STAT 306: Finding Relationship in Data

The following is a guide to the course STAT 306: Finding Relationships in Data, as given in the second summer term of the academic year 2024-25. 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, 6pm, in CEME Floor 1 Room 1202.
Assessment: By the completion of the labs activities (10%), a midterm test (20%, at 6pm on **29th July**), 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). Please ensure your student ID identifies you (either your name as it appears in Canvas or, should you have concerns about privacy, the first five digits of your student number).
The usual university rules for extenuating circumstances, academic misconduct, and plagiarism apply. Dates for the setting and completion of the assignments are indicated below:

	Set	Hand-in
Assignment 1	11th Jul.	23rd Jul.
Assignment 2	30th Jul.	8th Aug.

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

	Opens	Closes
HW1	4th Jul.	13th Jul.
HW2	11th Jul.	20th Jul.
HW3	18th Jul.	27th Jul.
HW4	25th Jul.	3rd Aug.
HW5	1st Aug.	10th Aug.

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. 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 2. The final project is a report submitted during the last week of term.

There is an interim stage proposal for review, however, during week 5. Due dates for both components are below:

	Proposal	Report
Group project	1st Aug.	11th Aug.

Teaching methods: Classes of approximately three hours duration will occur twice a week, with online pencasts describing related materials being available from the course web page in advance. Pre-class activities are set before each class and an accompanying quiz due by 4pm on 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 sessions every week. On the following days a lab activity will count toward the final grade:

	Date
1. Week 2	8th Jul.
2. Week 2	10th Jul.
3. Week 3	15th Jul.
4. Week 3	17th Jul.
5. Week 4	22nd Jul.
6. Week 4	24th Jul.
7. Week 5	31st Jul.
8. Week 6	5th Aug.
9. Week 6	7th Aug.

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.
Extreme environmental conditions: In the event that environmental conditions mean that in–person classes are impossible, it is expected that lectures and, if practical, labs will be offered synchronously via Zoom. Please see the Canvas page for announcements.
Programme of work: The study time should total around sixteen hours per week. So in addition to the contact hours, it is essential that learners spend approximately ten 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): 6 hours

WeBWorK: 4 hours
Lab: 2 hours
Reading/reviewing/exercises: 2 hours
Project/assignments: 2 hours
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, 9.1-9.7, 12.1-12.7, 13.3 are covered.)
Pardoe, I. (2020): *Applied Regression Modeling*, (3rd edit.). Wiley. (2nd edition (2012) also helpful.)

Further information will appear on the course web page.
Sensitive content: 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.

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. Least squares estimation for the simple linear model.

2. Residuals. Properties of the model. Confidence intervals for the slope and an expected response.

3. Prediction intervals. Distribution theory; why the t distribution?

4. Matrix formulation of linear models. Properties of least squares estimators in matrix form.

5. Properties of residuals and the Residual SS. Dummy variables in linear models.

6. More on categorical variables in linear models. Quadratic models and curve fitting.

7. Examining case studies. Review Activity.

8. Mid–term test. (29th July) Model selection, including Mallows' C_p statistic.

9. Leverage, influence, outliers, and the "hat" matrix. Transformations.

10. Continuous interactions. A case study. Introducing logistic regression.

11. Further logistic regression. PCA. Model selection in logistic regression. (Introducing Poisson regression. PCR.)

BD