Linear models

Preparing for the example

- Let's look at the example from the ISL book (the book is available for you on Graham in the projects folder); I'm actually not going to ask you to do anything with this data, but, as practice for a typical workflow, copy the "Advertising" data from the **projects/def-emd/MayStorm24** folder into your scratch and cd into your scratch
- Moving forward, we will be installing a number of Python packages, so, as further practice for a typical workflow, create a virtual environment for yourself; conda provides one set of tools for doing this, but I usually use venv, which comes with Python

```
python -m venv ~/python-env/bootcamp
source ~/python-env/bootcamp/bin/activate
```

- Now you can install the only package you will need for this section, using pip, **pandas**
- Now open a Python in interactive mode (don't worry about doing this on a compute node - we won't be doing anything intensive)

```
import pandas as pd
advertising = pd.read_csv("Advertising.csv")
advertising.head()
```

Linear models

- Linear models are at the core of some of the most popular machine learning approaches today, notably **neural networks** and **support vector machines** (the latter of which we won't talk about)
- You need to understand linear models
- It is easier to introduce the concept of a linear model using *linear regression*, which many of you are familiar with
- Remember that the difference between classification and regression is that a regression model is designed to predict a continuous value, while a classifier is intended to predict a discrete class

- As we will see later, we can find intelligent ways to convert discrete classes to continuous values and vice versa, in order to be able to work with some of the same models
- A **linear model** is a model that has the following form:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- Keeping this simple, let's start by imagining the case where we only have one predictor variable

$$Y \approx \beta_0 + \beta_1 X_1$$

- We have a **regression** model if we assume that Y is a continuous value
- We say $\approx$, that is, approximately equal to, because we assume we have no hope of predicting the exact value of Y
- For the above problem, the book proposes the following linear model:

$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV}$$

- We introduce a new notation, $\hat{\square}$, to indicate that something is an estimate
- Thus we have the following (why?)

$$\widehat{\text{sales}} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{TV}$$

- How to find these estimates? The book proposes to minimize the **residual sum of squares**, which is the usual approach, i.e., minimize this quantity:

$$\sum_i (y_i - \hat{y}_i)^2$$

- As they show in the book, this optimization problem actually has an analytic solution (i.e., no need for trial and error)
- This is related to the fact that, in the hypothesis space for a linear regression, we just have one global optimum: if you consider all the solutions (pairs of β_0, β_1), the graph of how good they are in terms if RSS looks like a bowl (you may sometimes hear the word **convex** because it also looks like a contact lens)
- Sadly, most problems we study in modelling related to language don't have this property and thus require some kind of search in order to find a good solution (although it would be fascinating to discover that some of them could be could be); as such, one must in general be careful to assess whether poor results are due to issues with **optimization** rather than problems with the underlying model

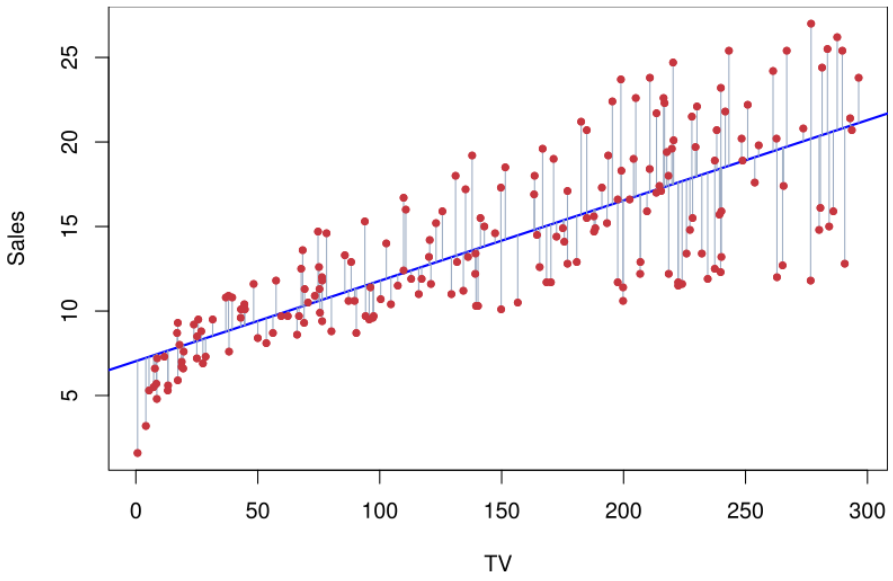

Figure 1: image

Linear model coefficients

- While a common characteristic of machine learning is that people throw a bunch of data into giant models and don't try and understand what they are doing, the coefficients of a linear model are easy to understand with a bit of reflection
- Empirically, they also appear to be **important** to understand: students who have trouble interpreting linear model coefficients (i.e., the β values) typically have some important issues with their understanding of linear models - and then have trouble making decisions about how to solve their modelling problems
- Thus it is worth explaining them before moving on to multiple regression: in this example from the book, the coefficients are in the first column
- Note that these are two separate models, fit with two different predictor variables
- What do their coefficients mean?

Multiple predictor variables

- Consider now the previous example of two separate regression lines. How could I now predict sales from both the *radio* and *newspaper* variables if I know both? (Think about it)
- What if radio and newspaper spending aren't independent of each other in

Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

Figure 2: image

the training data set? The clearest example is as simple as this: when our media budget expanded, we always just spent more on both; our company never sat down and did the experiment of manipulating one without the other

- Unless one sets out to do a controlled (balanced) experiment, one will run into the same problem: not all combinations of predictor variables have been seen, so there is no actual evidence about how they impact the output independently
- It's for this reason that the coefficients change when the book fits the larger model which is as follows:

$$\widehat{\text{sales}} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{TV} + \hat{\beta}_2 \times \text{radio} + \hat{\beta}_3 \times \text{newspaper}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

Figure 3: image

- More generally, each coefficient needs to be interpreted as the *effect of the given variable* ***holding all the other predictor variables constant***
- If there are other predictors we don't include in our model, we are really considering the data set at a sort of "average" value of the other predictors,

and only in very specific cases would that be without risk (i.e., would it not make a difference to what model we obtain)

- Before moving on, in addition to this justification for adding multiple predictors to a model, and the interpretation of each coefficient, make sure that you can visualize what is going on - this image, from the book, is only two, rather than three, predictors, to make it easier to plot

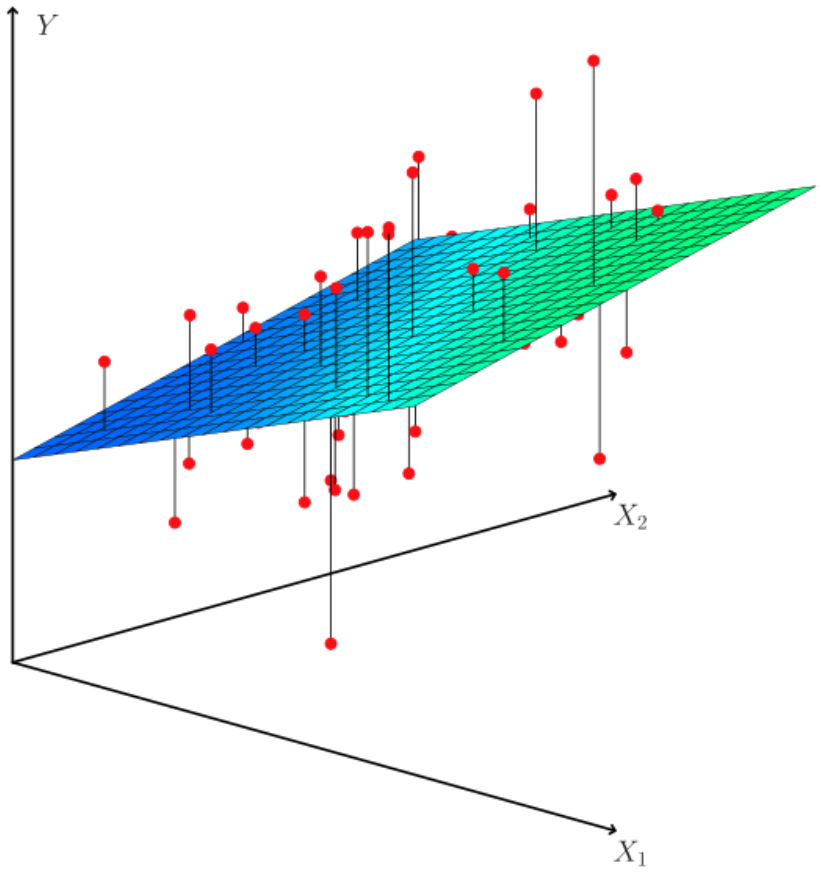

Figure 4: image