 100) number of levels but are ordered.
    - *Nominal categorical* data are data where there are multiple categories, but they aren't ordered.
    - Categorical predictors should read into R as factors, so your categories should include letters, and not just numbers. In general, try to avoid coding nominal or ordinal categorical variables as numbers.
    - Label your categorical predictors in the way you plan to use them in your analyses
    - *Missing data* are data that are missing and you don't know the mechanism. Missing data in the predictor variables are allowed, and you should code missing values in your tidy data set as `NA`. It is critical to report if there is a reason you know about that some of the data are missing. You should also not impute/make up/throw away missing observations on the predictor values in your tidy data set.
5. any other variables you need to share with me (typically this would only include things you had to use in order to get to your final choice of outcome and predictors.) Most people will not need to share any additional variables.

I will need to be able to take your submitted `.csv` file and run your eventual Markdown file (Task E) against it and obtain your results, so it must be completely clean. Because it is a `.csv` file, you'll have no highlighting or bolding or any other special formatting. If you have missing values, they should be indicated as NA in the file. If you obtain the file in R, and then write it to a .csv file, you should write the file without row numbers if you already have an identification variable. To do so, you should be able to use `write.csv(dataframeinR, "newfilename.csv", row.names = FALSE)` where you will substitute in the name of your data frame in R, and new (.csv) file name. Don't use the same name for your original data set and your tidy one.

7.3 The Codebook

For almost any data set, the measurements you calculate will need to be described in more detail than you will sneak into the spreadsheet. The code book contains this information. At minimum it should contain:

1. Information about the variables (including units! and codes for any categorical variables) in your tidy data set
2. Information about the summary choices or transformations you made or the development of any scales from raw data

By reading the codebook, I should understand what you did to get from the raw data to your tidy data, so add any additional information you need to provide to make that clear.

7.4 The Study Design

Here is where I want you to put the information about the experimental study design you used. You can and should reuse (and edit) the information you provided as part of the Proposal in this Codebook. The material you need here consists of three parts from the proposal, updated to mirror your current plan. Specifically, you should provide:

1. Your research question describes your outcome, your key predictor and other predictors, and the population of interest. It is probably easiest to follow one of these formats[1].

- What is the effect of `*your key predictor*` on `*your outcome*` adjusting for `*your list of other predictors*` in `*your population of subjects*`?
- How effectively can `*specify your predictors*` predict `*your outcome*` in `*your population of subjects*`? or

2. A thorough description of the data collection process, with complete details as to the nature of the variables, the setting for data collection, and complete details of any apparatus you used which may affect results that **has not already been covered** in the codebook materials.

3. Specification of the people and methods involved.

    a. Who are the subjects under study? How many are included in your final tidy data set?
    b. When were the data gathered? By whom?
    c. What caused subjects to be included or excluded from the study?

[1] You are welcome to move the clauses around to make for a clearer question.

Chapter 8

Task D Instructions

Task D requires you to share your data for Study 2. The model for this Task is Jeff Leek's Guide to Data Sharing. Specifically, Task D requires that you submit the following to Dr. Love via email by the deadline. Please make your email's subject: `431 TASK D for YOUR NAME`

1. a direct link to the raw data set (without any need for me to sign up for anything) or a .csv copy of the raw data set called `yourname-raw.csv`
2. a single .csv file with a name of your choice containing a clean, tidy data set for Study 2, along with
3. a Word or PDF file containing both
    a. a **codebook** section which describes every variable (column) and its values in your .csv file,
    b. a **study design** section which reminds (and updates) us about the source of the data and your research question.

8.1 The Raw Data Set

You need to show me the raw, de-identified data. The data are raw if you:

- Ran no software on the data and Did not manipulate any of the numbers in the data
- You did not remove any data from the data set other than to de-identify it and eliminate protected information and anything else that you cannot share
- You did not summarize the data in any way

A direct link (without me having to sign up for anything) is preferred. If this is not possible, send a .csv file of the raw data set, called `yourname-raw.csv`. Note that you should not send me any variables you have no chance of using in your analyses, but may include some variables you haven't made a final decision on.

8.2 The Tidy Data Set

Your .csv file should include only those variables you will actually use in your analysis of Study 2. Your .csv file should include one row per subject in your data, and one column for each variable you will use. Your data are tidy if each variable you measure is in its own column, and each different observation of that variable is in its own row, identifed by the subject `id`.

You need to provide:

1. a header row (row 1 in the spreadsheet) that contains full row names. So if you measured age at diagnosis for patients, you would head that column with the name `AgeAtDiagnosis` or `Age.at.Diagnosis` instead of something like `ADx` or another abbreviation that may be hard for another person (or you, two years from now) to understand.

2. a study identification number (I would call this variable `id` and use consecutive integers to represent the rows in your data set) which should be the left-most variable in your tidy data.
3. a quantitative outcome with a meaningful name using no special characters other than a period (`.`), hyphen(`-`) or underscore (`_`) used to separate words, which should be the second variable in your data.
    - If you have any missing **outcome** values, **delete those rows** entirely from your tidy data set.
4. at least four predictor variables, each with a meaningful name using no special characters other than `.` or `_` to separate words, and the predictors should be shown in columns to the right of the outcome.
    - *Continuous* variables are anything measured on a quantitative scale that could be any fractional number.
    - *Ordinal categorical* data are data that have a fixed, small (< 100) number of levels but are ordered.
    - *Nominal categorical* data are data where there are multiple categories, but they aren't ordered.
    - Categorical predictors should read into R as factors, so your categories should include letters, and not just numbers. In general, try to avoid coding nominal or ordinal categorical variables as numbers.
    - Label your categorical predictors in the way you plan to use them in your analyses
    - *Missing data* are data that are missing and you don't know the mechanism. Missing data in the predictor variables are allowed, and you should code missing values in your tidy data set as `NA`. It is critical to report if there is a reason you know about that some of the data are missing. You should also not impute/make up/throw away missing observations on the predictor values in your tidy data set.
5. any other variables you need to share with me (typically this would only include things you had to use in order to get to your final choice of outcome and predictors.) Most people will not need to share any additional variables.

I will need to be able to take your submitted `.csv` file and run your eventual Markdown file (Task E) against it and obtain your results, so it must be completely clean. Because it is a `.csv` file, you'll have no highlighting or bolding or any other special formatting. If you have missing values, they should be indicated as NA in the file. If you obtain the file in R, and then write it to a .csv file, you should write the file without row numbers if you already have an identification variable. To do so, you should be able to use `write.csv(dataframeinR, "newfilename.csv", row.names = FALSE)` where you will substitute in the name of your data frame in R, and new (.csv) file name. Don't use the same name for your original data set and your tidy one.

8.3 The Codebook

For almost any data set, the measurements you calculate will need to be described in more detail than you will sneak into the spreadsheet. The code book contains this information. At minimum it should contain:

1. Information about the variables (including units! and codes for any categorical variables) in your tidy data set
2. Information about the summary choices or transformations you made or the development of any scales from raw data

By reading the codebook, I should understand what you did to get from the raw data to your tidy data, so add any additional information you need to provide to make that clear.

8.4 The Study Design

Here is where I want you to put the information about the experimental study design you used. You can and should reuse (and edit) the information you provided as part of the Proposal in this Codebook. The material you need here consists of three parts from the proposal, updated to mirror your current plan. Specifically, you should provide:

1. Your research question describes your outcome, your key predictor and other predictors, and the population of interest. It is probably easiest to follow one of these formats[1].

- What is the effect of `*your key predictor*` on `*your outcome*` adjusting for `*your list of other predictors*` in `*your population of subjects*`?
- How effectively can `*specify your predictors*` predict `*your outcome*` in `*your population of subjects*`? or

2. A thorough description of the data collection process, with complete details as to the nature of the variables, the setting for data collection, and complete details of any apparatus you used which may affect results that **has not already been covered** in the codebook materials.

3. Specification of the people and methods involved.

    a. Who are the subjects under study? How many are included in your final tidy data set?
    b. When were the data gathered? By whom?
    c. What caused subjects to be included or excluded from the study?

[1] You are welcome to move the clauses around to make for a clearer question.

Chapter 9

Task H (The Portfolio) Instructions

Task H requires you to provide a written portfolio of materials, which you will also make use of in your final presentation.

9.1 Logistics

- Submit your portfolio via email to Dr. Love. Make your email's subject: `431 TASK E FOR YOUR NAME`.
- The portfolio should be contained in a single .zip file that contains each of the elements below.
- Name your .zip file YOURNAME-TASK E.zip
- The .zip file should contain
  - (for Study 1) a .csv, a .Rmd and a .doc/.docx/.pdf file generated from that .Rmd file, and
  - (for Study 2) a .Rmd and a .doc/.docx/.pdf file generated from that .Rmd file.

9.2 Materials for Task E

- [`.csv` file] A clean, tidy data set for Study 1, which will require combining the two data sets you are provided, dealing with any missing data and any necessary combination into scales on the variables in which you are interested.
- [`.Rmd` and `.doc/.docx/.pdf` files] The six required analyses for Study 1, as both a Markdown file and Word/PDF that work with the clean and tidy data set for Study 1.
- [`.Rmd` and `.doc/.docx/.pdf` files] The eight required analyses for Study 2, as both a Markdown file and Word/PDF that work with the clean and tidy data set you submitted in Task D.

9.3 The Six Required Analyses for Study 1 (The Survey)

The required analyses for the Project Survey that need to be in your Portfolio are:

1. A two-group comparison of population means (could use paired or independent samples)
2. An analysis of variance with Tukey HSD pairwise comparisons of population means across K subgroups, where $3 \leq K < 7$
3. A regression model to amplify the indepedent samples comparison in a or b by incorporating a quantitative covariate.
4. A 2x2 Table and resulting analyses for comparison of two population proportions in terms of relative risk, odds ratio and probability difference
5. A two-way JxK contingency table where $2 \leq J < 7$ and $3 \leq K < 7$ with an appropriate chi-square test

6. A three way 2 x 2 x J contingency table analysis whch will expand your 2x2 table from #4 and where $3 \leq \text{J} < 7$

A demonstration of an appropriate analysis for each of these pieces will be provided at https://sites.google.com/a/case.edu/love-431/home/projects/class-survey.

9.4 The Eight Required Analyses for Study 2 (Your Data)

For your portfolio presentation in Study 2 (Your Data) complete these steps:

0. Identify all the variables in your tidy data set that have missing (NA) values. Delete all observations with missing outcomes, and use simple imputation to impute values for the candidate predictors with NAs. Use the resulting imputed data set in all subsequent work.
1. Obtain a training sample with a randomly selected 80% of your data, and have the remaining 20% in a test sample, properly labeled, and using `set.seed` so that the results can be replicated later. Use this training sample for Steps 2-6 below.
2. Using the training sample, provide numerical summaries of each predictor variable and the outcome (with `Hmisc::describe`), as well as graphical summaries of the outcome variable. Your results should now show no missing values in any variable. Are there any evident problems, such as substantial skew in the outcome variable?
3. Build and interpret a scatterplot matrix to describe the associations (both numerically and graphically) between the outcome and all predictors. Use a Box-Cox plot to investigate whether a transformation of your outcome is suggested. Describe what a correlation matrix suggests about collinearity between candidate predictors.
4. Specify a "kitchen sink" linear regression model to describe the relationship between your outcome (potentially after transformation) and the main effects of each of your predictors. Assess the overall effectiveness, within your training sample, of your model, by specifying and interpreting the R^2, adjusted R^2 (especially in light of your collinearity conclusions below), the residual standard error, and the ANOVA F test. Does collinearity in the kitchen sink model have a meaningful impact? How can you tell? Specify the size, magnitude and meaning of all coefficients, and identify appropriate conclusions regarding effect sizes with 90% confidence intervals.
5. Build a second linear regression model using a subset of your four predictors, chosen by you to maximize predictive value within your training sample. Specify the method you used to obtain this new model. (Backwards stepwise elimination is a likely approach in many cases, but if that doesn't produce a new model, feel free to select two of your more interesting predictors from the kitchen sink model and run that as a new model.)
6. Compare this new (second) model to your "kitchen sink" model within your training sample using adjusted R^2, the residual standard error, AIC and BIC. Specify the complete regression equation in both models, based on the training sample. Which model appears better in these comparisons of the four summaries listed above? Produce a table to summarize your results. Does one model "win" each competition in the training sample?
7. Now, use your two regression models to predict the value of your outcome using the predictor values you observe in the test sample. Be sure to back-transform the predictions to the original units if you wound up fitting a model to a transformed outcome. Compare the two models in terms of mean squared prediction error and mean absolute prediction error in a Table, which Dr. Love will **definitely want to see** in your portfolio. Which model appears better at out-of-sample prediction according to these comparisons, and how do you know?
8. Select the better of your two models (based on the results you obtain in Questions 6 and 7) and apply it to the entire data set. Do the coefficients or summaries the model show any important changes when applied to the entire data set, and not just the training set? Plot residuals against fitted values, and also a Normal probability plot of the residuals, each of which Dr. Love **will be looking for** in your portfolio. What do you conclude about the validity of standard regression assumptions for your final model based on these two plots?

A demonstration of an appropriate analysis for each of these pieces is available at https://sites.google.com/a/case.edu/love-431/home/projects/your-data

Chapter 10

Task I (Your Presentation) Instructions

The presentation schedule is found at https://goo.gl/PivgQx. Arrive at Dr. Love's office (Wood WG-82L) at least 5 minutes early. If the door is open, please be sure that Dr. Love knows you are there.

You will give your final presentation in a 15-minute meeting with Dr. Love. This will involve materials from both of your studies, in a fairly regimented way, described below.

- You are welcome to bring either a printed presentation or (better) a functioning laptop which you can use to show me the key results as you describe them for each of the analyses in Study 1 and in Study 2 that you wind up discussing.
- You are welcome to show me results in the context of a Powerpoint-style presentation, if you prefer to develop one, or to show me results straight from your Markdown-created Word or PDF files in your portfolio. Whatever works for you - so long as I can see what you are talking about as you are talking, we'll be fine.
- The computers in my office will be busy while we are meeting, so I will NOT be able to pull up your portfolio or data while we are talking. You will have to be able to do that.
- It is 100% appropriate for you to ask questions before the presentation, of Dr. Love or the TAs. Please do. At the presentation, there will be a little time for Dr. Love to address any lingering questions after the main presentation, and he's eager to hear your questions at that time, too.

10.1 Study 1 Presentation (about 5 minutes, total)

In Study 1, you will first select your most interesting / intriguing result out of your six main analyses and present that, in about 90 seconds. In those 90 seconds, you should be showing me the highlights, specifically:

a. What question were you investigating?
b. What conclusion did you draw about that question?
c. What statistical method led you to that conclusion?

I will then ask you to present the results of one of the other five main analyses, in a similar way. You will need to come prepared to present this information for any of your six Study 1 analyses at a moment's notice, as you will not know in advance which of your other five main analyses I will ask for.

10.2 Study 2 Presentation (about 10 minutes, total)

In Study 2, you will start with telling me about the most important finding of your little study in four minutes. In these 4 minutes, you will tell me:

a. What your research question was
b. Why it was interesting to you (parts 1 and 2 combined should take no more than 30 seconds)
c. What your better model has to say about the answer to your research question
    - This should include a description of the predictors that wound up in your (final) model and the direction of each of their effects on your outcome. Show me the model as you're telling me about this.
    - This should also include a sense of how well the model predicted overall (R^2 is one good choice)
    - This should also include how well the residual plots for your final model fit regression assumptions. Show me the plots as you're telling me about this.
d. Your conclusions about rational next steps to learn more from these data, or what specific new data you now wish you'd had when you started the study.

For most of the remaining time, I will ask you about your study, and try to help you think through any problems you had in obtaining or interpreting analyses. You should come prepared to share any of the 8 steps of your analysis at a moment's notice, as we may want to look at any part of your work.