Lecture 1

Introduction

Disclaimer: *These lecture notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students of stat 305A at Stanford University. Copyright: A. B. Owen 2026.*

This is a course about the linear model. It is also a course about applied statistics and statistical thinking. We will cover the linear model and use it to illustrate applied statistics.

The linear model has

$$\mathbb{E}(Y) = X\beta, \quad \hat{\beta} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y, \quad \text{and} \quad \mathrm{var}(\hat{\beta}) = (X^\mathsf{T}X)^{-1}\sigma^2.$$

All the rest is commentary. Almost. In addition to a lot of commentary, there will be some other formulas. This should definitely not be your first course on the linear model. It is designed to be your last.

In this course, statistical thinking involves working through the consequences of the linear model theory we will learn and making connections to the motivating problems. Those problems usually originate in a domain like science, engineering, social science or business. A challenge for applied work is that there is not always a clean separation between statistical expertise and domain expertise. Important choices can require consideration of both statistical and domain knowledge.

This lecture is the most philosophical one in the course. The next one is almost purely about technical details. Typical lectures can touch both of those but are dominated by practical points.

1.1 A basic illustration

As an example of statistical thinking, consider the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $x_i \in \mathbb{R}$ with $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. We will study this model later. The notation should be familiar and will be discussed later. We will see that a smaller sample variance for x_i leads to a larger sample variance for $\hat{\beta}_1$. That has downstream consequences: a wide uncertainty interval for β_1 and possibly no evidence that $\beta_1 \neq 0$. If our data provide no evidence that $|\beta_1|$ is large enough to matter, then we will want to understand how that happened before concluding that x is unimportant. One of several possibilities to check is that the sample x_i did not vary much. We should see if that happened. The domain we are working in might actually require a narrow distribution for x. Box notes the possibility that x is known to be important and must be tightly controlled, so the data do not show that x is unimportant. Or, we may have observed a wide range of x values but our data selection might only gather y for x in a narrow range.

The above process is simply working the formulas backwards from our data to possible causes. It is not particularly difficult to do this, yet it could easily be overlooked. Another process is prospective. If we have some influence over which x values will be in the data set, then there is some statistical value in arranging for the x_i to vary widely. The cost to do this, or whether we can do it at all, both depend on the domain.

1.2 Generalization

Let's start with an extreme proposition: the reason we have a field of statistics is that we want to reason about data we don't have using data that we do have. For that to be possible, we need a logical connection between those two kinds of data. We do that through a model. The model consists of a collection of distributions on the data. From the known data we learn something about which distributions are best. We then use that knowledge to reason about the unknown data.

Figure 1.1 shows a diagram. On the left we see arrows pointing from the model to the two data sets. Our inference reverses the direction of the leftmost arrow providing a path from known to unknown data. We learn about the model by studying examples. This is inductive reasoning going from examples to a general statement. It is shakier than deductive reasoning but we have no choice. We will also use Ockham's razor, also called parsimony, which is about building in a preference for simpler models.

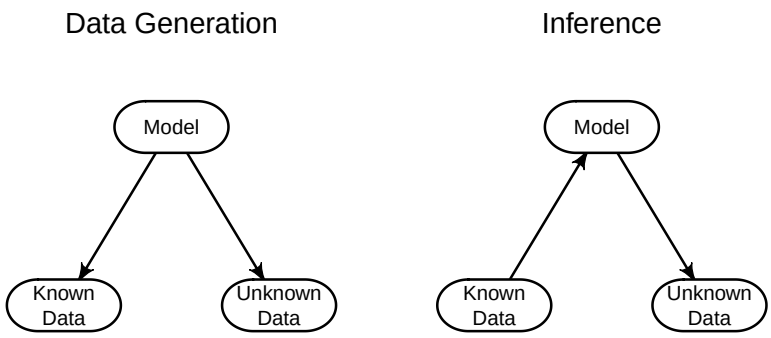


Figure 1.1: Our model generates both known and unknown data. In inference we use induction to learn about the model via the known data and then deduction to reason about the future data from the model.

The model is an assumption and it is generally false. As George Box has said: 'All models are wrong but some are useful'. So which models are useful, when are they useful, and for what purposes are they useful? It is on us to decide these things and people reasonably differ. The best experts can say diametrically opposed things. This is a sharp difference from the mathematics that we will draw on, where rigorous proofs settle important issues.

We can reasonably prefer some wrong models to others. If small errors in our model lead to small changes in our conclusions, then the model is robust. Other things being equal we prefer robustness. Robustness is a kind of continuity.

We don't necessarily judge assumptions by how believable they are. We will see examples where unrealistic assumptions have benign consequences while other assumptions we might not even realize we are making can make us very overconfident.

Sometimes we can directly observe the consequence of our model choices in a guess-and-check way. The observed performance of a model is informative at least for the ways we judge it.

We want to understand when the things we learn from the known data generalize to the unknown data. We will consider several kinds of generalization, such as:

1) from held-in data to randomly held out data,

2) from a sample to future samples from the same distribution,
3) from one or more data sets to other similar data sets, and
4) from observational data to data where we have intervened.

1.3 There's x and there's y

Much of applied statistics is about understanding how some variable y depends on some variable x or some vector of variables $\boldsymbol{x}$. This x might take two values, such as 0 and 1, for treated/untreated or new/old or success/failure. It might take $k \geqslant 3$ distinct values. Sometimes those distinct values are ordered (e.g., small, medium, large) and other times they are not. Perhaps $x \in \mathbb{R}$ or $\boldsymbol{x} \in \mathbb{R}^p$. Rarely, there isn't a variable x, just some y values. We will study these cases.

In this course $y \in \mathbb{R}$. In other courses y is binary or categorical or even a vector $\boldsymbol{y} \in \mathbb{R}^k$. Many of the techniques for $y \in \mathbb{R}$ extend naturally to those other settings, so this course is foundational.

The setting we will consider most is where we are given a point $\boldsymbol{x}$ and want to say something about the corresponding y. We will have historical pairs $(\boldsymbol{x}_i, y_i)$ for $i = 1, \ldots, n$ with which to do this.

If we are given $\boldsymbol{x}$ and want to guess y by some value $\hat{y}$ we ordinarily would like to have $\hat{y} = \mu(\boldsymbol{x}) := \mathbb{E}(y \,|\, \boldsymbol{x})$. This choice minimizes our mean squared error. We can write

$$\begin{aligned}
\mathbb{E}((y - m(\boldsymbol{x}))^2 \,|\, \boldsymbol{x}) &= \mathbb{E}((y - \mu(\boldsymbol{x}) + \mu(\boldsymbol{x}) - m(\boldsymbol{x}))^2 \,|\, \boldsymbol{x}) \\
&= \mathbb{E}((y - \mu(\boldsymbol{x}))^2 \,|\, \boldsymbol{x}) + 2\mathbb{E}((\mu(\boldsymbol{x}) - m(\boldsymbol{x}))(y - \mu(\boldsymbol{x})) \,|\, \boldsymbol{x}) \\
&\quad + \mathbb{E}((\mu(\boldsymbol{x}) - m(\boldsymbol{x}))^2 \,|\, \boldsymbol{x}) \\
&= \mathrm{var}(y \,|\, \boldsymbol{x}) + 2(\mu(\boldsymbol{x}) - m(\boldsymbol{x}))\mathbb{E}(y - \mu(\boldsymbol{x}) \,|\, \boldsymbol{x}) \\
&\quad + (\mu(\boldsymbol{x}) - m(\boldsymbol{x}))^2 \\
&= \mathrm{var}(y \,|\, \boldsymbol{x}) + (\mu(\boldsymbol{x}) - m(\boldsymbol{x}))^2.
\end{aligned}$$

Now it is clear that $m(\boldsymbol{x}) = \mu(\boldsymbol{x})$ is optimal. Subtracting and adding the answer as above provides quite clear proofs when it is possible and the above decomposition into a bias $\mu - m$ and a variance runs through applied statistics. We could also have found that $\mu(\boldsymbol{x})$ is optimal via calculus, differentiating through the expectation. In more theoretical courses one would need to justify that step.

If $\boldsymbol{x}$ is random then it remains optimal to take $m(\boldsymbol{x}) = \mu(\boldsymbol{x})$. It is intuitively obvious that making the best choice for every random $\boldsymbol{x}$ is globally

optimal. Also

$$\begin{aligned}\mathbb{E}((y-m(\boldsymbol{x}))^2) &= \mathbb{E}(\mathbb{E}((y-m(\boldsymbol{x}))^2 \,|\, \boldsymbol{x})) \\ &= \mathbb{E}(\mathrm{var}(y \,|\, \boldsymbol{x}) + (\mu(\boldsymbol{x}) - m(\boldsymbol{x}))^2) \\ &= \mathbb{E}(\mathrm{var}(y \,|\, \boldsymbol{x})) + \mathbb{E}((\mu(\boldsymbol{x}) - m(\boldsymbol{x}))^2),\end{aligned}$$

which can never be below $\mathbb{E}(\mathrm{var}(y \,|\, \boldsymbol{x})) = \mathrm{var}(y) - \mathrm{var}(\mu(\boldsymbol{x}))$ and it equals that bound for $m = \mu$.

Our actual loss is not necessarily the square of our error. We like conditional means and squared error because they pair nicely with each other, and because they commonly lead to simple and interpretable expressions and computations. We might reasonably have chosen to minimize $\mathbb{E}(|y - m(\boldsymbol{x})| \mid \boldsymbol{x})$ over m and the minimizer is the conditional median of y given $\boldsymbol{x}$. Later when we consider quantile regression we will look at loss functions that are minimized by arbitrary quantiles of y given $\boldsymbol{x}$.

1.4 Example linear models

Here we give some examples of models for y given $\boldsymbol{x}$. We may write $y = \mu(\boldsymbol{x}) + \varepsilon$. Then calling $\mu(\boldsymbol{x})$ the signal and ε the noise we will have linear approximations to $\mu(\boldsymbol{x})$. If $\mu(\boldsymbol{x}) = \mathbb{E}(y \,|\, \boldsymbol{x})$ then $\mathbb{E}(\varepsilon \,|\, \boldsymbol{x}) = 0$. We will have a lot to say later about models for the noise.

Most of the examples here will be covered at length in later parts of the course. We will also have a unified notation for them.

Example 1 (Plain regression)**.** Here

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

for $x_i \in \mathbb{R}$. Here β_1 is a slope and β_0 is an intercept.

In class, I sketched the example that Weisberg cites about Forbes' work where x_i is the boiling point of water and y_i is the air pressure. That data fits tightly around a regression line with one relatively large error ε_i and slight curvature. If barometers are awkward and thermometers are easy to use, then this model lets you use a thermometer to measure air pressure.

I also sketched a hypothetical example where (x, y) pairs are scattered in an ellipsoidal region with near zero slope. Something intermediate: child's height versus parental height.

Example 2 (Classical multiple regression)**.** Here y_i is the value of a home and $\boldsymbol{x}_i$ contains a lot of features like the area (in square feet), number of

bedrooms, number of bathrooms, school district and so on. The formula is

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i.$$

One of the features could be the square of the home's area. This illustrates that the predictors x_i are not necessarily the raw observations but can be functions of them. That model would not be linear in area. The model is linear in β which is where the term linear model comes from.

Example 3 (Polynomial)**.** Now let $x_i \in \mathbb{R}$ and take $x_{ij} = x_i^j$. Then we have a polynomial model

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_i^j + \varepsilon_i.$$

This is clearly not linear in x_i for $k \geqslant 2$.

Example 4 (Two groups)**.** Here we take the plain regression model with a binary predictor

$$x_i = \begin{cases} 0, & \text{group 0} \\ 1, & \text{group 1.} \end{cases}$$

The groups could be control versus treated in an experiment or male versus female or nickel versus copper. Then $\mathbb{E}(y_i \,|\, x_i = 0) = \beta_0$ while $\mathbb{E}(y_i \,|\, x_i = 1) = \beta_0 + \beta_1$ and β_1 will usually be the scientifically more interesting variable.

Example 5 (k groups)**.** For a categorical variable $z \in \{1, 2, \ldots, k\}$ with $k \geqslant 2$ we can run the multiple regression model on predictors $x_{ij} = 1\{z_i = j\}$. This is sometimes called the one-hot encoding. The resulting representation is a 'cell-mean model' because $\beta_j = \mathbb{E}(y \,|\, z = j)$. Another common representation includes an intercept term along with $k - 1$ of the indicator variables. The interpretation of β_j then depends on which indicator was left out.

Example 6 (Multiple lines)**.** Now let $x_i \in \mathbb{R}$ and $z_i \in \{1, 2, \ldots, k\}$. The model

$$y_i = \beta_1 x_i + \sum_{j=1}^{k} \beta_{j+1} 1\{z_i = j\} + \varepsilon_i$$

places $\mathbb{E}(y)$ on one of k parallel lines.

$$y_i = \sum_{j=1}^{k} 1\{z_i = j\}(\beta_j + \beta_{k+j} x_i) + \varepsilon_i$$

removes the restriction that the lines are parallel. These models can be quite useful for extrapolating when x has a limited range for one or more of the values of z.

Example 7. Sometimes one of the predictors is an older or previous value of y_i. We might have

$$y_i^{\text{new}} = \beta_0 + \beta_1 x_i + \beta_2 y_i^{\text{old}} + \varepsilon_i.$$

If we freeze $\beta_2 = 1$ then we get

$$\Delta y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where $\Delta y_i = y_i^{\text{new}} - y_i^{\text{old}}$ is the difference in y values. We are then modeling the change in y_i. I gave an example of some of the 50 states implementing a policy that the others did not. In some settings we prefer to estimate β_2.

Example 8 (Plain product interaction)**.** The model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$$

can capture synergies such as (in a causal interpretation) the 'effect' of x_1 depending on x_2 and vice versa. Here $\mathbb{E}(y \,|\, \boldsymbol{x})$ is a 'ruled surface' linear in x_1 for each x_2 and conversely. In a full quadratic model we add quadratic predictors x_{i1}^2 and x_{i2}^2.

Example 9 (Multi-phase models/splines)**.** We can write $z_+ = \max(z, 0)$ for the positive part of z. Then the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - t)_+ + \varepsilon_i$$

is a two-phase linear model. It is continuous with a slope change at t. The model is not linear in t. A more flexible piecewise linear model has

$$y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{k} \beta_{j+1} (x_i - t_j)_+ + \varepsilon_i.$$

There are extensions to piecewise quadratic or piecewise cubic models. For example

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \sum_{j=1}^{k} \beta_{j+3} (x_i - t_j)_+^3 + \varepsilon_i,$$

and other more numerically stable alternatives.

Example 10 (ReLU)**.** A rectified linear unit (ReLU) model has

$$y_i = \sum_{j=1}^{k} \beta_j (\boldsymbol{x}_i^\mathsf{T}\theta_j - \omega_j)_+ + \varepsilon_i.$$

It is linear in β for fixed θ_j and ω_j.

1.5 Notation

An observation consists of a (column) vector $\boldsymbol{x}_i \in \mathbb{R}^p$ and a value $y_i \in \mathbb{R}$. The data set is $(\boldsymbol{x}_i, y_i)$ for $i = 1, \dots, n$. We arrange the predictors and responses into

$$X = \begin{pmatrix} \boldsymbol{x}_1^\mathsf{T} \\ \boldsymbol{x}_2^\mathsf{T} \\ \vdots \\ \boldsymbol{x}_n^\mathsf{T} \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

We will write $X = \begin{pmatrix} X_1 & X_2 & \cdots & X_p \end{pmatrix}$ for $X_j \in \mathbb{R}^n$.

The $\boldsymbol{x}_i$ are vectors of features. They are not necessarily the underlying real world observations. When we want to distinguish them we suppose that there is some real world vector $\boldsymbol{u}_i \in \mathbb{R}^d$ of measured quantities and then $x_{ij} = \phi_j(\boldsymbol{u}_i)$ for functions $\phi_1, \dots, \phi_p$. For a scalar real world quantity such as time we might use t_i and $x_{ij} = \phi_j(t_i)$.

Our linear model is then

$$Y = X\beta + \boldsymbol{\varepsilon} \tag{1.1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\mathsf{T}$. Observation i has $y_i = \boldsymbol{x}_i^\mathsf{T}\beta + \varepsilon_i$.

The intercept is a notational nuisance. We can incorporate it by taking X_1 to be a vector of n ones. In some lectures we will want to treat the intercept differently from the other predictors. For example we might want to center X_j by subtracting $\bar{x}_j = (1/n)\sum_{i=1}^n x_{ij}$ from each x_{ij}. Centering X_1 would turn it into a column of zeros and we won't want to do that. Sometimes we will use

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i.$$

We will call out the difference instead of putting a 0'th column into X.

1.6 Prerequisites

The proper background for taking this course is

1) linear algebra and multivariable calculus
2) one or more prior courses on the linear model
3) strong preparation and intuition on probability
4) courses in statistical theory (at least stats 200)
5) knowledge of R
6) hands-on experience modeling data

With hard work one of those deficits can be made up. With two or more, I recommend you wait a year if you can, or audit the course. Hands-on experience will help you see why some topics are interesting.

The probability background does not have to include measure theory. You should know about continuous and discrete random variables, the usual distributions, expectations and covariances and Bayes theorem. These are not advanced concepts but using them should be effortless for you.

This should not be the first course where you see $(X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$ and $(X^\mathsf{T}X)^{-1}\sigma^2$.

1.7 References and influences

This note is based on experience teaching and applying statistics. The main early influences have been writings by Box, Cox, Tukey, Friedman and Efron after a start by Lawless and the Kalbfleisch brothers at Waterloo. Other influences and examples: Buja, Hastie, Tibshirani and Gelman.