NUTR 307 -- Regression Analysis for Nutrition Policy  
Spring 2019

Class Meetings:  
Mondays/Wednesdays, 1:30 - 3:00 pm  
Jaharis, Behrakis Auditorium, Boston Campus

Instructor(s):  
Parke Wilde, Jaharis #134  
617-636-3495, parke.wilde@tufts.edu

Office hours:  
Wed 3:00-4:00 p.m. (open office hours) and 4:00-5:00 p.m. (appointments). Or schedule appointment by email. Jaharis #134

Teaching Asst.:  
Sara John: Sara.John@tufts.edu  
Kate Schneider: Kate.Schneider@tufts.edu

Office hours:  
Sara: Thurs 8:30-10:00 am or by appt. Kneeland #835  
Kate: Fri 11:30 am-1:00 pm or by appt. Kneeland #835

Graduate Credits:  
3 credits

Prerequisites:  
Biostatistics. NUTR 207 or NUTR 209 or permission of the instructor. For a review of material covered in NUTR 207, see the text used by Prof. Sean Cash, Statistical Methods for the Social Sciences, 5th edition, A. Agresti, 2017.

Students should be familiar with mathematics at the pre-Calculus level. As suggested for NUTR 207, see the Columbia primer or the Educational Testing Service’s math preparation materials (links on the course site, under the folder of “Readings and Links”). No calculus and matrix algebra needed.

Course Description:  
Part two of a one-year, two-semester course sequence in statistics. This course is intended for students whose main focus is non-experimental or survey-based research. The course covers research design, simple linear regression, multiple regression, analysis of variance, non-linear functional forms, heteroskedasticity, complex survey designs, and real-world statistical applications in nutrition science and policy. Students will make extensive use of Stata.

1 Please put “NUTR 307” in email subject line for correspondence related to this course. NOTE: Students cannot receive credit for both NUTR 307 and its second semester counterpart NUTR 309.
Course Objectives:

Goal:
Master the basic statistical methods most useful for non-experimental and survey-based research in nutrition science and policy.

Topics:

Review
1. Review. Review the basic building blocks of statistics: discrete and continuous variables, expectation, conditional expectation, variances, covariances, and the normal distribution.
2. Simple linear regression. Review the theory and practice of simple linear regression, with one dependent variable and one independent variable.

Models
3. Multiple linear regression. Learn the practice of multiple linear regression. Intuition and understanding will be developed without matrix algebra, mainly by analogy with simple linear regression.
4. Analysis of Variance (ANOVA). Study problems where the dependent variable is continuous while key explanatory variables are binary or categorical. For example, a binary variable may identify whether a subject is in a study group or control group.
5. Brief introduction to Logistic Regression (Logit) and Linear Probability Models. Learn statistical analyses that are appropriate when the dependent variable is binary.
6. Panel data. Introduce statistical methods for longitudinal or panel data.
7. Introduction to complex surveys. Correctly interpret surveys that have complex sampling designs, using sampling weights and design-corrected standard errors.

Solving problems
8. Confounding variables and problems with functional form. Learn solutions to the most serious specification problems, which cause bias in all parameter estimates.
9. Heteroskedasticity. Recognize that heteroskedasticity biases standard errors but does not bias regression parameter estimates. Learn to use robust standard errors in large samples.
10. Non-normality. Recognize that non-normality causes problems for hypothesis tests of regression parameter estimates in small samples, but not in large samples.
11. Instrumental variables and program evaluation.
12. Introduction to research design.

Texts or Materials:

Course text. Introductory Econometrics: A Modern Approach (fourth edition through sixth edition), by Jeffrey M Wooldridge (Mason, OH: Thomson South-Western). Purchase is not mandatory. Six copies (editions 4-6) are on reserve at
HHSL library. If you purchase used, the 5th and 6th editions seem to have the most copies available and range from about $40-80. Older editions are close enough if you find them.

**Applied readings.** Several empirical research articles and reports will illustrate how the methods learned in this class are applied to real-world problems in: 1) agriculture, food, and the environment, and 2) food policy and applied nutrition.

**Academic Conduct:**

Each student is responsible for upholding the highest standards of academic integrity, as specified in the Friedman School’s Policies and Procedures manual (http://nutrition.tufts.edu/student/documents) and Tufts University policies (http://uss.tufts.edu/studentAffairs/documents/HandbookAcademicIntegrity.pdf). It is the responsibility of each student to understand and comply with these standards, as violations will be sanctioned by penalties ranging from failure on an assignment and the course to dismissal from the school.

Collaboration on the problem sets is encouraged, but beware. For some strange reason, watching a classmate complete a problem set feels like completing it oneself, but without much benefit. It is best to work separately first.

**Poll Everywhere (in-class response system)**

Poll Everywhere is a classroom response system that your instructor will use to ask questions during lectures and see the classroom’s responses in real-time. We will use it approximately every other class. Poll Everywhere access is free but you must set up your account with your Tufts email address. Please follow the instructions here, also listed in the Week 1 module on the course website prior to the second class, and bring a web-enabled device to every class so you can participate. To borrow a device, see the TA at least a day in advance.

**Statistical software:**

Stata statistical software will be used for this course. Stata 15 is available on the computers in the Jaharis student room, in the computer lab in Sackler 510, and on laptops at the HHSL library. Students can check the availability of computers in Sackler 510 on the white board behind the 5th floor library desk. Stata is also available on the computers at the Eaton computing lab on the Medford campus. If you’d like Stata on your own computer, Stata is available for purchase at: https://access.tufts.edu/software/stata. One option is the one-year license of Stata/IC version 15, priced around $125.
Laboratories:
There will be one mandatory laboratory, divided into 2 sections (each student attends just one section), on Jan. 30 from 3:30-5:00 pm & Feb 4 from 11:30am-1:00 pm in Sackler 216A. For selected other weeks, TA office hours will be scheduled in the Sackler computer lab in Sackler for Stata-related questions.

Course website:
Students who are registered for the course through SIS have been added to the course site automatically. If you need to be added to the course site manually, the TAs will be happy to do so.

Assessment and Grading:
Homework (30%)
Quizzes (45% for 3 quizzes)
Final examination (25%)

Assignments and Submission Instructions:
Weekly problem sets:
Exercises in statistics are like practice in learning a musical instrument. Some people enjoy them, and some do not, but they are utterly unavoidable. Weekly problem sets are submitted through the course site, due Sunday evening (any time before that date is fine). There is a 10% score penalty for late submission. It is encouraged to work on homework from Thursday through Friday (or perhaps Saturday) each week, for better access to TA consultation, and also for sanity.

This course uses real data sets relevant to several Friedman School programs, especially Agriculture, Food, and the Environment (AFE) and Food Policy and Applied Nutrition (FPAN). All necessary software will be provided in the student computer room in Jaharis, and in the Sackler computer library, and datasets will be available online or provided on the course website.

Quizzes and examination:
There will be three 45-minute quizzes and a 90-minute final examination during finals week (May 6). Please plan ahead and avoid these dates for field trips, conference travel, and best friends’ weddings. The questions for quizzes and the exam will be similar in style and substance to the problem sets. I have no desire to demonstrate originality in quizzes and examinations. For quizzes, calculators allowed, but no cell phones.

Accommodation of Disabilities:
Students with documented disabilities are entitled to academic accommodation appropriate to their needs. If you require accommodations for this course, please contact me confidentially prior to the end of the second week of classes.
**Course Schedule** (subject to future modification):

<table>
<thead>
<tr>
<th>Week # Class Dates</th>
<th>Topic</th>
<th>Readings</th>
<th>Lab</th>
<th>HW/Quiz</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Jan-16,23</td>
<td><strong>Multiple regression: estimation</strong></td>
<td>Wooldridge, Ch 3</td>
<td>-</td>
<td>HW1</td>
</tr>
<tr>
<td>2 Jan-28,30</td>
<td><strong>Multiple regression: inference</strong></td>
<td>Wooldridge, Ch 4</td>
<td>Lab</td>
<td>-</td>
</tr>
<tr>
<td>3 Feb-4,6</td>
<td><strong>Large samples, functional form, goodness-of-fit</strong></td>
<td>Wooldridge, Ch 5, Ch 6</td>
<td>-</td>
<td>HW2</td>
</tr>
<tr>
<td>4 Feb-11,13</td>
<td><strong>Analysis of variance (categorical variables)</strong></td>
<td>Wooldridge, Ch 7</td>
<td>-</td>
<td>Quiz 1 (Feb 13)</td>
</tr>
<tr>
<td>5 Feb-20,21</td>
<td><strong>Putting it together: reading regression analysis (I)</strong></td>
<td>Two articles on course site.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>6 Feb-25,27</td>
<td><strong>Heteroskedasticity</strong></td>
<td>Wooldridge, Ch 8</td>
<td>Lab office hours</td>
<td>HW3</td>
</tr>
<tr>
<td>7 Mar-4,6</td>
<td><strong>Specification problems</strong></td>
<td>Wooldridge, Ch 9</td>
<td>-</td>
<td>HW4</td>
</tr>
<tr>
<td>8 Mar-11,13</td>
<td><strong>Longitudinal or panel data</strong></td>
<td>Wooldridge, Ch 13</td>
<td>-</td>
<td>Quiz 2 (Mar 13)</td>
</tr>
<tr>
<td>Recess</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>9 Mar-25,27</td>
<td><strong>Complex surveys</strong></td>
<td>Stata documentation on course site.</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>10 Apr-1,3</td>
<td><strong>Logit</strong></td>
<td>Wooldridge, Ch 17</td>
<td>Lab office hours</td>
<td>HW5</td>
</tr>
<tr>
<td>11 Apr-8,10</td>
<td><strong>Putting it together: reading regression analysis (II)</strong></td>
<td>Two articles on course site.</td>
<td>-</td>
<td>HW6</td>
</tr>
<tr>
<td>12 Apr-17,22</td>
<td><strong>Instrumental variables and propensity scores</strong></td>
<td>-</td>
<td>-</td>
<td>Quiz 3 (Apr 22)</td>
</tr>
<tr>
<td>13 Apr-24,29</td>
<td><strong>Introduction to research design</strong></td>
<td>-</td>
<td>-</td>
<td>HW7</td>
</tr>
</tbody>
</table>

**Final Examination.** May 6 (Mon). Please plan ahead for this date.
Detailed Week–by-week Schedule
* This schedule is subject to modification at the instructor’s discretion.

***

Preparation for NUTR 307

Topic: Principles of Statistics

Objectives:

- Charts
  - Be able to interpret charts showing values of a single variable. Recognize which axis shows the variable of interest (such as the vertical axis in a vertical bar chart).
  - Be able to interpret histogram charts for a single variable. Recognize that the horizontal axis typically shows the variable of interest, and the vertical axis shows the frequency with which values of this variable occurs.
  - Be able to interpret charts showing values for two variables (such as scatter plots).

- Mathematics review
  - Know how to convert a proportion into a percentage, and vice versa.
  - For any continuous random variable, given a value at time 0 and time 1, be able to explain the percentage change.
  - Given a percentage at time 0 and time 1, understand the distinction between a percentage change and a percentage-point change.
  - Know the meaning of a logarithm (base 10) and a natural logarithm (with a base of e).

- Random variables
  - For a continuous random variable, understand the meaning of “normally distributed.”

- Samples and populations
  - Know and understand the equation for a population mean (which also happens to be the equation for a sample mean, useful for estimating the population mean). Given sample data, be able to estimate a population mean.
  - Know the equation for a population variance. Know the equation for a sample variance, when it is used to estimate a population variance. Be able to compute a sample variance. Understand that a variance is an average of squared deviations from means. Given some sample data, be able to estimate a population variance.
  - Know the equation for a sample standard deviation, when used to estimate a population standard deviation. The standard deviation
is simply the square root of a variance. Given some sample data, be able to estimate a population standard deviation.

- Know the equation for a sample covariance, when used to estimate a population covariance. Understand that the covariance of x and y is an average of the products of the deviations from means for x and y. Given some sample data on x and y, be able to estimate a population covariance.
- Know the equation for a correlation. Given some sample data on x and y, be able to estimate a population correlation.
- Know the meaning of a standard error of a mean. Be able to explain the difference between a standard error of a mean and a standard deviation of a random variable.

- Hypothesis testing
  - For a simple test of the difference in two means, know how to state a null hypothesis. Know how to state a one-tailed or two-tailed alternative hypothesis.
  - Know the equation for a t-statistic. Given a t-statistic for a hypothesis test, decide whether to reject the null hypothesis.
  - Know the meaning of a p-value. Given a p-value for a hypothesis test, decide whether to reject the null hypothesis.
  - Given a statistic and a confidence interval, conduct a hypothesis test.

- Simple linear regression
  - Be able to interpret the slope coefficient of a simple linear regression model (using the correct units in an interpretation sentence).

Reading: Statistics Review Materials on Course Site

Activities for this week: Self-assessment using “Preparation for NUTR 307.”

***

Week 1:

Topic: Multiple regression: estimation

Objectives:

- Recognize the wide diversity of useful applications for regression tools in non-experimental studies in food policy, nutrition policy, applied nutrition, agriculture, food, and environmental studies.
- Given a regression coefficient, be able to provide a correct interpretation sentence, including the correct units.
- Unbiasedness is the “number one” good quality we wish to find in a regression model. Know the meaning of unbiasedness and the assumptions required for unbiased regression estimates.
• Use the “3-variable game tool” to explore the consequences of failed assumptions, focusing specifically on circumstances that cause bias in a coefficient for a simple regression model.
• Understand how R-square is computed and what it means.
• Be able to interpret regression results as they appear in either Stata output or published tables from journal articles.

Reading: *Wooldridge, chapter 3*

Activities for this week: Homework 1 (Due Sunday, Jan 27, any time)

***

Week 2:

**Topic:** Multiple regression: inference

**Objectives:**
• Precision is the “number two” good quality we wish to find in a regression model. Understand how to assess the precision of a coefficient estimate.
• Understand each part of the equation for the variance of a regression coefficient estimate var(\( \hat{\beta}_1 \)) -- though you need not memorize this equation. Know the factors that make regression coefficients more or less precise.
• Know the meaning of the standard error se(\( \hat{\beta}_1 \)), and recognize that it can be computed as the square root of var(\( \hat{\beta}_1 \)). Know what var(\( \hat{\beta}_1 \)) and se(\( \hat{\beta}_1 \)) tell us about the precision of estimates.
• Be able to carry out a hypothesis test using four steps: (1) state a null hypothesis, (2) state an alternative hypothesis, (3) compute a test statistic whose distribution is known under the assumption that the null hypothesis is true, and (4) state a conclusion regarding the truth of the null hypothesis in the population.
• Find the information needed for hypothesis tests from published results tables or default Stata output when possible (t statistics and F statistics).
• Be able to request additional hypothesis tests from Stata when necessary (F statistics).
• Be able to compute confidence intervals from regression results, and know the interpretation of confidence intervals.
• Know the 5 assumptions needed for computing the “usual” regression standard errors, and the 6 assumptions needed for conducting the “usual” hypothesis tests in small samples.

Reading: *Wooldridge, chapter 4*

Activities for this week: Lab if you signed up for the Jan. 30 session (from 3:30-5:00 pm in Sackler 216A)
Week 3:

**Topic:** Large samples, functional form, goodness-of-fit

**Objectives:**
- Know the central limit theorem and be able to explain how the central limit theorem makes it easier to satisfy the assumptions of the regression model in large samples.
- Know the meaning of consistency and the assumptions required for consistent regression estimates.
- Be able to identify which difficulties in regression analysis are or are not remedied by large sample sizes.
- Recognize that the actual conduct of hypothesis tests is similar in small and large samples.
- Be able to explain quantitatively how regression coefficients change if the scale of a dependent or explanatory variable changes.
- Interpret regression coefficients when the dependent or explanatory variable is in logarithmic form.
- Use quadratic terms to estimate non-linear relationships between an explanatory variable and an outcome variable.
- Use interaction terms to estimate the effects of two variables that have synergistic effects.
- Use regression results to predict the value of an outcome variable, when given particular values of the explanatory variables.
- Be able to describe the precision of such predictions for (a) a particular observation or (b) the estimated expected value of the outcome variable conditional on the explanatory variables.

**Reading:** Wooldridge, chapters 5 & 6

**Activities for this week:**
- Lab if you signed up for the Feb 4 session (11:30am-1:00 pm in Sackler 216A)
- Homework 2 (Due Sunday, Feb 10, any time)

***

Week 4:

**Topic:** Analysis of variance (categorical variables)

**Objectives:**
- Understand the relationships between two categorical variables, including the meaning of joint distributions and conditional distributions.
- Interpret regression coefficients when an explanatory variable is binary or categorical.
- Using an example of one continuous explanatory variable and one binary explanatory variable, explore once again the circumstances in which regression coefficients may be biased.
• Estimate regression models with categorical explanatory variables using both ANOVA tools and regression tools in Stata, and confirm that the two approaches are equivalent.
• Use a linear probability model to estimate regression models when the outcome is a binary variable.

Reading: Wooldridge, chapter 7
Activities for this week: Quiz 1 (Wed, Feb 13)

***

Week 5:
Topic: Putting it together: reading regression analysis (I)
Objectives:
• In applied examples of published research, recognize whether the organization of quantitative material is typical or atypical.
• For descriptive statistics and regression results, be able to interpret tabular estimates (including locating additional information if needed for correct interpretations).
• Recognize common variations in the display of quantitative information.
• Read article conclusions and judge for yourself whether they are justified on the basis of statistical evidence presented.

Reading: Two articles on course website.
Activities for this week: Lab office hours may be available this week.

***

Week 6:
Topic: Heteroskedasticity
Objectives:
• Know the consequences of heteroskedasticity, without either overstating or understating these consequences.
• Test for the presence of heteroskedasticity in empirical examples.
• When sample size is large, use robust standard errors to compensate for the problem of heteroskedasticity.
• Understand the motivation for weighted least squares models in the presence of heteroskedasticity (but it is beyond the scope of this course to know weighted least squares models in detail).

Reading: Wooldridge, chapter 8
Activities for this week: Homework 3 (Due Sunday, Mar 3, any time)
Week 7:

**Topic:** Specification problems, functional form, and regression discontinuity

**Objectives:**
- Test for particular non-linearities or synergies by adding additional quadratic terms or interaction terms.
- Explain the consequences of measurement error in dependent and explanatory variables, recognizing how these consequences depend on assumptions about the nature of the measurement error.
- Understand the effects of non-response and identify situations in which non-response does and does not lead to biased coefficient estimates.
- Quantify non-response problems using response rates.

**Reading:** Wooldridge, chapter 9

**Activities for this week:** Homework 4 (Due Sunday, Mar 10, any time)

---

Week 8:

**Topic:** Longitudinal or panel data

**Objectives:**
- Understand how longitudinal or panel data are arranged and indexed with subscripts i and t.
- Know the consequences in Ordinary Least Squares (OLS) of ignoring unobserved individual-level time-constant confounding variables.
- Use fixed effects models to remedy the problem of unobserved individual-level time-constant confounding variables.
- Analyze panel data using the panel data tools in Stata.
- Understand the purpose of random effects estimators for panel data, including how this purpose differs from that of the fixed effects estimator.

**Reading:** Wooldridge, chapter 13

**Activities for this week:** Quiz 2 (Wed, Mar 13)

---

Spring Recess
**Week 9:**

**Topic:** Complex surveys

**Objectives:**
- Recognize the structural similarity between complex survey data discussed this week and the panel data discussed in the previous week.
- Understand both stratification and clustering in complex surveys.
- Be able to define design effects, effective sample size, and intra-cluster correlation coefficients.
- Know the consequences of estimating OLS with the usual standard errors when the data come from complex surveys.
- Use the survey procedures in Stata to compute correct standard errors for descriptive statistics and regression estimates.
- Understand how complex surveys design choices are related to the economics of cost-per-cluster and cost-per-observation.

**Reading:** Introduction, *Stata Survey Reference Manual (Release 13)*.

**Activities for this week:** Lab office hours may be available this week.

***

**Week 10:**

**Topic:** Logit, probit, and a second look at the linear probability model

**Objectives:**
- Know the meaning of probability, odds, and log odds, and be able to translate from one to another in quantitative examples.
- Interpret odds ratios as they appear in logit model tabular output.
- Understand the equation for a logit model.
- For logit model coefficients, give interpretation sentences in terms of both log odds and odds ratios.
- Compare results from logit models and linear probability models.

**Reading:** *Wooldridge, chapter 17*

**Activities for this week:** Homework 5 (Due Sunday, Apr 7, any time)

***

**Week 11:**

**Topic:** Putting it together: reading regression analysis (II)

**Objectives:**
- In applied examples of published research for panel data and logit analysis, be able to interpret tabular estimates (including locating additional information if needed for correct interpretations).
• Read conclusions of articles with panel data and logit methods, and judge for yourself whether they are justified on the basis of statistical evidence presented.

**Reading:** 2 articles on course website.  
**Activities for this week:**  Homework 6 (Due Sunday, Apr 14, any time)  

***

**Week 12:**  
**Topic:** *Instrumental variables and propensity scores*  
**Objectives:**  
• Understand the 2 critical assumptions for a valid instrument.  
• Be able to use appropriately both linear regression and instrumental variables to measure the effect of a program intervention or treatment.  
• Recognize the assumptions needed for propensity score matching, and distinguish them from the assumptions needed for instrumental variables.  
**Reading:** *None*  
**Activities for this week:**  Quiz 3 (Mon, Apr 22)  

***

**Week 13:**  
**Topic:** *Introduction to research design*  
**Objectives:**  
• Distinguish between the effect of a treatment on the treated and the average treatment effect in a population.  
• Recognize the implications of self-selection into a particular treatment or intervention, and identify circumstances when this self-selection does or does not cause bias in program impact estimates.  
• Consider novel sources of administrative data and other large-sample data files with distinctive features.  
**Reading:** *To be determined.*  
**Activities for this week:**  Homework 7 (Due Sunday, April 28, any time)  

***

**Final Examination:** May 6, 1:30 - 3:00 pm.