STAC67: Regression Analysis

Lecture 1

Thibault Randrianarisoa

September 9, 2026

University of Toronto Scarborough

Overview of today

- **Part 1 — Organisation of the course**
    - What this course is about, and what it will look like
    - Logistics, assessment, calendar
- **Part 2 — Introduction to regression**
    - What regression is, and where it came from
    - Covariance and correlation

Part 1 — Organisation of the course

Course information

Lectures	Wednesday 16:00–17:00, IA 2021
	Friday 13:00–15:00, IA 2021
Tutorials (from week 2)	TUT0001 — Tuesday 17:00–18:00, IA 3120
	TUT0002 — Wednesday 15:00–16:00, IC 208
Office hours	Wednesday 10:00–11:00 and Friday 15:00–16:00, IA 4064
Teaching assistants	announced on the course page

Tutorials begin next week, not this week.

Course logistics

Course website: the main source of information. Slides, annotated slides, tutorials and deadlines are posted there. Check it regularly.

Quercus: announcements and grades.

Piazza: for discussion. Ask about course content, assignments and the project here rather than by email.

- Your grade **does not depend** on your participation on Piazza.
- It is simply the fastest way to get an answer, from me, the TAs or your peers.

Email: reserved for private matters. Put `[STAC67]` in the subject and your student number in the body.

Literature

The course follows the chapter structure of the required text.

- **(Kutner)** Kutner, Nachtsheim & Neter, *Applied Linear Regression Models*, 4th edition. Older editions are fine.
- **(Sheather)** S. J. Sheather, *A Modern Approach to Regression with R*. A lighter, R-centred companion, free through the UofT library.

The suggested sections for each week are listed on the course page. Read them before the Friday lecture if you can.

What this course will look like

- **Wednesday (1h)** mostly new theory, developed at the board.
- **Friday (2h)** theory continued, then worked examples in R.
- The slides are deliberately sparse. I write on them during the lecture, and the **annotated version is posted after each class**.
- **Tutorials** are course reviews + some exercises, and the three quizzes are written there. Tutorials take place before that week's lectures, so a tutorial always revisits the previous week rather than previewing the coming one.
- Everything is computed in **R**. You will be expected to write R code and interpret R output on assignments, quizzes and tests.

Assessment

Evaluation	Weight	Details
Assignments	15%	Three assignments, 5% each. No late submissions.
Quizzes	15%	Three quizzes, written in tutorial. **Best two of three** count. No make-ups.
Midterm	15%	Scheduled by the Registrar, week of 19 October. Covers through week 6.
Case study project	20% or 0%	**Optional.** Groups of at most four. Checkpoint, report, presentation.
Final exam	35% or 55%	December examination period. Covers the whole term.

The project is optional, and it can only help you; see the next slide.

The case study project

Your grade is computed **both ways**, and you receive whichever is higher:

$$15\%\,A \;+\; 15\%\,Q \;+\; 15\%\,M \;+\; \max\big(\,20\%\,P + 35\%\,F,\;\; 55\%\,F\,\big)$$

- In words: your project replaces 20 points of your final exam's weight whenever that helps you, which happens exactly when your project mark beats your final exam mark.
- If you do not do the project, your final exam is worth 55%.
- The midterm is **not** optional and is not part of this calculation.

Brief and datasets are released on 9 October; groups and datasets are claimed by 6 November.

Use of generative AI

- You **may** use AI tools as learning aids: to gather information across sources, to help you understand material, provide new exercises/examples,...
- You **may not** use them for taking tests, or for completing course assignments.
- You are ultimately accountable for everything you submit. If you cannot explain a line of your report or your code, it does not belong there.

The full wording is in the syllabus. Ask me if a particular use is unclear.

Provisional calendar

Weeks	Topic
1–3	Simple linear regression: least squares, inference, ANOVA and R^2
4–5	Diagnostics I; the matrix formulation; properties of the estimators
6–7	Geometry of least squares; F-tests; multicollinearity; qualitative predictors
	Reading week: 26–30 October
8–9	Interactions and ANCOVA; polynomial regression; transformations
10–11	Model selection and validation; diagnostics II; weighted least squares
12	Ridge regression and the LASSO; project presentations

Tentative, the up-to-date version is always on the course page.

Part 2 — Introduction to regression

Regression: What is it?

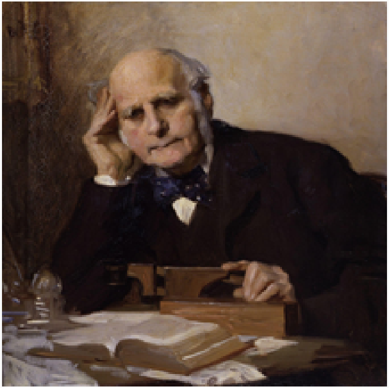
- Simply: The most widely used statistical tool for understanding relationships among variables.
- A conceptually simple method for investigating relationships between one or more factors and an outcome of interest.
- The relationship is expressed in the form of an equation or a model connecting the outcome to the factors.

Examples of Application

- **Epidemiology**: what are the social factors to contribute the death rate of Covid-19 in Canada?
- **Business**: determing price and marketing strategy:
  - Estimate the effect of price and advertisement on sales
  - Decide what is optimal price and campaign?
- Straight prediction questions:
  - What will be the interest rates be next month?
  - What will be the length of hospital stay of a surgical patient?
- Explanation and understanding
- Much more . . ..

Origin of Regression (History)

- In 1886, Sir Francis Galton (1822-1911) invented the term, **"regression"** after he analyzed the data on the heights of parents and their children.



{“Regression towards mediocrity in hereditary stature”. The Journal of the Anthropological Institute of Great Britain and Ireland, Vol 15, pages 246-263 (or Wikipedia!)}

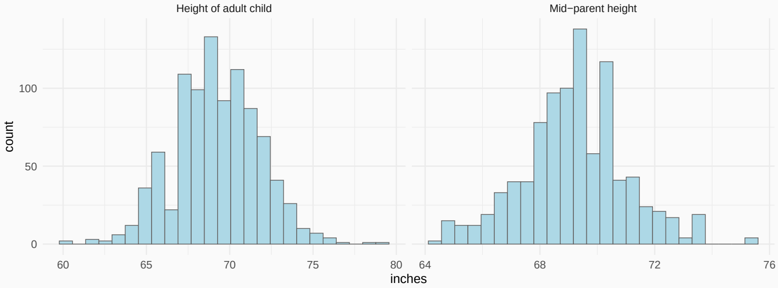
Galton's question

- Galton was a Victorian gentleman scientist (a cousin of Charles Darwin) who worked on everything from fingerprints to meteorology, but was best known in his lifetime for his work on inheritance/heredity.
- His question was concrete: how is height passed from one generation to the next?
- So he collected the heights of adult children and of their parents, and cross-tabulated them. The unremarkable conclusion came first: tall parents tend to have tall children.

The remarkable conclusion came only when he looked more carefully.

The GaltonFamilies data

- The `GaltonFamilies` dataset (package `HistData`) records the 934 adult children of 205 sets of parents (the modern form of the data Galton worked with).
- Galton put daughters on the same scale as sons by multiplying their height by 1.08. `midparentHeight` is already on that scale, so we do the same to the children.



Regression to the mean

Average the child heights within each parent band:

Parents	n	Mean parent	Mean child
72 and over	79	72.7	71.9
70-71	322	70.3	70.0
68-69	385	68.6	68.7
66-67	119	66.7	67.8
65 and under	29	65.0	66.2

- Tall parents do have tall children, but not as tall as themselves. Short parents have short children, but not as short as themselves.
- Every group is pulled towards the overall mean of 69.2 inches.

Galton called this "regression towards mediocrity". The method we are about to study inherited the name, but the name describes his finding, not the method.

From an idea to a method

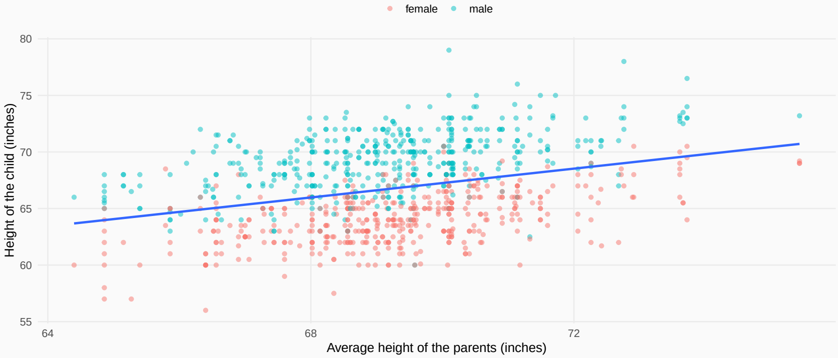
Galton had two ideas but not the mathematics for either:

- a function describing how one quantitative variable relates to another;
- a measure of how tightly that relationship holds, what he called the "closeness of co-relation".

These became **regression** and **correlation**. The formulas were supplied over the following years by **Edgeworth**, **Pearson** and **Yule**, leaning on results **Gauss** had developed almost a century earlier for a quite different problem from astronomy.

Least squares reaches us from astronomy; the name "regression" from heredity.

Data Visualization



Review: Covariance and Correlation Coefficient

Suppose we have observations on n subjects consisting of a **dependent** or **response variable** Y and an **independent** or **explanatory variable** X.

We want to measure both direction and strength of the relationship between Y and X.

Obs	Y	X
1	y_1	x_1
2	y_2	x_2
$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n

Covariance and Correlation

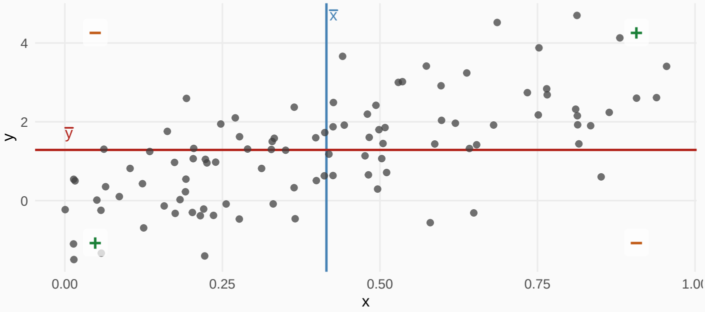
Covariance: $Cov(X,Y) = E((X-\mu_x)(Y-\mu_y))$, where
$\mu_x = E(X),\ \mu_y = E(Y)$

For $Z_x = \frac{X-\mu_x}{\sqrt{Var(X)}},\ \ Z_y = \frac{Y-\mu_y}{\sqrt{Var(Y)}}$, the correlation coefficient is given by

$$Cov(Z_x, Z_y) = \rho_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

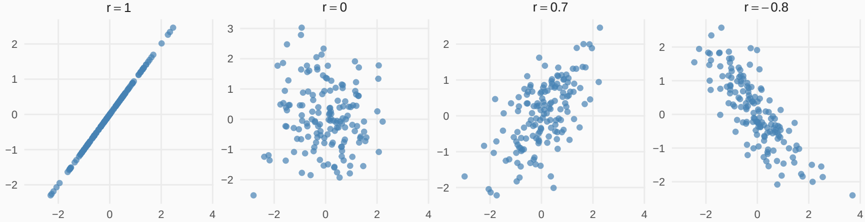
- $-1 \leq \rho_{xy} \leq 1$
- When the relationship is perfectly linear then $|\rho| = 1$.
- if two variables are independent then $\rho = 0$. (Note: the inverse does not hold)

Sample Covariance and Correlation



$$Cor(Y,X) = Cov(Z_y, Z_x) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{s_y}\right)\left(\frac{x_i - \bar{x}}{s_x}\right)$$

Correlation



Question: what are main differences between correlation and regression model?

A caution to carry through the course

A model never emerges from the data. We impose it.

- The regression model, and every assumption that comes with it, is a choice we make, not something we take from the data.
- A model is a representation: it displays the structure of something more complicated than itself, and it is wrong in ways we should be able to name.
- So we must be ready to justify the choice on **two** grounds: theoretical (does this form make sense here?) and empirical (do the diagnostics support it?).

We will spend weeks 4, 10 and 11 learning how to check the second one.