Statistical Learning – Advanced Regression

Introduction

R.W. Oldford

Winter 2017

Contents

---

1 Setting the stage

Statistics has been described as the discipline which reasons about uncertainty, often by using formal probability models. It was born however in reasoning with data, typically data about the "state" (hence the rather dull name of "statistics" that the discipline has carried for centuries). And, because the reasoning is often about aggregate charactersistics of whole collections of individuals, probability models can be of considerable value in such reasoning.

Statistical learning is a broad term used to encompass a number of statistical methods that are, generally speaking, computationally intensive. It can be marked by algorithms and/or models that may be complex or by data which is large in number and possibly complexity. Learning has long been associated with statistical reasoning (e.g. consider Bayes theorem as a means of updating probabilities as data are accumulated) and as research areas in artificial intelligence (e.g. "machine learning") became more data-driven it is not surprising that the methods and logic of statistics would be adopted. Conversely, by trying to simulate intelligence in a machine, artificial intelligence has brought a pragmatic focus on scalable solutions as well as creative "learning inspired" methods (e.g. artificial neural networks) to statistical learning.

In this course, we will be focusing on a very particular problem, namely the estimation of a functional relationship between a **measured response variate** y, and one or more **explanatory variates** $x_1, \ldots, x_p$

(p is the number of explanatory variates). We imagine that the response and explanatory variates are *approximately* related through an unknown (to be estimated/learned) function $\mu(\mathbf{x})$:

$$y = \mu(\mathbf{x}) + r$$

where $\mathbf{x} = (x_1, \ldots, x_p)^\mathsf{T}$ is a vector of p different explanatory variates and r denotes the remainder or **residual** of the response y that is not explained by $\mu(\mathbf{x})$.

There are many other names for the variates in this model, each of which invokes a slightly different take on the purpose of the model. These names come in pairs as shown in the following table:

response	explanatory
response	predictor
response	design
output	input
dependent	independent
endogenous	exogenous

Similarly, **variates** are also often called **variables**, **features**, or **attributes**. All of these are in common use and may occasionally be used interchangeably in these notes. Overall, these notes will try to consistently use **variate** throughout, reserving **variable** for mathematical variables as opposed to measured values which can *vary of their own accord* (typically beyond our control). So too will **response** and **explanatory** be used primarily, although other adjectives might be used occasionally for emphasis of that variate's role.

Finally, instead of **residual** traditional authors have often used **error** to name the remainder r of y that is not explained by $\mu(\mathbf{x})$. Unfortunately, "error"" connotes "mistake" which is not how r is viewed in these models. Throughout, we will use **residual** for r and reserve "error" for those instances where the notion of "mistake" more clearly applies.

On notation and nomenclature

Throughout these notes:

- lower case Latin letters (i.e. $a, b, c, \ldots, z$) will be used to represent values of variates (i.e. realized, though possibly unobserved)
- upper case Latin letters (i.e. $A, B, C, \ldots, Z$) will be used to represent **random** variates (i.e. a mathematical quantity which takes values by drawing them from some distribution).
- The use of random variates is indicative of the presence of a **generative model** for the data, that is one which assigns probabilities to the different values that the variate might take.
- Greek letters ($\alpha, \beta, \gamma, \ldots, \omega$) will be used to denote parameters, the values of which are **unknown**. These are typically associated with the unknown values of a generative model.
- **boldface** will be used to denote vectors and matrices (typically lower and upper case, respectively)
- all vectors (e.g. $\mathbf{x}$) will be taken to be column vectors (and so $\mathbf{x}^\mathsf{T}$ is a row vector)
- **estimands** are unknown values, such as parameters (e.g. θ) or realized but unobserved values (e.g. r_i), which are to be estimated.
- **estimates** of unknown quantities are distinguished by adding a *hat* as in $\widehat{\theta}$ or $\widehat{r}_i$. *Estimates are numerical values.*
- **estimators** are random variates and are denoted by adding a *wig* as in $\widetilde{\theta}$ or $\widetilde{r}_i$. *Estimators have a distribution.*

An example or two might help fix ideas.

1.1 Example: Advertising data

On the course website, in a directory called `Data`, there is a file called `Advertising.csv` taken from the website associated with the book *An Introduction to Statistical Learning* by James et al. (After downloading this file, you will need to read this into an `R` session using the `read.csv(...)` function and assign it to the variable `Advertising`).

The data is purported to consist of some company's advertising data in each of 200 different markets. Three variates, `TV`, `Radio`, and `Newspaper`, record the amount in thousands of dollars spent for advertising on these three media for each market. A fourth variate, `sales`, records the number of units (in thousands) sold in each marketing after the advertising. The first few rows of the data are shown below:

```
head(Advertising)
```

```
##      TV Radio Newspaper Sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 151.5  41.3      58.5  18.5
## 5 180.8  10.8      58.4  12.9
## 6   8.7  48.9      75.0   7.2
```

The numbers in any row are **realizations** of the variates in that market; the column heading identifies the variate involved.

Here `Sales` is a response variate whose values might be related to one or more of the explanatory variates `TV`, `radio`, and `newspaper`. If such a relation exists, the company might be able to **predict** the number of units sold based on how much was spent on advertising in each of the three media. Such information could be invaluable for planning future budgets so as to maximize sales.

If **prediction** is the only consideration, then we might refer to the various media simply as **predictors** or predictor variates. If we knew $\mu(\mathbf{x})$, we might be able to predict the number of units sold in any new market for any combination of budgets to be spent on each of the three media. There would, of course, be some uncertainty in that prediction, unless all residuals are known to be zero – that is there is a precise functional relationship between the response variate and the explanatory variates.

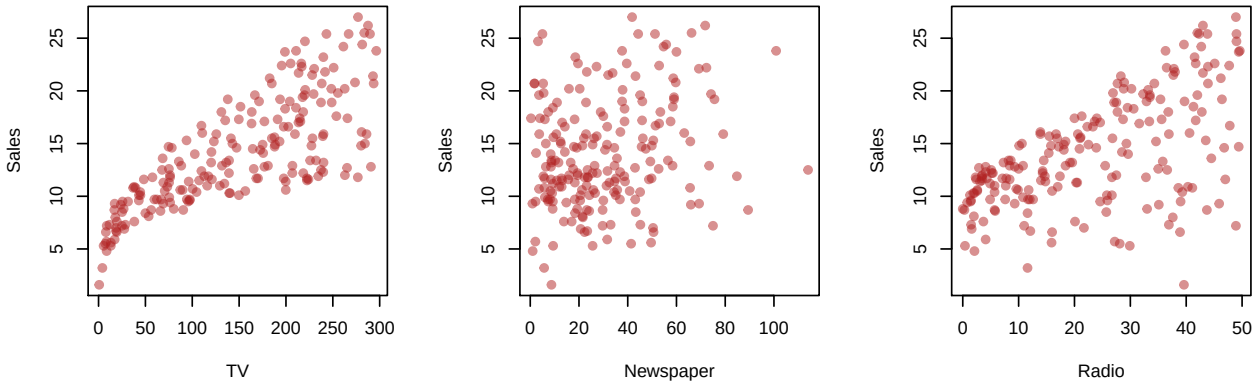
Alternatively, interest might lie in **explaining** the nature of the relation between the response and the explanatory variates. This would lead an understanding of how each of the **explanatory variates** affect the response and is often of the goal of any serious enquiry. Again, there would be some uncertainty in this explanation unless both $\mu(\mathbf{x})$ is known and the residuals are known to be zero.

In many applications, we are interested in both and so will almost exclusively use explanatory variate rather than predictor.

1.1.1 Fitting a model to data

We could, for example, consider modelling the response `Sales` as a function of any one of the three media budgets alone. A first step would be to look at the data itself.

```
savePar <- par(mfrow=c(1,3))
colour <- adjustcolor("firebrick", 0.5)
with(Advertising,
     { plot(TV, Sales, pch=19, col=colour)
       plot(Newspaper, Sales, pch=19, col=colour)
       plot(Radio, Sales, pch=19, col=colour)
       }
    )
```




```
par(savePar)
```

As the plots show, the number of units sold does seem to increase with the amount of money spent on each form of advertising. A very simple model to capture this increasing pattern might therefore be a simple straight line. That is, we posit the following functional relationship for simplicity for each explanatory variate separately:

$$y = \mu(x) + r$$

where

$$\mu(x) = \alpha + \beta x$$

and α and β are unknown parameters.

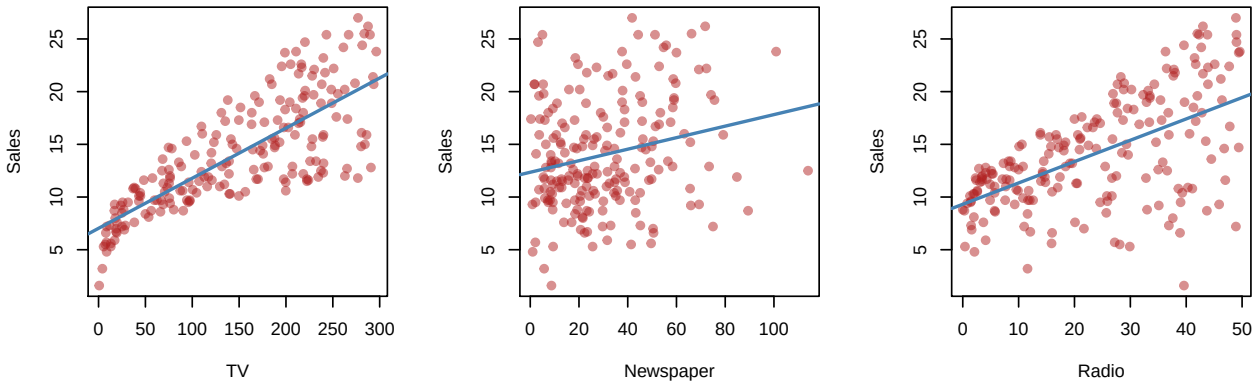
The data set `Advertising` contains observed realizations $y_1, \ldots, y_n$ for sales where $n = 200$ and i indexes the 200 markets as well as observed realizations for each of the explanatory variates x_1, x_2, and x_3 representing `TV`, `Newspaper`, and `Radio` respectively. The corresponding values of these expenditures in the ith market will be denoted x_{i1}, x_{i2}, and x_{i3} for $i = 1, \ldots, n$; the vector of realized explanatory variates for the ith market is denoted $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})^{\mathsf{T}}$.

We fit the straight line model above to each explanatory variate *separately* as follows:

```
fit1 <- lm(Sales ~ TV, data = Advertising)
fit2 <- lm(Sales ~ Newspaper, data = Advertising)
fit3 <- lm(Sales ~ Radio, data = Advertising)
```

and the fits can be overlaid on the plotted data:

```
savePar <- par(mfrow=c(1,3))
colour <- adjustcolor("firebrick", 0.5)
with(Advertising,
     { plot(TV, Sales, pch=19, col=colour)
       abline(fit1, col="steelblue", lwd=2)
       plot(Newspaper, Sales, pch=19, col=colour)
       abline(fit2, col="steelblue", lwd=2)
       plot(Radio, Sales, pch=19, col=colour)
       abline(fit3, col="steelblue", lwd=2)
       }
    )
```




```
par(savePar)
```

The fitted models themselves are not perfect but they do capture some sense of the increasing relationship between `Sales` and each explanatory variate, as seen in the plots.

For example, the fitted model for x_1 (the amount spent on `TV` advertising) would be as follows:

$$y = \widehat{\mu}(x_1) + \widehat{r}$$

where

$$\widehat{\mu}(x_1) = \widehat{\alpha} + \widehat{\beta} x_1$$

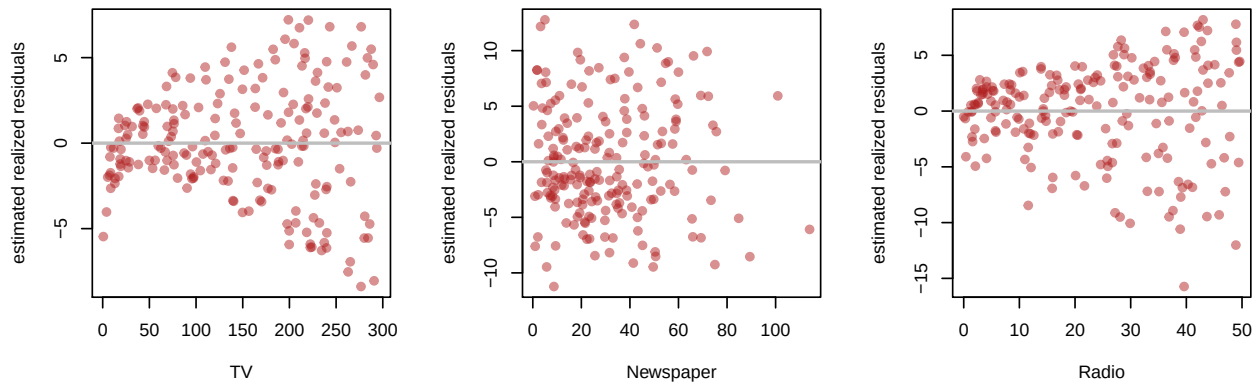
with $\widehat{\alpha} \approx 7.03$ and $\widehat{\beta} \approx 0.05$. These values are contained in `fit1$coefficients` and may be read (together with more information) from a summary of the fit in `R` as `summary(fit1)`.

All three fitted models (blue lines) suggest that as the amount spent x for whichever media increases the number of units sold increases ($\widehat{\beta} > 0$). The greatest increase is for `TV`, followed by `Radio` and then `Newspaper`.

Similarly, were $x = 0$ (i.e. 0 dollars spent on advertising in that medium) in some market there would still be many thousands of units sold (since $\widehat{\alpha} > 0$ for any of the three models). The simple straight line model for this problem is one where **both parameters are interpretable**. In many problems, the intercept parameter has no meaningful interpretation unless the explanatory variate has been centred properly (just as it meaningless here to predict values for sales when $x < 0$).

A closer look at the leftmost of these plots, however, reveals that the fitted function for `TV` might not adequately capture the relation between `Sales` and `TV` advertising. Rather than the straight line imposed by the model, the data seem to suggest a curve which actually drops down to zero units sold when zero dollars are spent on TV advertising! Rather than commit to a **globally defined** parametric function like $\mu(x) = \alpha + \beta x$, it might be desirable to have a function $\mu(x)$ that is **locally defined** and which can therefore adapt to local structure (here "local" is in the sense of neighbouring x values).

An examination of the **estimated realized residuals**, $\widehat{r}$, in the leftmost figure below



also shows this pattern as well (perhaps even more pronounced). Had the fitted function better captured this dependency of `Sales` on `TV` it should not have appeared in the residuals; that it can be seen in the residuals indicates that the functional form has missed this feature and might therefore be modified to better capture it. To a lesser extent, this may be true of expenditures on `Radio` as well.

Equally striking is that for both `TV` and `Radio` the residuals seem to spread out as the value of each explanatory variate x increases. This suggests that any prediction of `Sales` based on the amount of money spent on either `TV` or `Radio` advertising necessarily becomes **less and less certain** as the amount spent increases! Not that our least-squares fitted function models only the average of the ys; the increasing spread of the residuals observed suggests that their standard deviation might also increase with x.

1.2 Example: Facebook data

Again, in the course website's `Data` directory, there is also a file called `facebook.csv` which contains a subset of data collected by S. Moro, P. Rita and B. Vala as described in their recent paper :

> S. Moro, P. Rita and B. Vala (2016). "Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach". *Journal of Business Research*, *69*, pp. 3341-3351.

The data were downloaded from the University of California (Irvine) "Machine Learning Repository". (Again, you will need to read this into an `R` session using the `read.csv(...)` function and assign it to the variable `facebook`).

This study was conducted for a cosmetics company who had a Facebook page and wanted to see the effectiveness of their various postings on that page. Quoting their paper:

> "[…] we needed to collect a representative data set of published posts. All the posts published between the 1st of January and the 31th of December of 2014 in the Facebook's page of a worldwide renowned cosmetic brand were included. As a result, the data set contained a total of 790 posts published. It should be noted that Facebook is the most used social network with an average of 1.28 billion monthly active users in 2014, followed by Youtube with 1 billion and Google+ with 540 million (Insights, 2014)." - Soros et al (2016, p. 3342)

The data set found on the UCIrvine Machine Repository site contains only 500 of the 790 posts and a subset of the variates analysed by Soros et al (2016). The data uploaded to the course website is a further reduction of the 19 variates available to only 13. The variates are as follows:

- `share`: the total (lifetime) number of times the post was shared
- `like`: the total (lifetime) number of times the post "liked"
- `comment`: the total (lifetime) number of comments attached to the post
- `All.interactions`: the sum of `share`, `like`, and `comment`
- `Page.likes`: the number of "likes" for the facebook page at the original time of the posting
- `Impressions`: the total (lifetime) number of times the post has been displayed, whether the post is clicked or not. The same post may be seen by a facebook user several times (e.g. via a page update in their News Feed once, whenever a friend shares it, etc.).
- `Impressions.when.page.like`: the total (lifetime) number of times the post has been displayed to someone who has "liked" the page
- `Post.Hour`: the hour of the day at the original time of the posting (0-23)
- `Post.Weekday`: the day of the week at the original time of the posting (1-7) beginning with Sunday
- `Post.Month`: the month of the year at the original time of the posting (1-12)

- `Category`: the category of the post (as determined by two separate human reviewers according to the campaign associated with the post), one of `Action` (special offers and contests), `Product` (direct advertisement, explicit brand content), or `Inspiration` (non-explicit brand related content)
- `Type`: the type of content of the post, one of `Link`, `Photo`, `Status`, or `Video`
- `Paid`: 1 if the company paid Facebook for advertising, 0 otherwise

Here are the first few rows for each variate (in groups):

```
head(facebook[,1:4])
```

```
##   All.interactions share like comment
## 1              100    17   79       4
## 2              164    29  130       5
## 3               80    14   66       0
## 4             1777   147 1572      58
## 5              393    49  325      19
## 6              186    33  152       1
```

```
head(facebook[,5:6])
```

```
##   Impressions.when.page.liked Impressions
## 1                        3078        5091
## 2                       11710       19057
## 3                        2812        4373
## 4                       61027       87991
## 5                        6228       13594
## 6                       16034       20849
```

```
head(facebook[,7:10])
```

```
##   Paid Post.Hour Post.Weekday Post.Month
## 1    0         3            4         12
## 2    0        10            3         12
## 3    0         3            3         12
## 4    1        10            2         12
## 5    0         3            2         12
## 6    0         9            1         12
```

```
head(facebook[,11:13])
```

```
##      Category   Type Page.likes
## 1     Product  Photo     139441
## 2     Product Status     139441
## 3 Inspiration  Photo     139441
## 4     Product  Photo     139441
## 5     Product  Photo     139441
## 6     Product Status     139441
```

Again, the rows contain **realizations** (each row is a single post) of the variates named by each column.

In this collection of variates, there are certain obvious **explanatory variates**, namely all those in last two groups (from `Paid` to `Page.likes`), which precede all other variates in time.
There are also certain obvious **response variates** such as all those in the first group (containing all "interactions"). We might model any of these as a function of any of explanatory variates.

The remaining group, the second group of `Impressions`, also contains natural response variates that might be modelled as a function of the explanatory variates. However, *depending on the purpose of the analysis*, we might choose to model one or more of the interactions as function of these impressions since the impressions are the actual visual presentations to the facebook users who might subsequently interact with the post. Whether a variate is explanatory or response will depend on the context (e.g. which precedes which in time) and the purpose of the analysis.

A quick view of the range of values that are realized in the data for each group of these variates is given by the `summary(...)` function.

For the interaction variates

```
summary(facebook[,1:4])
```

```
##  All.interactions     share             like           comment
##  Min.   :   0.0   Min.   :  0.00   Min.   :   0.0   Min.   :  0.000
##  1st Qu.:  71.0   1st Qu.: 10.00   1st Qu.:  56.5   1st Qu.:  1.000
##  Median : 123.5   Median : 19.00   Median : 101.0   Median :  3.000
##  Mean   : 212.1   Mean   : 27.27   Mean   : 177.9   Mean   :  7.482
##  3rd Qu.: 228.5   3rd Qu.: 32.25   3rd Qu.: 187.5   3rd Qu.:  7.000
##  Max.   :6334.0   Max.   :790.00   Max.   :5172.0   Max.   :372.000
##                   NA's   :4        NA's   :1
```

we immediately see a number of things from this summary. For example, there are posts which have zero interactions. By far the greatest number of interactions are "likes", followed by "shares" and then "comments." Considering comments alone, 25% of the posts have at most one comment with half of them having at most three, and there is at least one post that has as many as 372 comments. The NAs at the bottom show that four posts have no information on the number of "shares" and one post has nothing on the number of "likes" it received.

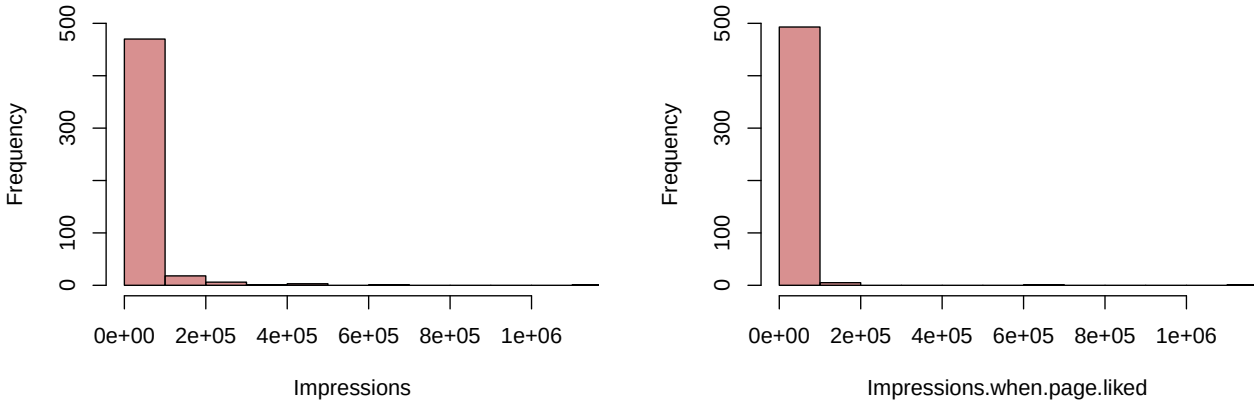
For impressions variates

```
summary(facebook[,5:6])
```

```
##  Impressions.when.page.liked  Impressions
##  Min.   :    567             Min.   :    570
##  1st Qu.:   3970             1st Qu.:   5695
##  Median :   6256             Median :   9051
##  Mean   :  16766             Mean   :  29586
##  3rd Qu.:  14860             3rd Qu.:  22086
##  Max.   :1107833             Max.   :1110282
```

we see that some posts have been presented more than a one million times to users. Closer inspection of the quartiles reveals that the distribution of realized values for each is also skewed, bunching up on low values ("short" left tail) and stretching out on the high values ("long" right tail).

A histogram makes this obvious.



Clearly, we can expect to see outliers in these data.

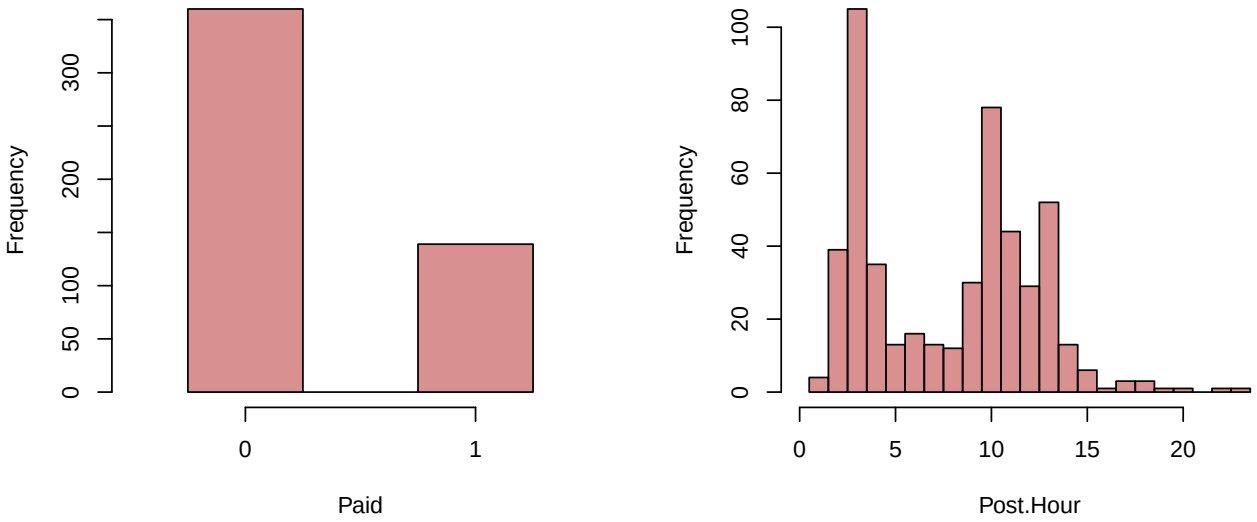
For the first group of basic explanatory variates we have

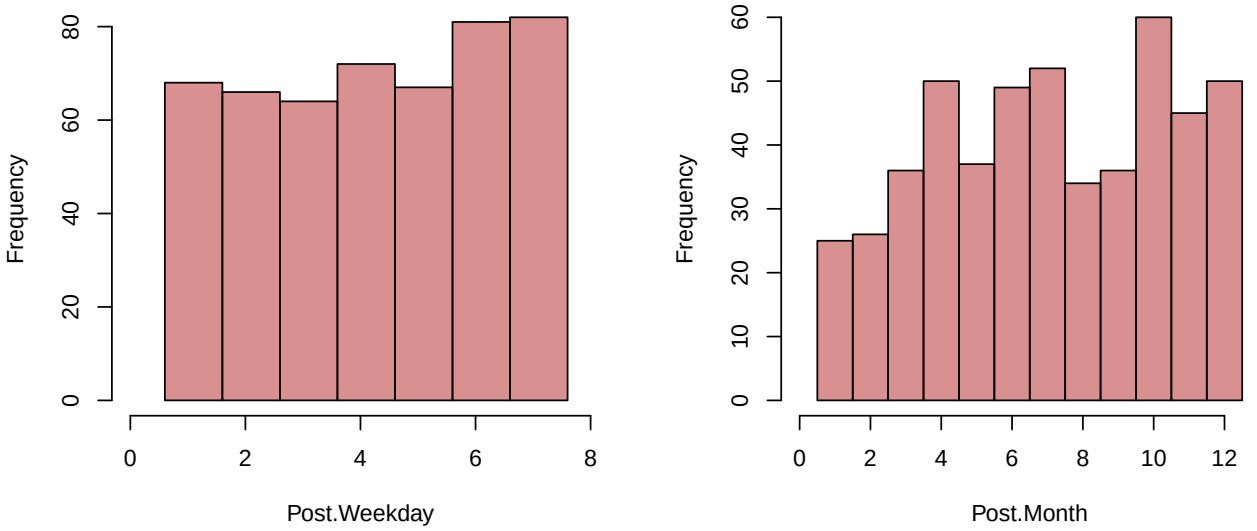
```
summary(facebook[,7:10])
```

```
##       Paid           Post.Hour      Post.Weekday    Post.Month
##  Min.   :0.0000   Min.   : 1.00   Min.   :1.00   Min.   : 1.000
##  1st Qu.:0.0000   1st Qu.: 3.00   1st Qu.:2.00   1st Qu.: 4.000
```

```
##  Median :0.0000   Median : 9.00   Median :4.00   Median : 7.000
##  Mean   :0.2786   Mean   : 7.84   Mean   :4.15   Mean   : 7.038
##  3rd Qu.:1.0000   3rd Qu.:11.00   3rd Qu.:6.00   3rd Qu.:10.000
##  Max.   :1.0000   Max.   :23.00   Max.   :7.00   Max.   :12.000
##  NA's   :1
```

from which we can see that for at least 25% of the posts Facebook was paid to provide advertising; for at least one post it is unknown whether Facebook was paid or not.
About 75% of the postings were made before 11 AM, with 25% between 1 and 3 AM, and another 25% between 9 and 11 AM. The other two variates (`Post.Weekday` and `Post.Month`) appear to be more uniformly distributed (perhaps as one might have expected). Again, histograms tell a more complete story:





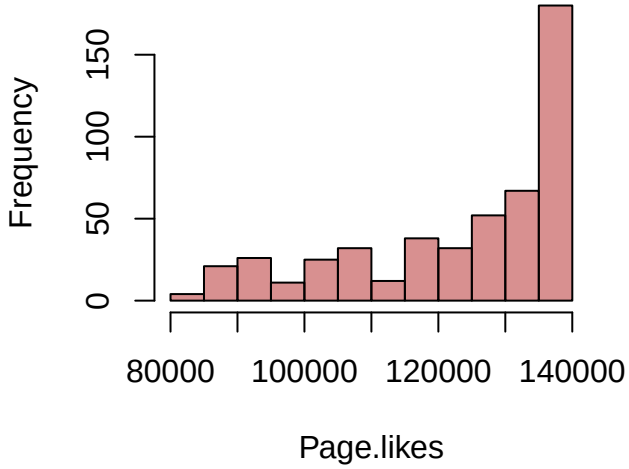
Finally, the `Category` and `Type` of the post as well as the number of likes for the company's Facebook page.

```
summary(facebook[,11:13])
```

```
##         Category       Type       Page.likes
```

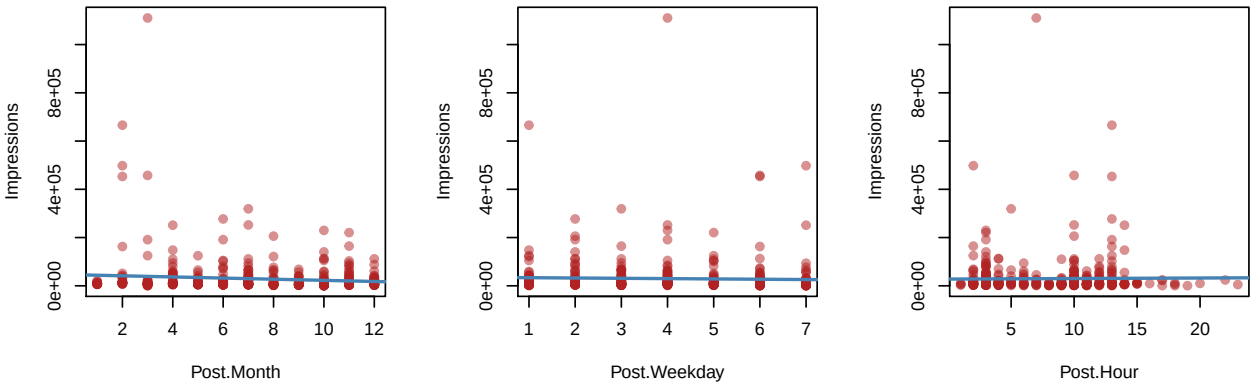
```
##  Action     :215   Link  : 22   Min.   : 81370
##  Inspiration:155   Photo :426   1st Qu.:112676
##  Product    :130   Status: 45   Median :129600
##                    Video :  7   Mean   :123194
##                                 3rd Qu.:136393
##                                 Max.   :139441
```

The most common category is that of an `Action` where the post is something like a contest or special offer. By far the most common `Type` of posting is that of photo. `Page.likes` indicate that many mode postings were made after the page was well established with a great many likes. Again, the histogram is more informative.



1.2.1 Fitting a model to data

Suppose we consider modelling the response `Impressions` as a function of any one of the three posting times and fit a straight line model in each. The results are shown below:



The previously observed skewness of the distribution of impressions now makes it very difficult to assess the nature of the fit.

At left it seems that the fitted line has a negative slope so that the number of impressions seems to decrease from January to December. However, it is very difficult to say for sure. Moreover, the outliers present in the early months of the year and these alone may have pulled the fitted line up towards them, artificially creating a negative slope – least-squares estimation is notoriously sensitive to outlying points. For the other two plots it is near impossible to see whether there is any non-zero slope in fitted line.

Indeed, the t-test associated with testing the hypothesis $H : \beta = 0$ in each of these models returns evidence against this hypothesis only for `Post.Month` model (leftmost plot) with an **observed significance level** of

about 0.02. For the other two models, the same test produced **no evidence against the hypothesis** that the slope is indeed zero.

Unfortunately, these t tests are based on the following **generative model** for our response variate given the value of the explanatory variate x

$$Y = \mu(x) + R = \alpha + \beta x + R$$

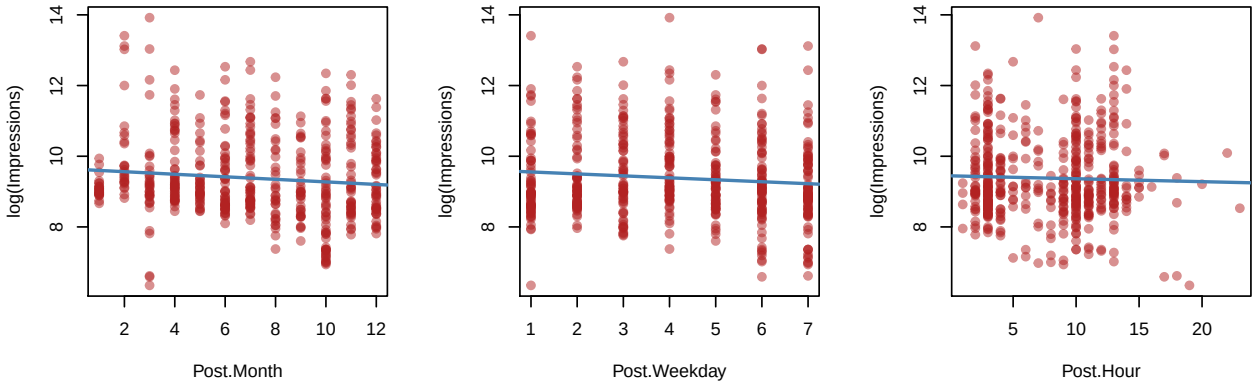
where Y and R are now **random variates**. The usual assumption is that the residual R follows a Gaussian, or Normal, distribution with mean zero and standard deviation σ.

The skewness observed for `Impressions` as the response variate puts the lie to any assumption of a symmetric distribution like $N(0, \sigma^2)$ as a generative distribution for the residuals here. It follows that our t test is itself suspect, and perhaps not to be trusted.

Instead of using `Impressions` as the response variate, we might use `log(Impressions)`. This should have the effect of making the response more symmetric and so pull in the large outliers. We make this **transformation** and fit each straight line model as before, except now take the response variate y to be the common logarithm ($\log_{10}$) of `Impressions`.

```
fit1 <- lm(log(Impressions) ~ Post.Month, data = facebook)
fit2 <- lm(log(Impressions) ~ Post.Weekday, data = facebook)
fit3 <- lm(log(Impressions) ~ Post.Hour, data = facebook)
```

The fitted lines are now overlaid on the corresponding data:



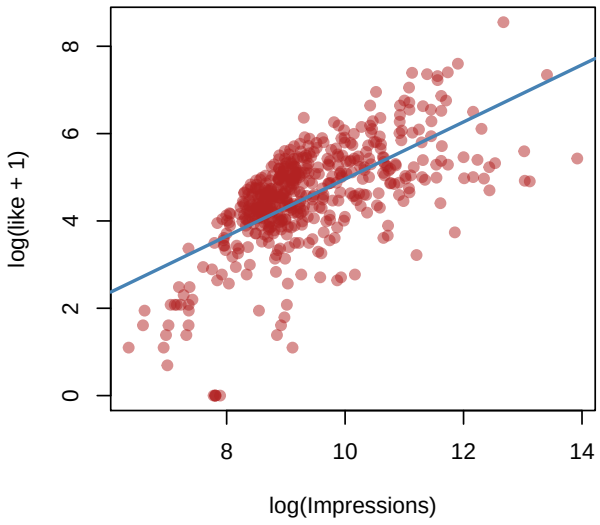
The log transformed are indeed more symmetrically distributed and the fitted slopes more obvious. In all three models now the slope `log(Impressions)` decreases with time. Recall that all posts occurred in the 2014 calendar year. The slope is strongest over months, less strong over days of the week (which are accumulated over all 52 weeks), and least of all for hours (accumulated over 365 days).

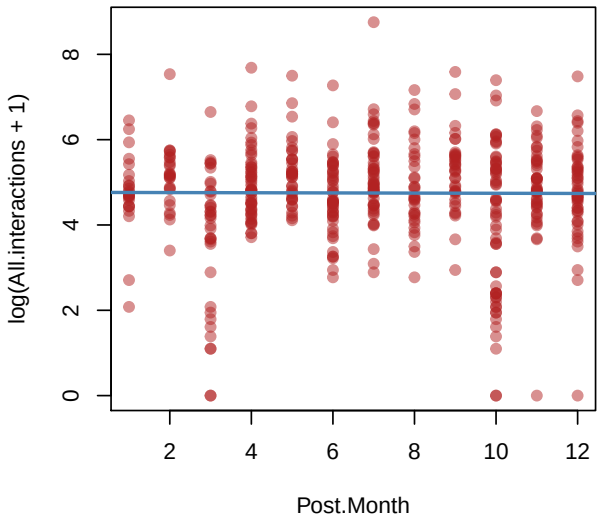
For this data, the t-test is significant for the first two plots, but not for the third. Again, this needs to be taken with a grain of salt, for while the estimated realized residuals are now much more symmetric than they were earlier, they still do not look as if they were generated by a normal distribution. Perhaps another transformation might produce more nearly normally distributed looking residual estimates.

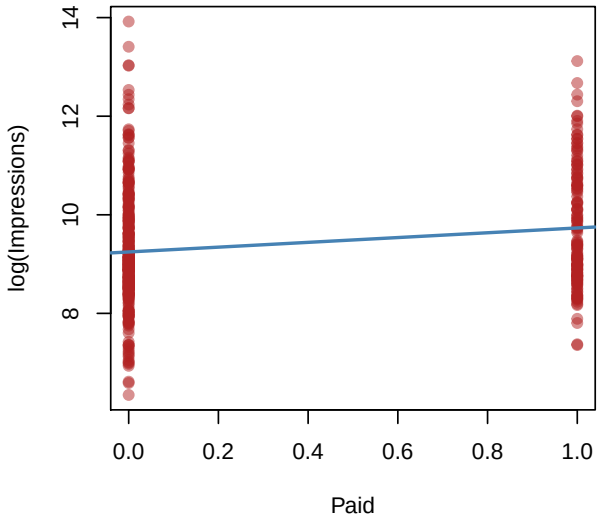
Some concern also still remains about the effect the various outlying points might have on the fitted lines. A fiting mechanism that was **less sensitive** or **more robust** to outlying points might be of value.

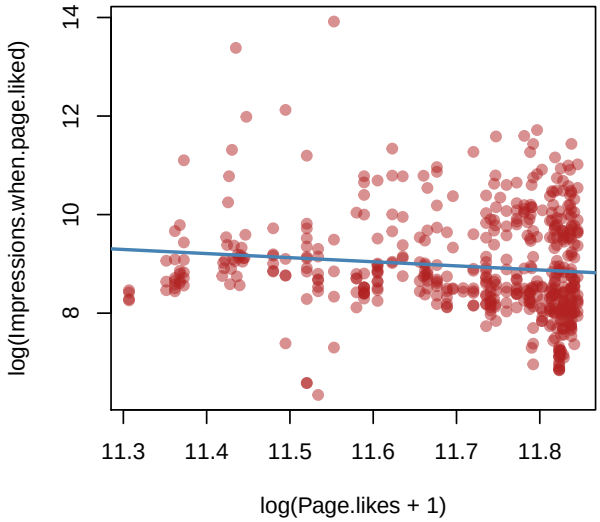
Again, rather than determine a **global model** for $\mu(x)$, it might be wiser find a means whereby the value of $\mu(x)$ was **fitted locally** in x and so could reproduce the ups and downs that can be seen in these plots (particularly for month and weekday).

Other response variates might be modelled as a function of other explanatory variates. Consider, for example, the plots and fitted lines below.









Again, for these plots the same concerns arise namely

- the effect of outliers on fitting
- global versus local modelling
- model complexity
- is there a reasonable generative model?

2 Statistical generalizability

Statistical thought necessarily presumes that we are interested in some non-empty collection of individuals. These might be individual markets, Facebook posts, people, things, places, sessions, time periods, words, documents, browser events, whatever. The important thing is that they are identifiable as individuals, each distinguishable from another, and that we have some number of them.

For the sake of clarity, let's refer to each such individual thing/person/time period/etc by the generic term **individual**. In a set of N such individuals we can imagine identifying each one and labelling it, say by an

index $i = 1, \ldots, N$. The label is arbitrary and only required to be unique to each individual. In this way it serves to identify individuals and to distinguish one individual from another.

So far, we are imagining at most a collection of individuals with little in the way of comparing one to another beyond their identity. To compare individuals we further imagine that there are one or more other characteristics or features of each individual that could **vary** from individual to individual. We will call such a characteristic or feature a **variate** and imagine it as a function that when applied to an individual will return a **value** for that variate. While not strictly necessary, it is typically possible that more than one individual could have the same value for any particular variate.

For example, if the individuals were persons, then "eye colour" might be a variate that returned the value "blue" or "brown" and so on, depending on the actual eye colour of that individual. A different variate for the same individuals might be "height" which could take values such as "190 cm" or "182 cm", et cetera.

Mathematically, we could introduce notation to represent such variates as follows. A variate x, say, would be some function $x(\cdot)$ which when applied to an individual i returns a value $x_i = x(i)$ say. Then if, as in our example, i denotes a person and x is "eye colour" then x_1 might be "blue" and x_2 "brown" for persons 1 and 2. A second variate, say y, would be used to denote the variate "height in centimetres" and we might analogously have $y_1 = 190$ and $y_2 = 182$.

Note that, to be meaningful, variates must be defined rather carefully. As we will see later in some applications, this is not always easy. In fact, sometimes even defining the individuals in a meaningful and useful way can be challenging!

A variate that is only ever determined on a single individual (and on no other) is not terribly interesting, or possibly even meaningful. But once we have the value of that variate for many individuals we have a means of comparing them to one another and things can quickly become very interesting indeed.

As mentioned earlier, depending on our purpose, some variates will be called **explanatory variates** and others will be called **response variates**.

2.1 Populations

In statistics, a set of individuals about which we would like to draw some conclusion is typically called a **population** of individuals. The word population has the same meaning as a set in this context in that each individual is unique, disinguishable from any other, and appears only once in the population. Like the word "statistics", "population" is a historical artefact from the discipline's earliest days when the largest and most interesting collections were related to the "state" and its "populace".

2.1.1 Population attributes

Once we have a population in hand, we attempt to summarize it. Any such summary will be called a **population attribute**. Notationally, we will use $\mathcal{P}$ to denote a population and, as with variates, population attributes can also be thought of as a function, this time of a population $\mathcal{P}$ rather than an individual. When we want to emphasise this we will write an attribute as $a(\mathcal{P})$.

Minimally, we can always imagine two possible summaries of any population, namely the count of how many individuals there are in that population, call it $N_{\mathcal{P}}$ say, and the set of labels that identify the individuals, for example being $\{1, 2, \ldots, N_{\mathcal{P}}\}$. These might be simply denoted as $a_1(\mathcal{P}) = N_{\mathcal{P}}$ and $a_2(\mathcal{P}) = \{1, 2, \ldots, N_{\mathcal{P}}\}$.

If we also have values of a variate, say $x_i = x(i)$ for every individual $i \in \mathcal{P}$, then there are numerous other summaries we might have for $\mathcal{P}$. For example, we might be interested in the following traditional population attributes (or "statistics"):

$$a_3(\mathcal{P}) \;=\; \overline{x} \;=\; \textstyle\sum_{i\in\mathcal{P}} x_i / N_{\mathcal{P}}$$

$$a_4(\mathcal{P}) \;=\; SD(x) \;=\; \sqrt{\frac{\sum_{i\in\mathcal{P}} (x_i - \overline{x})^2}{N_{\mathcal{P}} - 1}}$$

being the population average and standard deviation for the variate x on the population $\mathcal{P}$.

Other population attributes might be vector valued. For example, the `fivenum(...)` function in `R` produces a vector of five numbers as a summary of a collection of variate values. The attrribute is

$$\mathbf{a}_5(\mathcal{P}) = (min, Q_1, Q_2, Q_3, max)^\mathsf{T}$$

containing the minimum, the three quartile values, and the maximum of the values of some variate for all individuals in the population $\mathcal{P}$.

In this course, interest will often lie in population attributes which are **functions of explanatory variates**. This population attribute would be expressed as a **parametric curve**

$$a_6(\mathcal{P}) = (x, \mu(x))$$

relating the response y to a single explanatory variate x as in

$$y = \mu(x) + r$$

or as a **parametric surface**

$$a_7(\mathcal{P}) = (\mathbf{x}, \mu(\mathbf{x}))$$

with

$$y = \mu(\mathbf{x}) + r.$$

Clearly some population attributes are more interesting than others. Determining which population attributes are most interesting in any given problem is a challenge. Defining attributes of general interest that capture important, interesting, and useful features of a population has been a piece of statistical science for more than a century.

As such, we note in passing that population attributes are themselves a summary of the data, a reduction of the data as it were, and so follow the long and widely held view expressed early on by R.A. Fisher (1922, p. 311) that "… the object of statistical methods is the reduction of data." This is not to be confused with the idea that we choose to settle for small amounts of data, but rather that we should look for reliable means of extracting useful information from any amount of data, however voluminous. The goal remains to reduce data to its useful information, the interesting attributes of a population.

Note also that population attributes need not be restricted to those having numerical values (either scalar or vector). They could in fact be any picture, or visualization, that presents salient information about the population. That is, a population attribute of great interest might simply be **pictorial attribute**, such as a scatterplot of the pairs (x_i, y_i) for two variates evaluated on all individuals $i \in \mathcal{P}$.

2.1.2 Context: Facebook posts

To better appreciate the nature of populations, their individual elements, variates, and attributes, let's revisit the example of the Facebook posts made by our world renowned cosmetic brand.

As we have seen, there are a number of variates (both explanatory and response) which have been measured on every post made by the company on their Facebook page throughout 2014. They include `like`, `share`, `Impressions`, and so on. The researchers recorded values of 19 of these variates, though only 13 are presented here.

The **individuals** here are the posts themselves and each variate x (or y) can be thought of as a function on that post. The **variate** `like`, for example, returns the number of times a Facebook user clicked `like` on that post (wherever it appeared).

The values of many of these (e.g. `like`, `Impressions`, etc.) were determined by Facebook's performance measures. How reliable these determinations (or measurements) are is unknown to us. We do know, for example, that for four of the individual posts the number of likes was either not determined at all or simply

unknown, suggesting that the Facebook measurement systems in place at the time were not without error. The authors of the article provide no assessment of how accurate the Facebook determined values are.

The measuring procedure for at least one variate, namely `Category`, is described by Soros et al (2016). The determination of these categories and their value for each post was done manually. One employee of the cosmetics company looked at all 790 posts and for each post decided, in the context of the campaign at that time, whether it should be regarded as an `Action` post, an `Inspiration` post, or a `Product` post. The employee then recorded this value. Soros et al (2016, p. 3342) tell us that to minimize "the risk of misclassification due to typing error for being a manual procedure, another experienced professional in social media within the company validated this categorization for all the 790 posts [sic]." This was a clear attempt to reduce error in the determination (measurement) of this variate.

In Soros et al (2016), all 790 posts which appeared on the company's Facebook page are available for analysis. This collection of 790 posts constitute a **population**. Since this is the population available to be studied (on whose individuals we may determine all variate values), we will call it the **study population** and denote it by

$$\mathcal{P}_{Study}.$$

These are the population of individuals available for us to study, determining population attributes, etc.

Note, however, that this is not the population of greatest interest to Soros et al (2016). As quoted earlier, they go to some length to suggest that this collection of 790 posts is somehow "representative" of something more interesting to them. They write (p. 3341)
> Companies soon realized the potential of using Internet-based social networks to influence customers, incorporating social media marketing communication in their strategies for leveraging their businesses. Measuring the impact of advertisement is an important issue to be included in a global social media strategy [...] A system able to predict the impact of individual pub- lished posts can provide a valuable advantage when deciding to communicate through social media, tailoring the promotion of products and services. Advertising managers could make judged decisions on the receptiveness of the posts published, thus aligning strategies toward optimizing the impact of posts, benefiting from the predictions made.

They are hoping that whatever conclusions they draw from the study population of 790 posts in 2014, will apply in the future. That the "predictive tool" they develop here will also yield insight in the future, possibly with respect to posts in other social media like Google+ or YouTube. They clearly have in mind another population of posts that includes posts in the future. This population of posts, about which they would like to draw inferences, is called the **target population** and will be denoted by

$$\mathcal{P}_{Target}.$$

Finally, the authors have not made publicly available the whole study population of 790 posts. We have only a subset of these, a collection of 500 posts. For class, we have also removed 6 of the variates. This collection of 500 is called a **sample** and is always **a subset of the study population**, however selected. We denote the collection of individuals which constitute the sample as

$$\mathcal{S} \subseteq \mathcal{P}_{Study}.$$

In every respect $\mathcal{S}$ could be considered a population itself and might even sensibly be called a "sample population". Such nomenclature, while arguably legitimate, does unfortunately fly in the face of traditional statistical language and common English usage - **it is to be avoided** therefore and will not be used here. Nevertheless, treating $\mathcal{S}$ as its own population we can evaluate any population attribute on the sample in the same way we would for population $\mathcal{P}$.

Every statistical study has at least these three collections: the target population $\mathcal{P}_{Target}$, about which we would like to make inferences; the study population $\mathcal{P}_{Study}$, about which we can make inferences (since it is available to us); and the sample $\mathcal{S} \subseteq \mathcal{P}_{Study}$, which we actually perform calculations upon in order to make inferences about the study population.

2.2 Error

Having recognized that there are at least two populations (target and study), a sample, variates whose value have to be measured/determined for every individual, and attributes which may involve substantial calculation, it should be clear that the are several sources of error.

2.2.1 Sample error

If our purpose were to know the value of an attribute $a(\mathcal{P}_{Study})$ when $\mathcal{P}_{Study}$ is not available to us, we might use the value of $a(\mathcal{S})$ in its place. In this way, $a(\mathcal{S})$ would be an **estimate** of $a(\mathcal{P}_{Study})$. To emphasise this relationship we could write

$$a(\mathcal{S}) = \widehat{a}(\mathcal{P}_{Study}) = a(\widehat{\mathcal{P}}_{Study})$$

as an estimate of $a(\mathcal{P}_{Study})$. The second equality emphasises that we are explicitly thinking of $\mathcal{S}$ as an estimate of $\mathcal{P}_{Study}$ to achieve this.

There are two things to note from this simple relationship.

First, any difference between the actual values of the estimate $a(\mathcal{S})$ and the thing being estimated (the **estimand**) $a(\mathcal{S})$ is an **error**. In this particular case we call this error the **sample error**. The error will depend both on the actual sample and on the attribute being evaluated. With some abuse of notation we write

$$\text{sample error} \quad = \quad a(\mathcal{S}) - a(\mathcal{P}_{Study}).$$

For numerical attributes, this is easily determined mathematically; for pictorial attributes it is not precise and meant to be taken notionally (with possibly more precise quantification to be added). In the latter case, it at least suggests an awareness that the concept of sample error must apply here as well.

Second, that the sample error would be zero (or non-existent) when the sample $\mathcal{S}$ is replaced by the population $\mathcal{P}_{Study}$ means that the estimation is in some sense **consistent**. That is, if we actually had the population itself, then we would get the right answer – the estimation is consistent. More technically, this type of consistency is sometimes called **Fisher consistency** in the statistical literature, named after the great statistical scientist Ronald A. Fisher who identified it as an important criterion for estimation in 1922.

In the context of the Facebook posts, any population attribute (e.g. summary or fitted function, etc.) that can be determined for our sample of 500 posts could also have been determined from the 790 available to Soros et al (2016). The extent by which the two attributes differ is the sample error.

If our sample of 500 had been randomly selected from the 790 then we would have a statistical mechanism in place that would allow us to describe probabilistically how large the sample error might be. Unfortunately, we know only that the 290 omitted posts were omitted "due to confidentiality issues". The confidentially issues might be such that the 290 omitted posts are very different from the 500 available to us, so that the sample error might be quite large. Alternatively, the 290 missing posts might be very much like the 500 in hand and the sample error small. We are trusting that the posts excluded by Soros et al (2016) will contribute little to the sample error.

It might also be the case that the original investigators had access to several years of posts by the cosmetic company. If so, then this complete set of posts would be the study population and their sample would be just those posts made in 2014.

2.2.2 Study error

In the Facebook posts context, the target population is not well defined by Soros et al (2016). It might be all Facebook posts now and in the future by this particular cosmetic company, it might be the same by any cosmetics company, or perhaps any company whatsover. It might even be any social media post by any company anywhere any time. We don't know.

We do know, however, that the more distant is the target population from the study population the less reliable would inferences about the target population be when based on the study population.

A new source of error can now be identified. We wish to know the value of $a(\mathcal{P}_{Target})$ say but at this time can only possibly know the value of $a(\mathcal{P}_{Study})$. Any difference in these two values is due to the population available, $\mathcal{P}_{Study}$, not being identical to that of the target population $\mathcal{P}_{Target}$. We call this difference the **study error** and express this as

$$\text{study error} \quad = \quad a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target}).$$

Note that if as before we really only have a sample, namely the $\mathcal{S}$ the two page spreads taken only from the first book, our total error so far can be decomposed into different sources as

$$\begin{aligned} a(\mathcal{S}) - a(\mathcal{P}_{Target}) &= (a(\mathcal{S}) - a(\mathcal{P}_{Study})) + (a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})) \\ &= \text{sample error} \quad + \quad \text{study error} \end{aligned}$$

The existence of a target population that is potentially quite different from a study population is not at all unusual. Anytime, for example, that we would like to draw inferences or make predictions about the future we are in this situation – we can only study or examine things that exist and are available now, at the time of the study.

Moreover, in a case like that of the Facebook posts, the target population might in fact be changed by our analysis. Facebook users, or advertising competitors, may change their behaviour in light of how the advertiser changes behaviour as a result of the study. Thus will in effect change the target population, potentially dramatically increasing the study error. A well known example of the same worry is having future financial transactions, which we might be trying to model and predict now, change as investors and systems change in reaction to our modelling and predictions!

2.2.3 Measurement Error

A final common source of error which must always be kept in mind is that of measurement error. Recall that for Facebook posts, our values of the variates come to us at best third hand. First from Facebook's performance measures (which are not without error) or the categorization by the cosmetic company employees, second by the transcription and public storage of these values on the ICIrvine machine learning repository by the authors Soros et al (2016), and finally by the author of these notes who downloaded the data, removed some more variates, and replaced the numerical values for `Category` and `Type` which were used on UCIrvine by the corresponding nominal values (`Link`, `Status`, `Photo`, etc.) as determined from the paper Soros et al (2016). In each and every instance there is an opportunity for a **measurement error** to be made along the path to the realized values with which we have worked.

Measuring systems typically have at least three sources of error: the instrument or gauge used, the method used to conduct the measurement (including the entire process), and the human operator (if any) that is conducting the measurements.

2.3 Inductive inference

As should be clear by now, when we are trying to draw conclusions about attributes of a target population from a sample of measurements, there is a pretty clear inductive path that is always being followed.

At each stage there is a source of potential error. In this path the most worrisome step is that from study population to target population. These can be very different. And, unless the individuals in the study population have been selected from those of the target population according to some probability mechanism where it is possible for any individual in the target population to appear in the study population, any argument about the magnitude of this error is outside of statistical inference. It will require appealing to

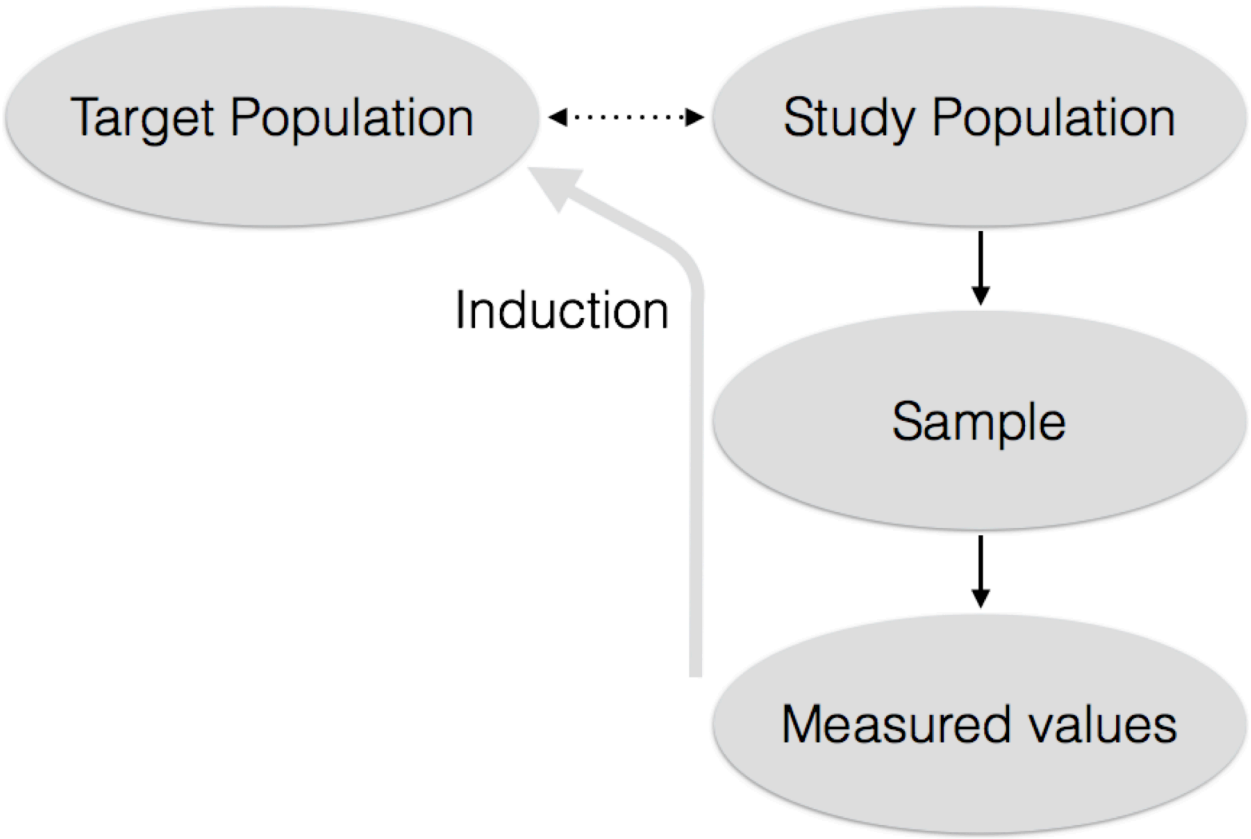


Figure 1: Path of Inductive inference

some other reasoning that links the two populations (or at least their attributes). One need only think of the study population as a collection of lab rats and the target population as a collection of humans to appreciate the potential difficulties in making the case.

Statistical reasoning can be brought to bear (and typically is) on discussing the potential magnitude of a sample error when a probabilistic mechanism for selecting a sample from the study population can be reasonably argued.

For example, suppose we have a mechanism that will choose a single sample $\mathcal{S}$ say from a set of possible samples, say $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M\}$. If that choice is random so that each sample, $\mathcal{S}_j$ say, will be selected only with some probability p_j (with $\sum_{j=1}^{M} p_j = 1$) then (at least notionally) for any attribute of interest, we could determine its **sampling bias** as

$$\text{sampling bias} \quad = \quad \left(\sum_{j=1}^{M} a(\mathcal{S}_j) \times p_j\right) - a(\mathcal{P}_{Study})$$

Clearly we would like to have a mechanism that made this bias small. Similarly we could define a measure of the **sampling variability** for the mechanism and attribute of interest. This might, for example, be the standard deviation of the attribute values over the samples selected randomly according to their probabilities.

Proper statistical sampling provides an insurance policy that what is learned about the attributes of the samples may be applied to the attributes of the (study) population with confidence. There is no guarantee that this application will be without error, however by careful planning the probability that the error is small can be made large.

Absolute certainty about a population attribute requires the entire population, anything less will not do. Fortunately, this degree of certainty is rarely required. There is of course one population attribute that will require the entire population and that is when the attribute is to a single particular individual in the population - the proverbial needle in a haystack. Or, Where's Waldo in this picture? If the population is every character appearing in the picture, then we will need to be able to, in principle, examine all characters - any sample of characters that does not contain Waldo must fail. And we must also be certain that we can recognize Waldo when we see him - without error! That is that our measuring system for recognizing Waldo must be without error - in real applications this is a tall order.

Ensuring that measuring systems are sufficiently reliable will itself require statistical study. All aspects of the system (gauge, method, operator) will need to be well studied to ensure that potential measuring errors have little effect of import on our conclusions.

Only in the case where the target population is identical to the sample and the variates are measured without error can we be certain about this induction. Otherwise some uncertainty inevitably remains.

2.3.1 The broader scientific context

In Scientific Method, Statistical Method, and the Speed of Light, published in the journal *Statistical Science*, an overview of scientific method and the nature of statistical method is described using, as illustration, the history of the determination of the speed of light. The paper makes the following points about statistical (and scientific) generalizability:

- "As regards induction, for statistics the problem can be neatly separated into two pieces … Ultimately, interests lies in the target population, as it is nearest to the broad scientific concerns of the problem. This population may be infinite, possibly uncountably so, and its definition can involve phrases like"all units now and in the future." Drawing conclusions about this population will often require arguments that are extra-statistical for they will be based on the similarities of, and differences between, the target population and the study population. Such arguments may ultimately be unable to avoid assuming Hume's "uniformity of nature" principle and hence what philosophers mean by the "problem of induction"."

- "Such weighty problems dissipate when focus shifts to drawing conclusions about the study population. Such is its definition that all study populations are finite in size and random selection of units to form a sample is possible. Random selection provides the strongest grounds for inductive inference. When, for whatever reason, random selection has not been employed then either the case that it has been near enough approximated, or that the sample is itself similar in its attributes of interest to the study (or target) population must be made. The latter is much like making the case for the transfer of conclusions from the study to the target population and so can be just as difficult. In either case, the arguments will to a large extent be extra-statistical."
- "The critical reader might suppose that the structure we propose is designed to relegate all the difficult problems to the realm of the"extra-statistical." But this is not sweeping them under the rug. Just the opposite. They are exposed as potentially weak links in the chain of inference about which statistics has nothing to say. (This does not preclude further statistical studies being carried out to address some of these problems (e.g., further investigation of study error).)"

2.4 What if?

Given a population and some attributes, we could ask a number of "what if" questions. For example, what if we hypothesized that the variate values in our population were generated by some mechanism? Possibly a randomizing mechanism? Could we tell whether the realized values of the variates were consistent with that mechanism? Or perhaps that they were inconsistent with that mechanism? How inconsistent are they?

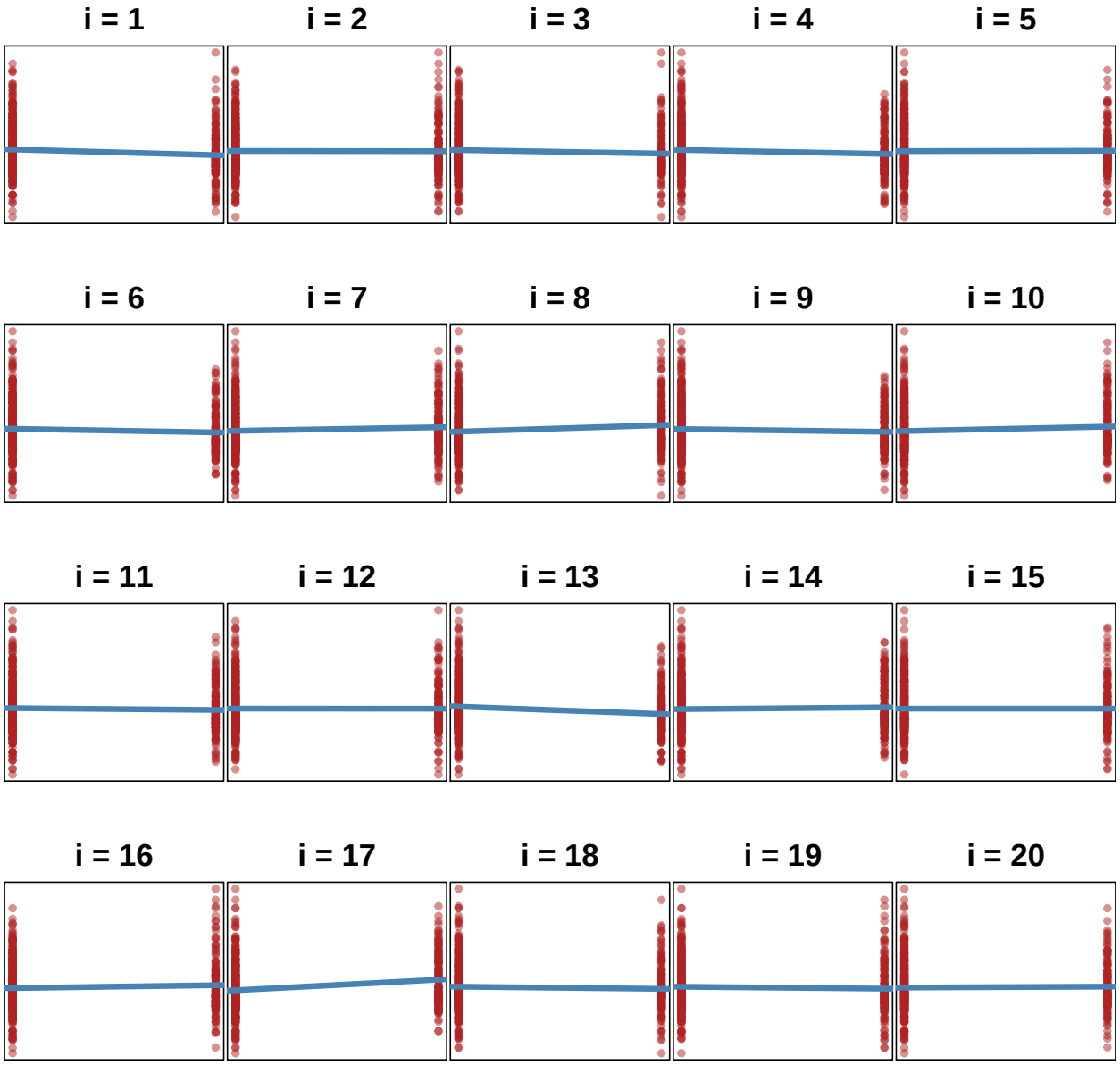
In the case of the Facebook posts, we might wonder whether the distribution of the `Impressions` was the same whether the Facebook had be paid to advertise the posts or not. Are the known values consistent with the hypothesis that there is no difference? Or not?

More formally, we imagine a **generative model** for the values of any variate. Given an individual i, the model generates values for each variate for i. The values generated are **realizations** of that variate.

We use **random variates** to represent a mechanism that randomly generates the values. Notationally, we denote a random variate by a capital letter, say X, and its realizations by a lower case letter, as in x. The random variate X would then follow some distribution, say $F_X(x)$ which might be completely or only partially known. For an individual i, X_i would be the random variate whose distribution generates values with particular probabilities, and x_i would be the variate value actually produced or realized.

The beauty of having a *generative model* is that we can *generate* values and see whether these values resemble (by some measure) the original data.

For example, consider the following plots:




```
## $trueLoc
## [1] "log(4.0238875743422e+112, base=29) - 60"
```

Of these 20 plots, only one is that of truly observed data. Its true location has been obfuscated and appears as the string in the above printed expression. The other 19 are independently generated plots of data whose variate values come from the hypothesized generative model. Each data set has the same number of points.

The plots are different scatterplots of some response variate against the value of a binary valued explanatory variate. Overlaid on each is the corresponding least-squares fitted line. Look over the plots carefully and try to choose which of these twenty looks has the fitted line with greatest absolute slope? Write down the number of the plot you select.

The true data is actually a plot of the `log(Impressions)` versus `Paid` variates from the `facebook` data. Under the hypothesis that paying for advertising had no effect on the number of impressions, the slope should be zero. Equivalently, we could randomly mix the impressions for posts that were paid for with those that were not. This is the generative model used here. All 20 plots use the actual `log(Impressions)` from the data, but only one preserves the correct value of `Paid` for each post. The rest randomly assign (without replacement) the values of `Paid` to each post.

Now evaluate the true location and compare it to the number you wrote down. Does it match? If the hypothesis is true, that paying Facebook has no effect on the number of impressions generated for a post, then you are effectively selecting a plot at random. The chance that you selected the plot of the truly observed data is therefore simply 1/20 or five percent. If you did, then either you were very lucky (probability 1 in 20 of selecting the true location) or the hypothesis is false.

This is a statistical significance test, much like those you have seen elsewhere (e.g. t-test, z-test, χ^2 test, etc.). A major difference here is that you are actually part of the test. If you do choose the plot of the observed data, then, the level of significance is 0.05 and you have evidence against the hypothesis. Had there been 100 plots, and you chose the data from them, the significance level would be 1/100 or 0.01 and the evidence against the hypothesis would be even stronger.

Of course, if you did not choose the plot, then the significance level is greater than 0.05 and there is no evidence against the hypothesis that there is no difference due to paying Facebook or not.

3 A data menagerie

For each data set in this section, consider the following questions:

1. What is the Target Population?
2. Could the target population change in response to the analysis?
3. What is the Study Population?
4. What is the sample?
5. What are the individuals in the sample?
6. What are the variates? How would they be determined/measured?
7. Which variates might be considered response? which explanatory?
8. What function would be an interesting population attributes? How might it be estimated?
9. What generative models, if any, might you be interested in for this data?

3.1 The Titanic

Figure 2: The great ship

From the `help(Titanic)` description in `R`:

"The sinking of the Titanic is a famous event, and new books are still being published about it. Many well-known facts—from the proportions of first-class passengers to the 'women and children first' policy, and the fact that that policy was not entirely successful in saving the women and children in the third class—are reflected in the survival rates for various classes of passenger.

These data were originally collected by the British Board of Trade in their investigation of the sinking. Note that there is not complete agreement among primary sources as to the exact numbers on board, rescued, or lost.

Due in particular to the very successful film 'Titanic', the last years saw a rise in public interest in the Titanic. Very detailed data about the passengers is now available on the Internet, at sites such as Encyclopedia Titanica."

The data set `Titanic` provides "information on the fate of passengers on the fatal maiden voyage of the ocean liner 'Titanic', summarized according to economic status (class), sex, age and survival."

The four variables are

No.	Name	Levels
1	Class	1st, 2nd, 3rd, Crew
2	Sex	Male, Female
3	Age	Child, Adult
4	Survived	No, Yes

3.2 Swiss fertility data

Figure 3: Swiss 1888 images

For example, consider the `swiss` data from `R`. This data consists of a standardized fertility measurement and five other standardizded socio-economic measurements taken on each of 47 different french speaking Swiss cantons/districts in 1888. Around this time Switzerland was entering a demographic transition period where its fertility was beginning to drop from the high level that is typical of underdeveloped countries.

The variates are:

- *Fertility*, here a 'common standardized fertility measure' denoted 'Ig' and scaled to be in [0, 100],
- *Agriculture*, as % of males involved in agriculture as occupation,
- *Examination* as % draftees receiving highest mark on army examination,
- *Education* as % education beyond primary school for draftees
- *Catholic* as % of 'catholics' (as opposed to 'protestants') in the canton, and
- *Infant.Mortality*, a rate, being the number of live births who die before reaching 1 year old per 1,000 live births.

Source info https://opr.princeton.edu/archive/pefp/switz.aspx

3.3 Great white shark encounters

Figure 4: Friendly great white shark

Data on known great white shark encounters with humans has been gleaned by Prof. P-J Bergeron of the University of Ottawa from a variety of tables appearing on the (now defunct) site http://sharkattackinfo.com/shark_attack_news_sas.html

Several variates recorded for $n = 65$ encounters where someone was bitten by a great white shark:

- *Year* - the year in which the encounter occurred
- *Sex* - sex of the victim (M = male, F= female)
- *Age* - age of the victim in years
- *Time* - time of the encounter (AM or PM)
- *Australia* - binary indicator whether encounter occurred in Australian waters (1 if in Australia, 0 otherwise)
- *USA* - binary indicator whether encounter occurred in US waters (1 if in USA, 0 otherwise)
- *Surfing* - binary indicator of whether the victim was surfing at the time of the encounter (1 if yes, 0 if no) – N.B. other unrecorded activities might be "free diving", "fishing", "pearl diving", etc.
- *Scuba* - binary indicator of whether the victim was scuba diving at the time of the encounter (1 if yes, 0 if no) – N.B. other unrecorded activities might be "free diving", "fishing", "pearl diving", etc.
- *Fatality* - whether the victim died after being attacked though not necessarily directly because of the attack (1 if yes, 0 if no)
- *Injury* - whether the victim was injured by the encounter (1 if yes, 0 if no)
- *Length* - the recorded length in inches of the shark thought to have encountered the victim.

There are numerous questions one might ask of this dataset. For example, does survival depend on the size of the shark involved? Is it safer to scuba dive or to surf? Are females more likely to be injured than males? What are the chances that I would survive an encounter by a "Jaws" size (25 feet, or 300 inches) great white shark?

The data will be available from the course web site.

Figure 5: Rats

3.4 Mammary tumours in rats

Here we consider data from a study of the development of mammary tumours in rats (as reported in Gail et al. *Biometrics* Vol. 36, No. 2 (Jun., 1980), pp. 255-266).

This study was a carcinogenicity experiment in which seventy-six animals were injected with a carcinogen for mammary cancer at day zero. All animals were then given retinyl acetate (Vitamin A) to prevent cancer (i.e as a "retinoid" prophylaxis) for sixty days.

After 60 days, 48 animals remained tumour-free. Of these, 23 were then randomly assigned to have continued retinoid prophylaxis treatment (the treatment group) and the remaining 25 were assigned to have no further treatment (the control group).

From the time of randomization on, the 48 rats were regularly examined for the development of new tumours, and multiple tumours could develop. At each time of examination the number of new tumours that were detected in each rat was recorded. Observation ended 122 days after the randomization.

The aim here is to estimate the expected number of tumours in the two groups and make treatment comparisons.

3.5 Fatty Acid Composition of Italian Olive Oils

Reference: Forina, M., Armanino, C., Lanteri, S., and Tiscornia, E. (1983) "Classification of Olive Oils from their Fatty Acid Composition", in Food Research and Data Analysis (Martens, H., Russwurm, H., eds.), p. 189, Applied Science Publ., Barking.

This data set records the percentage composition of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic) found in the lipid fraction of 572 Italian olive oils. The oils are samples taken from three Italian regions with a varying number of areas within each region. The regions and their areas are recorded as shown in the following table:

Region	Area
North	North-Apulia, South-Apulia, Calabria, Sicily
South	East-Liguria, West-Liguria, Umbria
Sardinia	Coastal-Sardinia, Inland-Sardinia

The data set `olive` is available from several `R` packages including `loon` and `RnavGraph`.

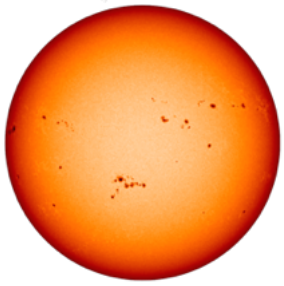
3.6 Sunspots

Reference Andrews, D. F. and Herzberg, A. M. (1985) Data: A Collection of Problems from Many Fields for the Student and Research Worker. New York: Springer-Verlag.

These are monthly mean relative numbers of sunspots from 1749 to 1983. They were collected at the Swiss Federal Observatory in Zurich until 1960, then at the Tokyo Astronomical Observatory.

The dataset is called `sunspots` in `R`.

Figure 6: Italian Olive Oils



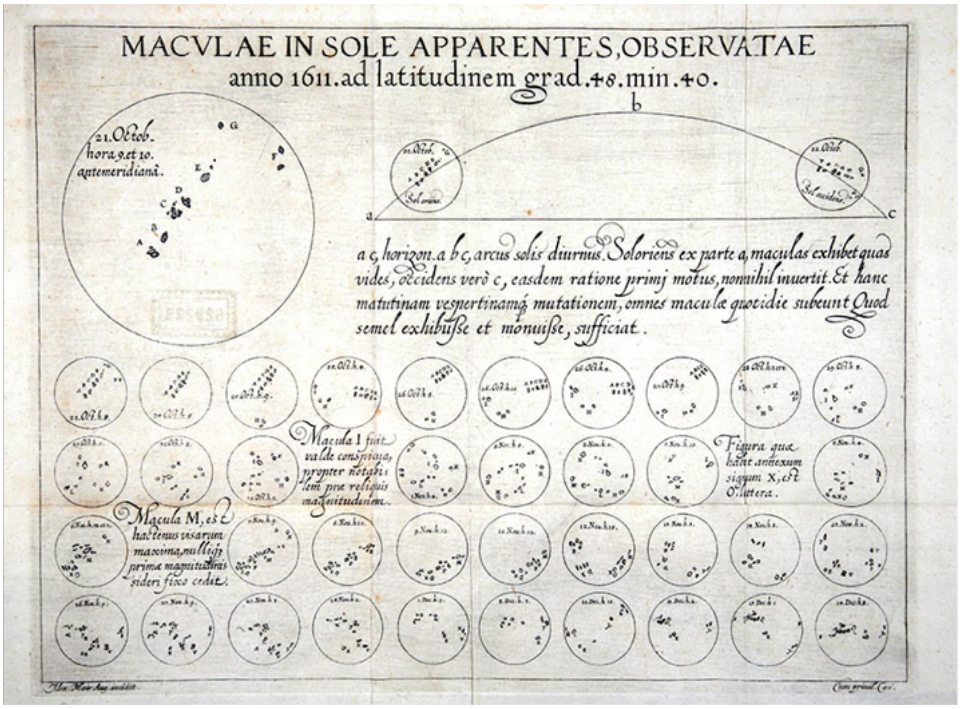
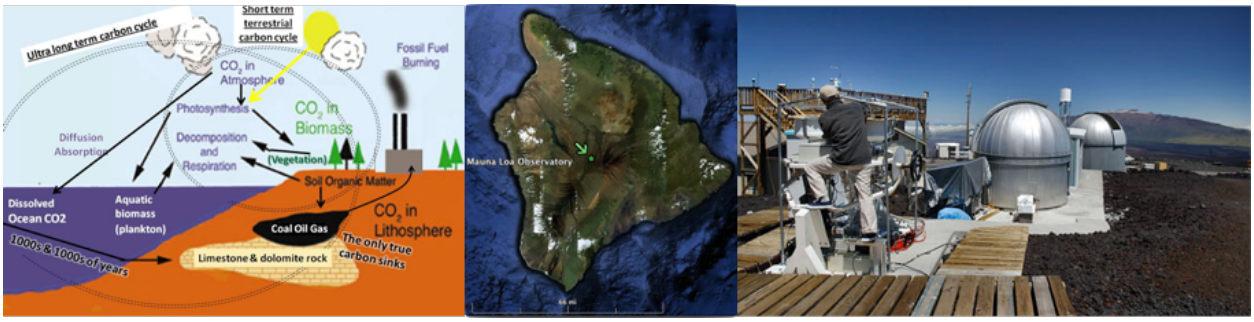


Figure 7: Sunspots

3.7 Atmospheric CO_2 concentrations



These are the atmospheric concentrations of carbon-dioxide over Mauna Loa in Hawaii. Concentrations are in parts per million and were reported in the preliminary 1997 SIO manometric mole fraction scale. The measurements were recorded monthly from and including January 1959 to and including December 1997.

It is available in `R` as the dataset `co2`.

3.8 NASA Earth surface data

Figure 8: Earth Surface image

The data are geographic and atmospheric measures on a very coarse 24 by 24 grid covering Central America. Measurements at each grid are: temperature (surface and air), ozone, air pressure, and cloud cover (low, mid, and high). All variate values are monthly averages, with observations for Jan 1995 to Dec 2000.

The latitude and longitude of each geographic grid are recorded as well.

There are in total 41,472 = (6 years * 12 months) * (24 * 24) individual locations in space and time and seven variates measured at each.

These data were obtained from the NASA Langley Research Center Atmospheric Sciences Data Center.

The dataset is available as `nasa` in the `R` package `dplyr`

4 Statistical learning - function estimation

Statistical learning is associated with a collection of statistical methods and problem contexts. The term "statistical learning" is fairly new but the underlying concepts and concerns are not.

As with other statistical contexts, interest often lies in determining a model based on observed data which will allow us to predict outputs from previously unseen inputs. Our observed data came from our sample, our unseen inputs from the study, or even target, population. In statistical learning we try to emulate this distinction between sample and population by breaking our sample into at least two pieces, one we call the **training set** the other the **test set**. We fit our models on the training set and evaluate them on the test set. This is key to an honest assessment of how well our fitted models generalize.

In this course we will be considering models that functionally relate some response variate to one or more explanatory variates, as in

$$y = \mu(x) + r.$$

These are often called **response models** for obvious reasons. They are also called **regression models**, largely for historical and traditional reasons.

For the most part, in this course we focus on models and methods where y is quantitative and has values on a real-valued scale where ratios are meaningful. That is not to say that cases where y is qualitative are not important. They are and are perhaps the most common models used in statistical learning. For example, oftentimes the values of the response are binary classes (e.g. "survived" or "died", as in the Titanic data) or multiple class (e.g. "region" for the olive oils data set). The objective is to determine a functional rule which will predict which class a new individual belongs to based on its associated explanatory variates. The statistical learning course which deals with these models and methods is STAT 441/841 Statistical Learning - classification. Here the goal is primarily prediction. Another statistics course which deals with a quite general modelling framework for this kind of data and which is largely concerned with **interpretability** of the models is STAT 431/831 Generalized Linear Models and their Applications.

Whatever the type of value y takes, models for this kind of statistical learning are often called **supervised learning**, "supervised" because we have the values of the response in hand when we create the model. The term "supervised" is perhaps most meaningful in the context where y is a class label, that is a categorical response. Here we imagine $\mu(\mathbf{x})$ as a classifier function that produces the class label of any individual based on the values $\mathbf{x}$ of its explanatory variates. Estimating $\mu(\mathbf{x})$ is "supervised" in the sense that our sample of individuals have both the values of the explanatory variates $\mathbf{x}$ **and** their label y (e.g. as seen STAT 441/841).

Oftentimes, we have no response y but only explanatory variates. In such cases, we may simply be interested in understanding the relationships, if any between the variates. We might be looking for functional relationships, we might be looking for lower dimensional structure, we might be looking for groups of like individuals on the basis of their $\mathbf{x}$ values. The last of these objectives is often sometimes cast as trying to "learn the class labels" without knowing what (or even how many of them there are) they are in any situation. This is traditionally called cluster analysis, but because if its similarity to classification and hence "supervised learning", it is now popularly known as **unsupervised learning**, "unsupervised" because the class labels are unknown. The term "unsupervised learning" is now applied more broadly as well. Of course, if we have some individuals in our sample for which y and $\mathbf{x}$ are known and other individuals for which only $\mathbf{x}$ is known (i.e. an example of the more general missing data problem in Statistics), the learning problem is sometimes called **semi-supervised learning**.

In all of the above, the concepts and concerns of statistical reasoning are inescapable and need to be addressed in any application.

Again, this course will be focused almost exclusively on the function estimation for the response model when y is real-valued. We will be concerned both about understanding the structure of the various models to help with their interpretation in a given context as well as their performance. Both **prediction** and **interpretability** will be important and in every case careful attention needs to be given to any proposed fitted model's **generalizability**.

The approach will be applied and so will use formal mathematics and algorithms in the context of actual data analysis. To that end, we will also be using modern interactive graphical tools now available in `R` both for analysis of data and to develop an understanding of the methodology. Expect to carry out analyses and to program in `R`.