# Stat 305A: linear models and more

## Overview

This course is about the linear model. It is also a course about applied statistics, using the linear model to illustrate important concepts. The structure is as follows: we work through linear models in increasing order of complexity, pausing to talk about statistical ideas along the way to understand how and when to use them.

In regression we're working primarily with real-valued responses. The main tool for regression is the linear model, in all its glory ranging from the humble one sample $t$ test to more elaborate methods like splines and wavelets. We also look at competing methods that are sometimes better than linear regression, because the focus is on the problems not the tools. The mathematics and computation involved in regression are comparatively simple. Applied statistics remains difficult because connecting methods appropriately to a given problem context is hard. That is the focus of this course.

## Goals

**1)** Learn about the linear model $Y = X\beta + \varepsilon$ in depth and detail.
**2)** Use applied statistics tools: bootstrap, cross-validation, permutations, and more.
**3)** Cross-cutting concepts: reproducibility, random effects, sparsity, Bayes, Causal inference and more

## Prerequisites

This is not a first course in linear models. It is designed to be one's **last** course on linear models and is aimed at first year statistics PhD students. Of course it is useful in other programs. Students enter with

**1)** Hands on experience handling data.
**2)** Strong preparation in probability, linear algebra, analysis.
**3)** Lots of statistics theory.
**4)** Experience computing with data. We will be using R.

Hard work can make up for a lack of one of the items above. For two or more gaps, you'll feel lost, and would get more out of some other course. Then taking this course a year later would be wiser.

## Access and accommodations

Here is a link to Stanford's office of accessible education:

`https://oae.stanford.edu/faculty-staff/syllabus-statement`

If you have an accommodation letter, send it to me as early as possible. Ideally within the first week or two. It can be difficult to find a room for an accommodation.

## Texts and materials

Will be posted in canvas.

## Grading

Homework 25%, midterm 30%, final 45%. Tests are closed to books and notes.
Do not book holiday travel that conflicts with the final! That may delay your grade by one year.

## More information

Here is a prior non-canvas page from an earlier year:
`https://artowen.su.domains/courses/305a/`

## Office hours

Art Owen Wed 11am–12pm, Sequoia 130
Jamie Liu, Wed 1:30pm–3:30pm, Sequoia 105
Jack Krew, Mon 3:00pm–5:00pm, CoDa B40

# Stat 305A $\widehat{\text{syllabus}}$

The hat over the syllabus indicates that there might be small alterations (student suggestions, news events, etc.). In the table below "Min" refers to a chapter in the scribed notes by Eric Min that I'm updating this quarter.

| Date | Min | Linear Model | Applied Statistics |
|---|---|---|---|
| Sep 22 | 1 | Linear model examples, notation | overview, generalization, connecting data to the world, Box's quote(s), signal and noise inferential tasks, sampling models for $x$ |
| Sep 24 | 2,3 | Random vectors, random matrices, moments, quadratic forms, Normal and related distributions, basics of Bayes | |
| Sep 29 | 4 | Linear least squares: algebra, geometry, calculus and computation. Distribution of least squares estimates. | Gauss-Markov theorem. Extra sum of squares. SVD. $t$-tests. $F$-tests |
| Oct 01 | 5,6.1–6.3 | One sample as example of one parameter: mean, variance, $t$-test, power, sample size | Staircase of inference, one-tail tests (avoid), practical vs statistical inference, confidence intervals (theory **and** practice) |
| Oct 06 | 7,8 | Two samples. Balance, begin ANOVA | Permutation tests, randomization, sign test, Wilcoxon |
| Oct 08 | 8 | More ANOVA, contrasts | Fixed and random effects (generalization again), multiple testing |
| Oct 13 | 9 | Plain linear regression. Pointwise and simultaneous confidence bands. Prediction bands. Regression through the origin (avoid). | Regression fallacy. Sufficient variation. Errors in variables. Calibration. Variance explained. |
| Oct 15 | 12,13 | Multiple regression | Interpretation, competing variables, correlation vs causation, partial correlation, Simpson, Cornfield |
| Oct 20 | 14 | Ridge, lasso, principal components regression | correlated predictors, regularization, cross-validation, GCV |
| Oct 22 | 16 | Non-Gaussianity, Omitted variables, Added variable plots, Generalized least squares. | Model error, Bias, Lack of fit and pure error, Nonlinearity, Correlated errors |
| Oct 27 | – | Variable selection, AIC, BIC, Robust regression, (catchup/review) | model choice, outliers |
| Oct 29 | | **Midterm** (topics up to Oct 27) | Midterm |
| Nov 03 | | **Democracy day** | **No class** |
| Nov 05 | 17 | Bootstraps. Sandwich. Newey-West. | Resampling, robust standard errors |
| Nov 10 | – | Smoothing and kriging | Nonparametric regression, kernels |
| Nov 12 | – | Regression discontinuity & instrumental variables | Causal inference |
| Nov 17 | – | Bayes for the linear model I | Basics of Bayesian inference |
| Nov 19 | – | Bayes for the linear model II | Prior choice, hierarchical models, spike and slab, random effects |
| Nov 24 | | **Thanksgiving recess** | **No class** |
| Nov 26 | | **Thanksgiving recess** | **No class** |
| Dec 01 | – | Quantile regression and conformal inference | Choosing a loss function, nonparametric prediction intervals |
| Dec 03 | – | Extras and a summary | TBD |
| Dec 10 | | **Final exam** | 3:30–6:30pm |