STAT 306: Finding Relationship in Data

The following is a guide to the course STAT 306: Finding Relationships in Data, as given in the first Winter term of the academic year 2020-21.
Aims: The course aims to provide learners with a toolkit for the understanding and application of a range of key methods in the field of linear models and multivariate data.
Objectives: On completing the course, students should be able to demonstrate an understanding of the techniques and applications of well–known ideas in linear modelling, including model fitting, model selection, model diagnostics, as well as basic ideas for generalised linear models and principal components analysis.
Learning outcomes: Detailed learning outcomes are provided on the course website, but broadly on successful completion of the course a learner will be able to:

1. Understand the principles of model fitting and inference for linear models involving a response variable with a single explanatory variable.

2. Understand the role of residuals in linear regression, including model diagnostics.

3. Appreciate and apply key concepts of linear modelling when there is more than one explanatory variable.

4. Understand and apply linear model theory to cases where at least one explanatory variable is categorical.

5. Critique studies that involve regression methods, including identification of any flaws and limitations to inferences.

6. Apply commonly used methods for model selection in a multiple regression context.

7. Use and interpret common approaches to identifying influential data points and outliers in a regression context.

8. Apply and interpret linear models that involve the transformation of one or more variable.

9. Apply and interpret a principal components analysis (PCA).

10. Understand and apply concepts from generalised linear modelling, including logistic and Poisson regression.

11. Apply linear modelling methods in the software R, using appropriate R functions, and interpreting the output.

Pre-requisites: One of MATH 152, MATH 221, MATH 223 and one of STAT 200, STAT 241, STAT 251, STAT 300, BIOL 300, COMM 291, ECON 325,

ECON 327, FRST 231, PSYC 218, PSYC 278, PSYC 366 and one of MATH 302, STAT 302.
Lecturer: Prof. B. Dunham (room ESB 3118, email: `B.Dunham@stat.ubc.ca`).
Lecture times: Tuesdays and Thursdays, 8.40am (with a mirrored session at 5pm).
Assessment: By the completion of the labs activities (10%), a sixty–minute midterm test (20%, at 8.20am on **29th October**), responses to clickers questions (5%), pre-classes quizzes (5%), online WeBWorK homeworks (10%), a $2\frac{1}{2}$–hour unseen examination (30%), two assignments (10%, each worth 5%) and a group project (10%). We will use iClicker Cloud (see `lthub.ubc.ca/guides/iclicker-cloud-student-guide/` for information). Note there are two courses set up in iClicker Cloud, one for each of the morning and evening sessions. You may attend whichever you wish. On creating your iClicker account please ensure your student ID identifies you (either your name or, should you have concerns about privacy, the first five digits of your student number).
It is necessary to pass the final examination to pass the course overall. The usual university rules for extenuating circumstances and plagiarism apply. Dates for the setting and completion of the assignments are indicated below:

	Set	Hand-in
Assignment 1	25th Sept.	16th Oct.
Assignment 2	9th Nov.	27th Nov.

For the on–line WeBWorK homeworks, the dates when each homework opens and closes are as follows:

	Opens	Closes
HW1	11th Sep.	20th Sep.
HW2	18th Sep.	27th Sep.
HW3	25th Sept.	4th Oct.
HW4	2nd Oct.	11th Oct.
HW5	9th Oct.	18th Oct.
HW6	16th Oct.	25th Oct.
HW7	30th Oct.	8th Nov.
HW8	6th Nov.	15th Nov.
HW9	13th Nov.	22nd Nov.
HW10	20th Nov.	29th Nov.
HW11	24th Nov.	4th Dec.

The homework sets can be accessed via the Canvas page under "Assignments". All questions set are of multiple choice or "fill in the blanks" format. All deadlines fall on Sunday evenings apart from HW11. Specific details regarding assessment regulations for the course can be found on the course web page.
There will be a group project in which students will work in pre-assigned groups on a data set of their selection. Further details will be available by week 3. The final project is a report submitted during the last week of term. There is an interim stage proposal for review, however, during week 10. Due dates for both components are below:

	Proposal	Report
Group project	13th Nov.	2nd Dec.

Teaching methods: Classes of approximately forty minutes duration will occur twice a week, with an online pencast describing related materials being available from the course web page in advance. A pre-class activity is set before each class and an accompanying quiz due on the morning of class days. In all sessions an in–class activity will replace at least part of the lecture component. A calculator or (preferably) R will be necessary for many of the in–class activities. Guided reading or other activities will be set at the end of one lecture to be completed prior to the next.
There will be required lab assignments most weeks. On the following weeks a lab activity will count toward the final grade:

	Commencing
1. Week 2	14th Sep.
2. Week 3	21st Sep.
3. Week 4	28th Sep.
4. Week 5	5th Oct.
5. Week 7	19th Oct.
6. Week 9	2nd Nov.
7. Week 11	16th Nov.
8. Week 12	23rd Nov.

There will also be office hours each week, commencing in week 2. Students can register for the Piazza forum via the link on the Canvas page.
Programme of work: The study time should total around eight hours per

week. So in addition to the contact hours, it is essential that learners spend approximately six hours per week on self–study for the course. A proposed workload for a typical week is as follows:
Classes (including pre-class activity, pencasts, quiz, class): 3 hours
WeBWorK: 2 hours
Lab: 1 hour
Reading/reviewing: 1 hour
Other/assignments: 1 hour
Feedback: After all assignments have been submitted and marked, individual feedback will be provided in the form of brief notes on marked work. Detailed written comments will also be provided on the course web–page where appropriate.
Recommended texts: There are a variety of books that cover at most of the material in this course, and it is suggested you try the UBC online library stock to find those that suit you. The course notes are
Joe, H. (2020): *Course Notes for STAT 306: Finding Relationships in Data.* which can be ordered from the UBC bookstore. Amongst other useful texts, both available via the library website, are
Chatterjee, S. and Hadi, A.S. (2006): *Regression Analysis by Example*, (4th edit.). Wiley (In particular chapters 1–6, 11, 12, 13.3 are covered.)
Weisberg, S. (2014): *Applied Linear Regression*, (3rd edit.). Wiley.

Further information will appear on the course web page.
Sensitive content: During this pandemic, the shift to online learning has greatly altered teaching and studying at UBC, including changes to health and safety considerations. Keep in mind that some UBC courses might cover topics that are censored or considered illegal by non-Canadian governments. This may include, but is not limited to, human rights, representative government, defamation, obscenity, gender or sexuality, and historical or current geopolitical controversies. If you are a student living abroad, you will be subject to the laws of your local jurisdiction, and your local authorities might limit your access to course material or take punitive action against you. UBC is strongly committed to academic freedom, but has no control over foreign authorities (please visit `http://www.calendar.ubc.ca/vancouver/index.cfm?tree=3,33,86,0` for an articulation of the values of the University conveyed in the Senate Statement on Academic Freedom). Thus, we recognize that students will have legitimate reason to exercise caution in studying certain subjects. If you have concerns regarding your personal situation, consider postponing taking a course with manifest risks, until you are back on campus or reach

out to your academic advisor to find substitute courses. For further information and support, please visit:
`http://academic.ubc.ca/support-resources/freedom-expression`

There follows a provisional guide to the lecture slots available. It is possible that the material covered in the classes will differ slightly from the description below.

1. Introduction and motivation. Exploring relationships between two variables.
2. Least squares estimation for the simple linear model.
3. Residuals. Properties of the model.
4. Confidence intervals for the slope and an expected response.
5. Prediction intervals.
6. Distribution theory; why the t distribution?
7. Matrix formulation of linear models.
8. Properties of least squares estimators in matrix form.
9. Properties of residuals and the Residual SS.
10. Dummy variables in linear models.
11. More on categorical variables in linear models.
12. Quadratic models and curve fitting.
13. Examining case studies.
14. Review Activity.
15. Mid–term test. (29th Oct.)
16. Model selection, including Mallows' C_p statistic.
17. Leverage, influence, outliers, and the "hat" matrix.
18. Transformations.

19. A case study.
20. Introducing logistic regression.
21. Further logistic regression.
22. Model selection in logistic regression. PCA.
23. Introducing Poisson regression.
24. Further Poisson regression.
25. Review session.

BD