Credits and Prerequisites

This course is 4 credits. It combines approximately 120 hours of reading, online activities, computer activities, and assignments. The enforced prerequisite for this course is ST516 Foundations of Data Analytics.

Course Content

- This course adds methods for modeling quantitative variables and making predictions to your data analytics tool-kit. Specific topics include:
  - Data analysis
  - Linear regression
  - Linear mixed effects models
  - Imputation for missing data
  - Prediction and cross-validation
  - Scaling up to large datasets

- Other key elements of the course are:
– Gaining experience using R to perform exploratory, inferential and predictive data analytics
– Using simulations to understanding the properties of estimators and the performance of predictions

University and Department Policies

Statement Regarding Students with Disabilities Oregon State University is committed to student success; however, we do not require students to use accommodations nor will we provide them unless they are requested by the student. The student, as a legal adult, is responsible to request appropriate accommodations. The student must take the lead in applying to Disability Access Services (DAS) and submit requests for accommodations each term through DAS Online. OSU students apply to DAS and request accommodations at our Getting Started with DAS (http://ds.oregonstate.edu/getting-started-das) page.

Statement of Expectations for Student Conduct These can be found at http://studentlife.oregonstate.edu/studentconduct/offenses-0

Learning Outcomes

• Identify the underlying assumptions of the multiple linear regression model and perform diagnostics to evaluate the validity of these assumptions.
• Recognize the situations in which to use different data analytic tools, and be able to appropriately implement those tools in R.
• Recognize the situations in which adding random effects to a statistical model is an appropriate approach; implement, evaluate and interpret mixed effects models in R.
• Address missing data using data imputation methods.
• Make and evaluate predictions using regression models.
• Identify issues with scaling multiple regression analysis to large datasets; implement methods for large data sets.
• Perform a complete data analysis—describing a dataset and research questions; selecting between candidate models; evaluating model assumptions and taking appropriate remedial action; interpreting and reporting results.
Course Policies

Canvas. This course will be delivered via Canvas, where you will interact with your classmates, your teaching assistant (TA) and the instructor. Within the course Canvas site you will access the syllabus, learning materials and tutorials. You will also be able to discuss issues/ask questions; submit assignments; take exams; and email other students, the TA and the instructor. To preview how an online course works, visit the Ecampus Course Demo (http://ecampus.oregonstate.edu/coursedemo). For technical assistance, see http://ecampus.oregonstate.edu/services/technical-help.htm.

Course Content Materials. Course content material will be available on a week-by-week basis. Each week, you will read material from the assigned materials and go through three sets of “lecture” notes. These notes will include some video and some R tutorials (more below about R), as well as overhead slide materials with key points from the weekly readings.

Computer labs. These are weekly exercises intended to teach you how to create plots; perform data exploration and statistical analyses; and perform computer simulations to understand the properties of estimators and the performance of predictions. Here is nothing to hand in from the labs, but the material covered in each lab will usually be needed to complete the homework assignment for that week.

For the computer lab component of the course you will use the package R, which is an open source programming environment. An easy way to access R is through the University Virtual Computer Lab (“umbrella”); see http://oregonstate.edu/is/mediaservices/scf/virtual-lab for instructions. Students may also download R for use on their own computers; it is freely available for Windows, OSX, and a variety of UNIX platforms at http://www.r-project.org.

Discussion Boards. There will be a discussion board set up for each week of the course. Please use these boards to interact with other students in the course, the TA and the instructor. As part of your course grade, you are expected to participate (in a substantive way) in the discussion boards each week. This can take the form of asking or answering questions about the course material, or simply writing a comment about something to make sure that you’re understanding a particular topic correctly. There is a separate discussion board set up for asking/answering “how to” questions about R.

Homework. The homework assignments will require that you use R to perform data analyses, and that you write summaries of those analyses. You should fashion your write-ups as reports to a supervisor: they should be typed and well-organized; raw computer output should not generally be included (unless it’s specifically requested), but rather summarized in words or using a table or plot. Plots or other output should be inserted into your write-up at the appropriate junctures. There will be a closing time after which each homework assignment will no be accepted (i.e., you will no longer be able to submit through Canvas)—please be aware of these times! You are encouraged to attempt all homework
problems on your own. After doing so, you are welcome to discuss your solutions with other students on the discussion board, though all work that you turn in must be written in your own words.

Exams. There will be one midterm exam at the end of the fifth week of the course and one final exam. All exams are open-book, open-note and they will require that you do some work in R. You must work on the exams by yourself.

Final Project. For the final project, you will perform a data analysis and write a report of your findings. A dataset and a series of questions will be provided, and you are asked to work on this project on your own.

Course grade. Homework, 30%; Discussion participation, 10%; Midterm, 20%; Project, 10%; Final, 30%.
# Course Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Week 1: Introductions; Simple Linear Regression</td>
<td>TBD</td>
</tr>
<tr>
<td>2</td>
<td>Week 2: Multiple Linear Regression</td>
<td>TBD</td>
</tr>
<tr>
<td>3</td>
<td>Week 3: Assumptions and Diagnostics</td>
<td>TBD</td>
</tr>
<tr>
<td>4</td>
<td>Week 4: Model Violations</td>
<td>TBD</td>
</tr>
<tr>
<td>5</td>
<td>Week 5: Advanced Topics, Model Interpretation</td>
<td>TBD</td>
</tr>
<tr>
<td>6</td>
<td>Week 6: Making Predictions</td>
<td>TBD</td>
</tr>
<tr>
<td>7</td>
<td>Week 7: Dealing with Missing Data</td>
<td>TBD</td>
</tr>
<tr>
<td>8</td>
<td>Week 8: Linear Mixed Effects Models</td>
<td>TBD</td>
</tr>
<tr>
<td>9</td>
<td>Week 9: Scaling up to Large Datasets</td>
<td>TBD</td>
</tr>
<tr>
<td>10</td>
<td>Week 10: Putting it all Together: Final Project</td>
<td>TBD</td>
</tr>
</tbody>
</table>