DSCI 561: Regression I

Contents

Many questions in data science are focused on how variables interact and are related. In particular, in DSCI 561 we will study the association of an outcome variable with other measured input variables using a linear regression. The course focuses on introducing new useful quantities, called the linear regression coefficients, how to estimate and interpret these quantities using a random sample, and how to establish these estimates' sampling distributions to measure their uncertainty. Since the input variable of a linear regressions can be a categorical variable defining groups (e.g., students vs non-students), we will see that some of the concepts covered in DSCI 561 are a continuation of some topics learned in DSCI 552.

The course emphasizes the distinction between generative and predictive models and uses estimated linear regression models to make inference about the population and to predict in-sample and out-of-sample observations. Students will be able to recognize when linear models are useful in practice and to establish connections with methods learned in other courses.

Slack Channel: https://ubc-mds.slack.com/messages/561_regr-1

Course Learning Outcomes

By the end of the course, students are expected to be able to:

1. Define, construct and interpret linear regression models to draw conclusions on practical examples.
2. Explain the difference between different types of explanatory variables and interpret their model parameters appropriately.
3. Explain the least square (LS) principle and apply different methods to find the optimal model.
4. Fit a linear regression in R, obtaining estimates of the regression coefficients and predictions of the response, especially using the `broom` package in R.
5. Find the sampling distributions of the coefficients' estimators using both asymptotic theory and bootstrapping, and use them to construct and interpret confidence intervals and prediction intervals.
6. Apply different tests of hypotheses (e.g., F-test, t-test) when they are useful in practice based on asymptotic theory.
7. Evaluate the trustworthiness of asymptotic results (sampling distributions and p-values) and their relation to bootstrapping results.
8. Compare nested models, select relevant variables, and explain how to evaluate the goodness of fit of the model. Distinguish different methods depending on the goal of the analysis: generative *versus* predictive models.
9. Measure multicollinearity, normality of residuals, (and heteroskedasticity), and identify whether and how the LS estimator becomes compromised in these settings.
10. Explain the role MLE plays in regression, and make extensions to model binary outcomes with logistic regression.

Teaching Team

Position	Name	Slack Handle	GHE Handle	Section
Lecture/Lab Instructor	Katie Burak	`@Katie`	`@katieburak`	1
Lecture/Lab Instructor	Payman Nickchi	`@Payman Nickchi`		2
Teaching Assistant	Ramin Rezaeianzadeh	`@Ramin Rezaeianzadeh (TA)`		1
Teaching Assistant	Anne-Sophie Fratzscher	`@Anne-Sophie (TA)`	`@afratz`	1
Teaching Assistant	Kaiyun Guo	`@Kaiyun Guo (TA)`		1
Teaching Assistant	Skylar Fang	`@Skylar Fang (TA)`		1
Teaching Assistant	Haley Oleynik	`@Haley Oleynik (TA)`	`@holeynik`	2
Teaching Assistant	Sidney Saint	`@Sidney (TA)`		2
Teaching Assistant	Samir Damji	`@Samir Damji`		2

Lectures

This course occurs during **Block 3** in the 2024/25 school year.

Lecture	Topic	Reading Material
1	Introduction to Linear Regression and population quantities	• ISLR 2.1, 3.1 • LMWR 2.1-2.3 • Great post about estimation vs prediction
2	Linear Regression with one explanatory variable. Estimation of regression coefficients and prediction. Different type of explanatory variables	• See the above resources • ISLR 3.5 for comparisons with kNN • IDS 8, 9 • This useful blog post
3	Linear Regression with one explanatory variable: inference and prediction. The sampling distribution of the coefficients′ estimators	• ISLR 3.1.1, 3.1.2, 3.3.1 • R4DS 23.1-23.3 (focused on computation but not rigorous conceptually) • OIS 7.1, 7.2, 8.1.1, 8.1.2 (good to build intuition but not too rigorous conceptually)
4	Linear model with many explanatory variables. Estimation, inference, and prediction	• Some of the above resources cover MLR • ISLR 3.2.1, 3.3.1
5	Comparison of nested models and goodness of fit	• ISLR 3.1.3, 3.2.2
6	Generative *versus* predictive models. Variable selection. Correlation *versus* causation problems	• ISLR 6.1

Lecture	Topic	Reading Material
7	The integrity of the model: check of assumptions, multicollinearity (and heteroskedasticity)	- ISLR 3.3.3
8	Maximum likelihood estimation. Logistic regression	- Data Analytics with R Chapter 7

Deliverables

This is an **assignment-based course**. The following deliverables will determine your course grade:

Assessment	Weight
Lab Assignment 1	12%
Lab Assignment 2	12%
Lab Assignment 3	12%
Lab Assignment 4	12%
iClicker	2%
Quiz 1	25%
Quiz 2	25%

Class Schedule & office hours

See calendar.

Reference Material

1. Part of the content follows closely Chapter 17 of Introduction to Data Science, Rafael Irizarry (book available at leanpub).
2. Introductory Econometrics: A Modern Approach, by Jeffrey Wooldridge (chapter 1 and 2.1, recommended).
3. Intro to Statistical Learning (ISLR), especially Chapter 3.
   - A modern and approachable take on statistics / machine learning.
4. Introduction to Data Science (IDS)
   - In particular Chapters 8 and 9
5. R for Data Science (R4DS), especially Part IV.
   - Practical and approachable book on the use of R for data science.
6. Linear Models with R (LMWR)
   - Comprehensive book on linear models.
7. OpenIntro Statistics (OIS)
   - Fairly accessible, seems to lean towards a traditional approach. Chapters 7 & 8 are relevant for linear regression.
8. Introduction to Data Science: Data Analysis and Prediction Algorithms with R (IDS)
9. Data Analytics with R
   - Chapter 7

Policies

Please see the general MDS policies.

Attribution

The course is built upon previous years' materials developed by previous instructors.

License