

HS237C Grading and Term Paper

Grading:

- 5% Class participation (including submission of paper proposals and descriptive statistics)
- 35% Homework assignments
- 60% Final paper

Attendance:

While class attendance is strongly encouraged, it is not required. However, students cannot expect to miss class on a regular basis and then have the TA or instructor give one-on-one instruction during office hours to make up for what was missed.

Homework:

Five homework assignments will be given, focusing on the empirical application of the models learned in class. You will be expected to run the models learned in class using the data we give you, and then interpret the results. You are encouraged to work together to run the analyses and discuss the problems, but you are expected to do the write-ups separately. ***Evidence that the write-ups were not done independently will be grounds for a reduction in the grade.***

Homework assignments will be graded \checkmark^+ , \checkmark , and \checkmark^- . The weight of each homework assignment in the overall grade varies depending on its length.

Paper proposal:

At the beginning of week 6, you will be asked to turn in a proposal summarizing your plans for the final paper. It should be approximately two pages (double-spaced) and contain sufficient information for us to evaluate the appropriateness of the methods, including the following:

- research question and specific hypotheses to be tested
- brief description of study design and data
- description of statistical methods, including specification of the regression equation
- discussion of any methodological limitations

Note: We strongly discourage hypotheses and analyses involving the testing of interaction terms, since substantial experience suggests that even the best students calculate the estimates incorrectly when looking at interaction terms in the context of nonlinear models. If you insist on examining interaction terms, we encourage you to work closely with the TA and the instructor to make sure that you are doing this correctly.

Descriptive statistics:

At the end of week 7, you will be asked to submit descriptive statistics (means and standard deviations for continuous variables, frequencies for categorical variables, minima and maxima where relevant) for the dataset you will use for your paper. If you do not have a clean analysis file created by this time, you are unlikely to be able to complete a class paper of acceptable quality by the deadline.

Paper:

Two hard copies of the paper and supporting materials (see below) should be submitted to the instructor by the due date. **Late papers are not acceptable and will be excused only in the most dire emergencies, so please do not count on being able to get an extension.**

A detailed template for the paper is available as a separate document. **Please pay careful attention to the relative importance placed on each section of the paper.** The paper needs only the briefest of introductions and does not need a background section, literature review, or references (except for any statistical references that may be needed). Please use your time and energy to improve the methods and results sections and not for convincing us that this paper is interesting or important. In HS237C, the goal of the class paper is a technical exercise to demonstrate that you learned how to properly apply and interpret at least one new statistical method, not to produce a publishable paper.

A detailed explanation of what is expected for this paper is included in the grading section below. The statistical methods used in this paper must go beyond what you learned in HS237B. **Please check with the instructor that the model/method you are proposing to use meets this criterion.** The difficulty of the statistical model used will be taken into account in the grading. Although your regression models must *control* for all relevant covariates, it is usually easiest to only *report* the estimates for a single regressor or subset of regressors of interest (e.g., income, or type of insurance) in the table and text, and just describe what the other covariates were in a footnote to the table.

Grading of Paper:

Points for the paper will be assigned according to the objectives outlined below. Please remember that your paper must be clearly written (with proper use of grammar and punctuation), and be sure to proofread and use spell check. If we do not understand your writing, you will lose points for each objective that is not clearly written.

Your research should meet the following objectives:	Possible Points:
Clearly define the research question and hypotheses (make sure that they are testable, given your study design and data). Include a discussion of the hypothesized direction of the relationship being examined.	5
Explain how the regression model and hypotheses are justified with a theoretical model.	5
Describe the data, variables and the statistical methods used in the analyses. A clear explanation of why they are appropriate for the problem is required. Include the specification of your regression models and how it corresponds to the theoretical model.	10
Provide a clear, user-friendly presentation of the results in both text and tables. The appropriateness of results presented (i.e., are the numbers presented interpretable by themselves) will also be considered.	15
Present the correct interpretation of results.	15
Adequately describe the methodological limitations of the research. Include limitations that cannot be addressed within the scope of the study and the likely nature of the bias.	10
Total possible points	60

Tables for Paper:

You must include the following tables in your paper:

- 1) A table summarizing your regression specification, with columns representing each of the following:
 - A list of all of the variables that theoretically influence your outcome measure, including any important interactions
 - The hypothesized direction of the effect of each variable
 - A brief rationale for this hypothesis; alternatively, this information can be moved to the text, if there is insufficient room in the table. Do NOT rely solely on prior empirical literature as the rationale for a hypothesized regressor effect, but rather, briefly explain the conceptual reason in your own words
 - The measured variable(s) (proxy or proxies) for the theoretical variable. If there is no empirical variable corresponding to the theoretical measure, write in N/A (and make sure you discuss the likely direction of the omitted-variable bias in the limitations section).

Do not use any form of stepwise regression analysis unless you have multiple empirical proxies for the same theoretical variable and your sample size is too small (or the collinearity too high) to include all of them in the empirical specification.

- 2) A table with descriptive statistics for the sample, showing each dependent and independent variable used in the analysis, the sample size for the variable, and the mean and standard deviation (or frequency) for the variable. Present unweighted data, so that the true distributions of the variables can be determined. If desired, the table can report these statistics for subsamples defined by values of the dependent variable or the primary regressor of interest.
- 3) One or more tables with estimates based on the regression models

These tables must include the following information:

- The dependent variable
- Which sample was used (if there is more than one sample)
- Sample size(s)
- An explanation of the estimates being shown, either in the title or footnote (e.g., "Relative risks and 99% confidence intervals")
- Footnote explaining what type of regression model was used and any other relevant information (e.g., "Estimates are derived from a multinomial logit regression with hospital-level clustering corrections.")
- Footnote explaining what the omitted categories for the outcome measure and/or the regressor of interest are (if applicable and unclear from the table).

- Footnote listing the other independent variables controlled in the regression (if unclear from the table itself).
- Coefficient estimates if easily interpretable; otherwise, you must present more informative estimates. For example, for a multinomial logistic regression, you might report relative risks, odds ratios (when appropriate), level changes in the predicted probabilities, or elasticities.
- Confidence intervals or standard errors around the estimates of interest

STATA programs and output:

You must separately include a clean and final copy of all STATA programs and log files used to run your regression models. You do not need to include programs and log files used to create the dataset or run the descriptive statistics. These programs will be reviewed and the accuracy of the code will be considered in assigning a final grade.

Note: One of the most common problems (and reasons for getting points marked down) is the failure to ensure that the sample size is the same for all of your analysis variables before you run your programs. In many cases this will end up leading to out-of-sample predictions and so forth. You can avoid this by making sure that your program to create the analysis dataset is entirely separate from the programs you use to run your statistical analyses. Your missing data should be dealt with already in the program to create the analysis dataset, either by choosing an imputation method or by dropping observations with missing data (complete-case analysis).