3 Looking at Data — Relationships

3.1 Scatterplots

Response Variable:
outcome of a study or an experiment

Explanatory Variable:
has an impact on the outcome; it may be given and not modifiable by humans (e.g., temperature and precipitation in August 2002) or it may be controlled in an experiment (e.g., amount of fertilizer used)

Note:
there may be multiple response variables and multiple explanatory variables!

Scatterplots:
If we have n **bivariate** measurements $(x_i, y_i),\ i = 1, \ldots, n$, we can draw a **scatterplot** of y vs x where each individual pair (x_i, y_i) is plotted as a (colored) symbol within a rectangular coordinate system.

In case of a **dependent (response) variable** and an **independent (explanatory) variable**, we plot the dependent variable on the vertical (y) axis and the independent variable on the horizontal (x) axis. We typically want to predict a y–value for a given x–value.

Example: "The Cost of Victories"

I have randomly drawn a random sample of $n = 8$ teams. The main reason why we want to work with a reduced data set is to reduce the amount of work for our hand calculations and hand drawings. Instead of working with these 8 teams, we could do exactly the same things with all 30 teams. Here are the numbers for our sample:

Team	Wins	Payroll (rounded to nearest Million)
Arizona	8	31
Toronto	17	49
Cleveland	20	56
Florida	14	33
Texas	24	55
Atlanta	29	60
San Diego	24	45
Cincinnati	19	22

We want to predict the number of wins for a given payroll. We first look at the **scatterplot** of the data (i.e., for these 8 teams) and will see later in this Section how to make the desired prediction.

Scatterplot:

Patterns in Scatterplots:

Scatterplots give us information on the form, direction, and strength of the relationship between two variables.

Form of the relationship:

Direction of a *linear* relationship:

Strength of the relationship:

Use of Scatterplots: (not in Textbook)

(i) Scatterplots are widely used in everydays statistical applications. They are useful to detect the type of a relationship between x and y values (e.g., linear, quadratic, exponential, logarithmic, etc. — or no relationship at all).
Scatterplots also allow to detect **clusters** in the data. A cluster is a collection of nearby points that is well separated from the remaining points. Often, we are interested to find one or multiple clusters in our data set.
Finally, scatterplots are very useful to detect **outliers**.

(ii) We can look at one scatterplot, e.g., x vs y, at a time or at multiple scatterplots, e.g., x vs y and y vs z, at a time.
But we can also look at all possible pairs of scatterplots, e.g., x_1 vs x_2, x_1 vs x_3, x_2 vs x_3, x_2 vs x_1, x_3 vs x_1, and x_3 vs x_2, at a time — in this case, we speak of a **scatterplot matrix**.

(iii) When we speak of **(scatterplot) brushing**, we mean that we are marking a subset of points with a different color and/or different symbol.

(iv) When we speak of **linked (scatterplot) brushing**, we mean that we are marking all related points in different scatterplots (or even in a scatterplot matrix) with the same color and/or symbol. This is very helpful to find relationships in multivariate data sets (i.e., ≥ 2 variables).

(v) See the Monmonier article for more details on **geographic brushing**. In short, geographic brushing means that we brush data in a geographic context.

(vi) GGobi is a software tool that allows for linked scatterplot brushing and contains some geographically inspired data sets. GGobi is freely available at `http://ggobi.org/`.

Arrangement of Variables in Scatterplot Matrix in Monmonier Article:

Note: Straight Lines and the SS( ) Notation

Recall:

The equation

$$y = b_0 + b_1 x$$

represents a **straight line**. The **slope** is b_1 and the **y–intercept** is b_0.

Example:

$y = 1 + 2x$

The SS( ) Notation:

We have already defined earlier what $SS(x)$ stands for:

$$SS(x) = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

Similarly, we can define $SS(y)$ and $SS(xy)$:

$$SS(y) = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

and

$$SS(xy) = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$