Chapter 11: Linear regression

Part 1: What is statistical learning?

https://dzwang91.github.io/stat324/

UNIVERSITY OF WISCONSIN–MADISON

Reference

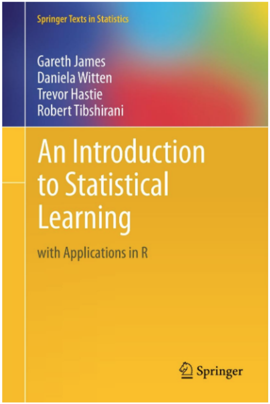


Figure: `http://www-bcf.usc.edu/~gareth/ISL/`

- Reading for today's lecture: Section 2.1.1

A motivating example: advertising data

- Suppose we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.
- The Advertising data set [1] consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.

[1] download the data set at `http://www-bcf.usc.edu/~gareth/ISL/data.html`

A motivating example: advertising data

- Suppose we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.
- The Advertising data set [1] consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.

```
> setwd("/Users/peterwang/Desktop/LinearReg")
> advertising=read.csv("Advertising.csv")
> head(advertising)
  X    TV radio newspaper sales
1 1 230.1  37.8      69.2  22.1
2 2  44.5  39.3      45.1  10.4
3 3  17.2  45.9      69.3   9.3
4 4 151.5  41.3      58.5  18.5
5 5 180.8  10.8      58.4  12.9
6 6   8.7  48.9      75.0   7.2
```

Figure: Advertising data

[1] download the data set at http://www-bcf.usc.edu/~gareth/ISL/data.html

A motivating example: advertising data

- It is impossible for our client to directly increase sales of the product. But they can control the advertising expenditure in each of the three media.
- Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales.
- In other words, our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.

A motivating example: advertising data

- It is impossible for our client to directly increase sales of the product. But they can control the advertising expenditure in each of the three media.
- Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales.
- In other words, our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.

We want to find a good function f such that

$$\text{Sales} \approx f(\text{TV, Radio, Newspaper})$$

Input variable and output variable

- The advertising budgets: input variables/predictors/independent variables/features
- Typically we use X to denote the input variables. For example, $X_1 =$TV budget, $X_2 =$radio budget, $X_3 =$newspaper budget

Input variable and output variable

- The advertising budgets: input variables/predictors/independent variables/features
- Typically we use X to denote the input variables. For example, X_1 =TV budget, X_2 =radio budget, X_3 =newspaper budget
- The sales: output variable/response/dependent variable
- Typically we use Y to denote the output variable. For example, Y = sale

Input variable and output variable

- The advertising budgets: input variables/predictors/independent variables/features
- Typically we use X to denote the input variables. For example, $X_1 =$TV budget, $X_2 =$radio budget, $X_3 =$newspaper budget
- The sales: output variable/response/dependent variable
- Typically we use Y to denote the output variable. For example, $Y =$ sale
- In general, suppose we have a quantitative response Y and p different predictors, $X_1, ..., X_p$. Then

$$Y = f(X_1, ..., X_p) + \epsilon$$

  - f is some fixed but unknown function of $X_1, ..., X_p$, and it's called regression function
  - f represents the systematic information that X provides about Y
  - ϵ is a random error term, independent of X and has mean 0

Example: income data

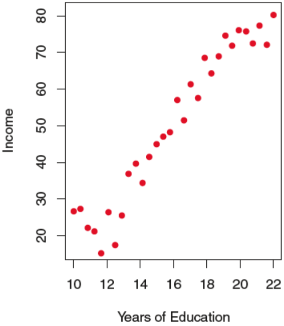


Figure: The red dots are the observed values of income (in tens of thousands of dollars) and years of education for 30 individuals.

Example: income data

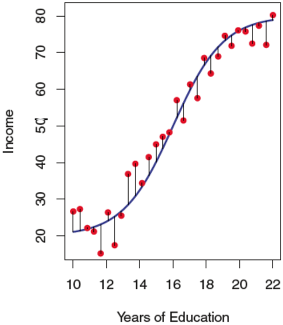


Figure: The blue curve represents the underlying relationship between income and years of education.

$$\text{Income} = f(\text{Years of education}) + \epsilon$$

Why estimate f: prediction

- In many situations, a set of input X are readily available, but the output Y cannot be easily obtained.

Why estimate f: prediction

- In many situations, a set of input X are readily available, but the output Y cannot be easily obtained.
- We can predict Y using

$$\hat{Y} = \hat{f}(X)$$

$$X \longrightarrow \boxed{\hat{f}} \longrightarrow \hat{Y}$$

Figure: $\hat{f}$ is treated as a black box

- $\hat{f}$ represents our estimate for f
- $\hat{Y}$ represents the prediction of Y

The accuracy of prediction

- We use mean squared error (MSE) to measure the accuracy of the prediction:

$$\text{MSE} = \mathbb{E}(Y - \hat{Y})^2$$

The accuracy of prediction

- We use mean squared error (MSE) to measure the accuracy of the prediction:

$$\text{MSE} = \mathbb{E}(Y - \hat{Y})^2$$

- MSE decomposition:

$$\begin{aligned}
\text{MSE} &= \mathbb{E}(f(X) + \epsilon - \hat{f}(X))^2 \\
&= \mathbb{E}[(f(X) - \hat{f}(X))^2 + 2(f(X) - \hat{f}(X))\epsilon + \epsilon^2] \\
&= (f(X) - \hat{f}(X))^2 + 2(f(X) - \hat{f}(X))\mathbb{E}(\epsilon) + \mathbb{E}(\epsilon^2) \\
&= \underbrace{(f(X) - \hat{f}(X))^2}_{\text{Reducible}} + \underbrace{Var(\epsilon)}_{\textit{Irreducible}}
\end{aligned}$$

The accuracy of prediction

- We use mean squared error (MSE) to measure the accuracy of the prediction:

$$\text{MSE} = \mathbb{E}(Y - \hat{Y})^2$$

- MSE decomposition:

$$\begin{aligned} \text{MSE} &= \mathbb{E}(f(X) + \epsilon - \hat{f}(X))^2 \\ &= \mathbb{E}[(f(X) - \hat{f}(X))^2 + 2(f(X) - \hat{f}(X))\epsilon + \epsilon^2] \\ &= (f(X) - \hat{f}(X))^2 + 2(f(X) - \hat{f}(X))\mathbb{E}(\epsilon) + \mathbb{E}(\epsilon^2) \\ &= \underbrace{(f(X) - \hat{f}(X))^2}_{\text{Reducible}} + \underbrace{Var(\epsilon)}_{\textit{Irreducible}} \end{aligned}$$

- Our goal for estimating f is to minimize the reducible error

Why estimate f: inference

- We are interested in
  - which predictors are associated with the response? Only a small fraction of the available predictors are substantially associated with Y, want to identify the important predictors
  - what is the relationship between the response and each predictor?
  - Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?
- Example: for the advertising data, we are interested in
  - which media contribute to sales?
  - how much increase in sales is associated with a given increase in TV advertising?

Next step of statistical learning

How can we estimate f?

Next step of statistical learning

How can we estimate f?

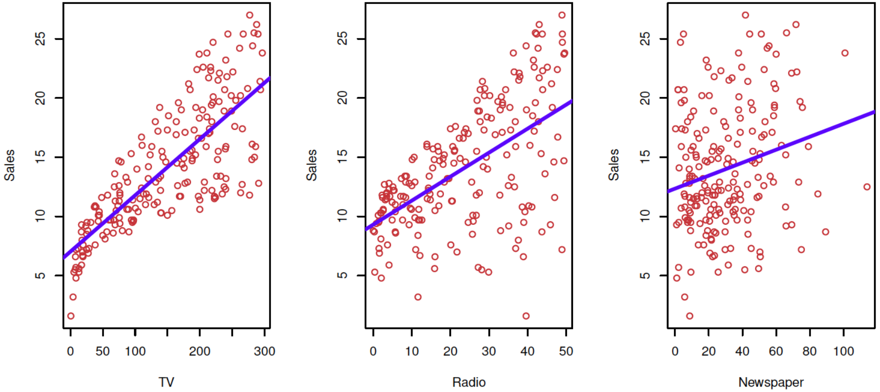
The most popular method: linear regression

Back to advertising data

What is the relationship between TV/Radio/newspaper advertising budgets and sales?

Back to advertising data

What is the relationship between TV/Radio/newspaper advertising budgets and sales?



“All models are wrong, but some are useful”
George Box

- True regression functions are never linear!

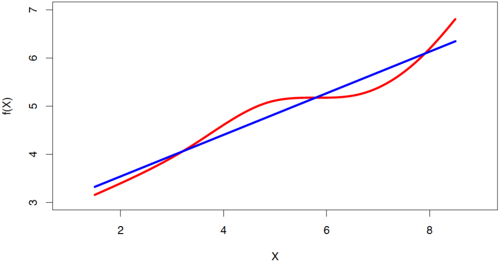


Figure: the red curve is the true regression function

- But linear regression is easy to implement and interpret!

What's the next?

We'll introduce more details on linear regression in next lecture.