CHAPTER 1

Introduction

Stat 305 is the first course in our applied statistics sequence. It focusses on regression problems, especially the linear model. We will get a deep understanding of linear regression and, in learning the limits of linear regression, get an introduction to some related topics.

Predictive Modelling

A lot of statistical methods can be organized around predictive modelling. There is some random variable that we wish to understand. This is called the response variable and is usually denoted by Y. The distribution of Y is related to one or more predictor variables, collectively called X. The more we understand about how X affects the distribution of Y the better off we are. The data set supplies us with a bunch of (X, Y) pairs and we want to say something about how the distribution of Y changes when $X = x$. The approach we take depends mostly on the type of variable that Y is (real or binary or categorical or some other) as well as the types for the components of X. It also depends on how exactly the (X, Y) pairs were obtained. Perhaps they were just observed, with independent identically distributed (IID) sampling being the simplest possibility, or perhaps they were taken by one of several different kinds of experimental design. Of course the goals that lead us or somebody else to gather the data in the first place affect the method as does subject matter knowledge relevant to the data at hand and previous similar data sets.

Of the broad factors described above the primary one is the type of response variable. Methods for the same response type usually have a lot in common with each other. We ordinarily use the same regression analysis to predict real valued response variables from IID sampling as we would in a designed experiment, and even the confidence interval calculations could be the same. Whether the

X values are real, or categorical or vector values changes how we encode them in our model, but after they're encoded it looks the same. But when the response changes from real to binary or categorical, some bigger changes are required.

The artist Paul Gauguin famously asks three big questions: "Where do we come from? What are we? Where are we going?". Applied statisticians learning about a new data set can start by asking: "What is Y? What X's are there? How were the (X, Y) pairs sampled?". The answers to these questions set the stage for deeper considerations about the goals of the investigation and what might already be known about the problem.

These notes are about the setting with $Y \in \mathbb{R}$ under all the usual assumptions on the predictors X and some of the most common assumptions about how the data are gathered. The central thread is about prediction and modelling but other issues come up too.

Statistics, mathematics and computation

To do applied statistics well, we need to use some mathematics. For this course the mathematical requirements are linear algebra, calculus, and some basic probability. We also need to be able to compute. For this course we'll be using the statistical language R.

It is important to keep in mind that statistics itself is not mathematics. Nor is it computing. Statistics is about learning from data. Mathematical tools let us understand how methods will work under idealized conditions. When the mathematics is too hard to do then computational methods, particularly simulations, can be brought to bear. Often we do both, and it's not always clear which is the answer and which is done as a check. The main task in statistics is in figuring out which mathematical assumptions or computational approaches best fit the problem. Actually applying those methods is now quite easy and getting easier as software improves.

Assumptions and models

The nastiest issue in applied statistics is the role of the assumptions we make about the data. Data analysis would be very straightforward if we could make a set of assumptions, verify that they hold true, derive the consequences for our data, and then apply and interpret those consequences. Unfortunately we're almost never sure that the assumptions are correct. Indeed we usually know that they're not correct. A small set of popular distributions get used over and over in statistics, but there's no a priori reason to expect a new set of data to follow one of our favorite distributions. We're left trying to get right answers from wrong assumptions.

George Box is widely quoted as saying "All models are wrong, but some are useful." The quote appears in varying forms. I think that Box is right, but that still leaves the hard job of judging which of the many wrong models at our disposal might be useful for a given problem. Describing how to make those choices, for real valued responses, is the main subject of these notes.

A good assumption is not necessarily one that we most readily believe. We will look at two sample problems in Chapter 3.2. The basic model there is that $Y \sim N(\mu_1, \sigma^2)$ in the one group and that $Y \sim N(\mu_2, \sigma^2)$ in the other. It is quite far fetched to suppose that the observations have exactly the normal distribution. It seems plausible to me that no normally distributed random variable has ever been observed. Yet somehow medicine and agriculture and many other sciences have made tremendous strides over the past century or so, using in many cases precisely that condition. It helps that answers which hold exactly for exactly normal data often hold approximately for approximately normal data. The data don't always have to even be nearly normally distributed because the central limit theorem will make some of our test statistics nearly normally distributed for moderately large sample sizes, even when the data themselves are not very close to normally distributed.

Hidden in that two sample setting above is another assumption, that σ^2 is the same in both groups. This is quite plausible, in fact the two groups might possibly be taken from just one population with the group variable unrelated to the values that Y takes. But if the two groups don't in fact have the same variance then the results of our modelling may be unreliable even when the sample sizes in the two groups are very large.

This example shows that the importance of an assumption depends not just on how likely it is to be correct but also on how much our results change when that assumption fails. Here the plausible assumption (equal variance) has to be watched carefully while the implausible one (normality) causes less trouble.

A recurring theme in these notes is that we judge the safety of a model assumption by embedding it in a bigger model to see how sensitive the results are. For the two sample problem we can suppose that $Y \sim N(\mu_1, \sigma_1^2)$ in one group and that $Y \sim N(\mu_2, \sigma_2^2)$ in the other and see what happens to a method derived assuming $\sigma_1 = \sigma_2$ when in fact $\sigma_1 \neq \sigma_2$. Or we can use a more general model with $Y \sim F_1$ in one sample and $Y \sim F_2$ in the other where F_1 and F_2 are two different distributions and see what happens for a method derived for normally distributed data when F_1 and F_2 are not normal.

Linear model

The central statistical technique for handling real valued responses is the linear regression model. The given data take the form of (x_i, y_i) pairs for $i = 1, \ldots, n$. From each $x_i \in \mathbb{R}^d$ we construct a vector of features $z_i \in \mathbb{R}^p$ and suppose that $y_i \doteq z_i'\beta$ for a coefficient vector β. We'll look at several ways to make the meaning of "$\doteq$" precise later.

As an example linear model, suppose that $x_i \in \mathbb{R}$ and that $z_i = (1, x_i, x_i^2)'$. Then the linear model has y_i approximated by a quadratic function $\beta_1 + \beta_2 x_i + \beta_3 x_i^2$ for $i = 1, \ldots, n$. If it seems odd that our first linear model is in fact quadratic in x, then that should serve to remind us what 'linear' means in the linear model. It means that the predictions are linear in the unknown coefficient vector $\beta \in \mathbb{R}^p$. They're also linear in the feature vector z_i. But the linear model is not necessarily linear in the predictors x.

Matrix and vector notation is very useful in linear models. First we pack the n observations up into

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{pmatrix}, \quad \text{and} \quad Z = \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1p} \\ z_{21} & z_{22} & \dots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \dots & z_{np} \end{pmatrix}.$$

Then we gather the coefficients into $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$. Now the linear model may be compactly written

$$Y \doteq Z\beta.$$

The matrix Z is called the design matrix. The term is most appropriate when x_i are taken from a designed experiment but it is used more generally. The matrix X can be thought of as a "run sheet". Row i of X has the settings for all d variables used in getting y_i. Column j has all the settings for the j'th experimental variable.

The linear model is often written in terms of $X\beta$ not $Z\beta$. That gets awkward because the design matrix is usually different from the run sheet. The connection between X and Z is closest for multiple regression (Chapter 5), except that we usually want the first column of Z to consist completely of 1s to put an intercept term in the model.

Sampling models

Fixed X random Y. The simplest and most widely studied sampling model for linear regression has the following form. The values of $x_1, \dots, x_n \in \mathbb{R}^d$ are fixed nonrandom values. So are $z_i \in \mathbb{R}^p$. We'll suppose that to make features out of x_i we invoke some function on x_i. We call this function ϕ with the mnemonic "phi for features". Then $z_i = \phi(x_i)$, with $\phi(x) = (1, x, x^2)'$ in the example above. Next we suppose that Y_i are independent random variables

$$Y_i = z_i'\beta + \varepsilon_i \tag{1.1}$$

where $\beta \in \mathbb{R}^p$ is an unknown parameter vector and ε_i are independent random variables with mean 0, and usually some more conditions. The actual y_i values we get are realizations of these random variables. We write $y_i = z_i'\beta + \varepsilon_i$. We don't have notation to separate the random variable ε_i from an observed value. We don't need it because we will not observe the ε_i.

Random X random Y. In a lot of applications nobody really fixed the x's and then observed the Ys. Very often (X, Y) pairs are observed together and X is just as random as Y. It is typical to use the 'fixed x random Y' models even in this setting. The reason that treating x as fixed can make sense is that we can do our data analysis conditionally on $X_i = x_i$ for $i = 1, \dots, n$. Once we have conditioned on them, they're fixed.

Conditioning arguments can get slippery. Here is an intuitive explanation. We can envision an (X, Y) pair as being generated in two steps. At the first step x is sampled from its marginal distribution F_X. Then y is drawn from the conditional distribution $F_{Y|x}$ of Y given that $X = x$. Maybe F_X depends on a parameter θ and $F_{Y|x}$ depends on β and σ. After the first step we should know something about θ. But in modelling, we mostly care about β and possibly σ. Unless there is some known connection from θ to β or σ then we don't learn anything about them until the second step happens. Between the two steps X is fixed, so a fixed analysis is appropriate, or at least not terribly wrong. In Chapter xxx we will look at an argument, going by the name of ancillarity, that says that the conditional analysis is the more appropriate choice even when we can do either the conditional or the unconditional analysis.

The variable X does not have to actually be observed ahead of Y in time for this argument to apply. Any joint distribution $F_{X,Y}$ can be represented as the marginal distribution of X and the conditional distribution of Y given X.

While it is usually simpler to treat X as fixed by conditioning, there are settings where it is better to treat the random (X, Y) pairs as samples from a joint distribution. For example we may want to use a cross-validation technique with some (X_i, Y_i) left out of the fitting. Cross-validation is easier to understand for random X's than for fixed ones.

Random X fixed Y. We seldom see regression problems with Y fixed and then X randomly sampled. In calibration problems (Chapter xxx) we look at predicting a variable that was fixed in our data set from one that was random. But we handle it by predicting X in a fixed X random Y setting.

Random X with fixed Y is very common in discrete data settings where for example one might compare features X for people with a disease $Y = 1$ ("cases") to those without the disease $Y = 0$ ("controls"). There it is common to look a fixed number of cases and controls.

Edge cases

Not every problem fits the ordinary linear model $Y \doteq Z\beta$ discussed above with random Y and either fixed or random X. Though most of this course is on that setting we need to recognize exceptions. Many of the exceptions require specialized methods that are outside the scope of this text.

Imperfectly observed data. Sometimes there is a disconnect between the distribution from which our data were sampled and the setting in which we wish to apply the conclusions. These disconnects can happen in many different ways.

1. extrapolation
2. sampling bias, concept drift
3. censoring and truncation

4. missing data

Multiple error sources. The main model in regression has one error term ε_i. Sometimes, especially in designed experiments, we have more than one error term. For example iron bars may be cut on machine i by person j. The bars have a target length of τ. The actual length might be $\tau+\alpha_i+\beta_j+\varepsilon_{ij}$ where α_i is a random error governed by the way machine i works, β_j is a random error capturing habits of person j and ε_{ij} is an error term.

We look at some simple versions of these random effects models in Chapter xxx. In an industrial setting like the one mentioned here the three errors might have variances σ_A^2, σ_B^2, and σ_E^2. The key to understanding the quality of the product may lie in measuring these variances and reducing the largest of them.

Sometimes we can cram this problem into the single error framework. But that is not always wise. We will look at some random effects models for problems like this, in Chapter xxx.

Hierarchical sampling. In examples like the one above it may happen that each machine is located in a different building and that none of the people using them works on more than one of the machines. Then only certain i-j combinations ever get looked at. We say that the person variable is nested inside the machine variable.

Similarly in the analysis of educational data we might have students from a sample of classrooms, that were chosen from a sample of schools selected from a sample of school boards. An analysis that ignores the sampling structure can lead to serious errors.

Causality

A model like $Y=\beta_0+\beta_1 x+\varepsilon_i$ strongly suggests that we can change Y by changing X. Taken at face value it says that increasing x by one unit increases Y by β_1 units if ε_i does not change, or by β_1 units on the average, if the change in X is accompanied by a new sampling of ε.

The face value interpretation is not necessarily right. It is one thing to learn about the distribution of Y given X in a setting where X and Y were jointly observed and quite another in a setting where we're changing X. The best way to be sure is to design a study where the investigator intervenes to assign X values. The conditioning argument that lets us treat random X's as fixed X's does not help here either.

Further references

Sacks et al (19xx) give a rationale for giving prediction a central place in statistical modelling.