**CORRELATION**

Two variables \((X,Y)\) are positively associated if when \(X\) is larger than its average, there is a tendency for \(Y\) to be larger than its average. Strong associations lead naturally to the idea of prediction: if \(Y\) is strongly associated with \(X\), then knowing only \(X\), I should have a pretty good idea about \(Y\).

Discovering, interpreting and exploiting associations is a major goal of research in almost any subject. Hence, quantification and verification of association is a major topic in statistics.

The *scatterplot* is a basic graphical methods for exploring the association between two variables. E.g. A scatterplot of heights of father and their adult sons – from a study by Karl Pearson & A. Lee (1903).

![Scatterplot of heights of father and their adult sons](image)

However, numerical measures of association are vital also. Why? The fact is that, while scatterplots convey a lot of *qualitative* information, our human visual perception system can on occasion be tricked.
The basic quantitative measure is the Pearson (product moment) correlation coefficient:

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

(Varying correlations):

- \( r = 0 \)
- \( r = 0.9 \)
- \( r = -0.7 \)
- \( r = -0.3 \)

Notes:

- \( x_i \) and \( y_i \) are standardized - no units - correlation is a dimensionless quantity.
- \( r > 0 \leftrightarrow \) positive association
• $r < 0 \leftrightarrow$ negative
• $-1 \leq r \leq 1$ (Cauchy-Schwartz inequality)
• $r = \pm 1 \leftrightarrow$ points lie exactly on a line: perfect association
• Some software uses $1/n$ instead of $1/(n-1)$. Don’t worry about it except if $n$ is less than 10, then use $(n-1)$.

Football shaped (i.e. bivariate normal) scatter plots can be summarized by 5 numbers:

Means: $(\bar{x}, \bar{y})$ (68,69)
SD’s: $(s_x, s_y)$ (3,3)
Correlation $r$ 0.5

Measure only linear associations, not association in general

\begin{itemize}
  \item \textbf{$r = 0.025$}
  \item \textbf{$r = 0.009$}
  \item \textbf{$r = -0.014$}
\end{itemize}

\textbf{NOT resistant to outliers}
No distinction between “explanatory” and “response”: Consider e.g. the correlation between daily temperatures at Stanford and in Berkeley. This doesn’t depend on whether we call Stanford X and Berkeley Y, or if we go at it the other way round, with Stanford Y and Berkeley X.

Both variables must be quantitative (different methods for categorical data)

Unaffected by linear changes in units: \( \tilde{x} = cx + d, \tilde{y} = ey + f, c, e > 0. \) Add 7 to each, multiply each \( y_i \) by 3 etc -scores are standardized.

Consider again the correlation between daily temperatures at Stanford and at Berkeley: This shouldn’t depend on whether we measure in Fahrenheit or Celsius, and indeed it doesn’t.
Recall: \( ^\circ F = \left(9/5\right) ^\circ C + 32. \)

Question: what if we multiply all the values of \( y \) by \(-1\)?

Correlations depend on the data used.

Depending on how we summarize/select from the height data, we can generate three wildly different sample correlations:

a) all data \( r = 0.5 \)
b) Fathers between 65 and 67 inches: 0.11
c) Averages over groups \( r = 0.99 \) correlation of averages effect

Effects not related to sample size \( n \):

Restriction of range reduces the variation in \( x \), and hence the variation in \( y \). If the variation in \( y \) is no larger than the variation in \( y \) given \( x \), then there is no predictive power left in the restricted range.

Correlations of averages are typically higher than correlations of individuals, and can give a misleading impression about the strength of an association at the individual level.


between individuals: \( r = 0.44 \)

compute average income and average education in each state

between state averages: \( r = 0.64 \)

The variation within each state is removed by taking averages and gives a misleading picture of the relationship for individuals, which is our primary interest.

Such correlations based on rates or averages (rather than individuals) are sometimes called ecological correlations.
LINEAR REGRESSION – Fitting a straight line to a scatterplot. Straight line notation:

\[ y = a + bx \]

- \(a\) = intercept
- \(b\) = slope \(= \frac{\Delta y}{\Delta x}\)

An interesting historical dichotomy that highlights two approaches to straight line fitting:

- Physical sciences – (typically) experimental data – least squares line
- Social sciences – observational data – regression

Formulas and some of the interpretations are the same.

The experimental science approach to fitting lines – the least squares idea – is conceptually simpler, and indeed came historically earlier – around 1805, so we start with that. Later we look at the regression idea, which in the case of bivariate normal data, is intimately related to the least squares approach.

**Least squares regression**: Fit a line to \((x_i, y_i)\) \(i = 1, \ldots, n\)

Minimize sum of vertical distances

\[ e_i = y_i - a - bx_i, \quad \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \]

(by calculus) minimize over \(a\) and \(b\).

The resulting fitted LS regression line has as ingredients:

- Means: \((\bar{x}, \bar{y})\)
- SD’s: \((s_x, s_y)\)
- Correlation \(r\)

Resulting equations:

\[ \hat{y} = a + bx \]

\[ b = r \frac{s_y}{s_x} \]

\[ a = \bar{y} - b\bar{x} \]

\(\hat{y}\) or \(\hat{y}(x)\) denotes ‘fitted’ or ‘predicted’ value of \(y\) given \(x\).

See “Derivation of Least Squares Regression Line” for details. Appendix 12.1 has a derivation not using calculus.
Another way to write the regression line (just substitute definitions of $a$ and $b$):

$$\hat{y} - \bar{y} = r \times \left( \frac{x - \bar{x}}{s_x} \right)$$

Change of 1 SD $x \leftrightarrow r$ SD in $y$

- Change of 1 SD $x \leftrightarrow r$SD in $y$ Note: expressed in standardized units.
- Line passes through $(\bar{x}, \bar{y})$

2. Residual Plots, Outliers
Residual = observed $y$ - predicted $y$: $e = y - \hat{y}$
Diagnostic plots (to check linear relation): residuals vs * explanatory variable $x$ * time(sudden change point) Can indicate

- non linear association of $y$ with $x$
- variation of $y$ increases/decreases with $x$
- may suggest a transformation of $y$: e.g. log($y$)

Outlier(s): lies outside overall pattern of the data
Outliers in $y$ - typically have large residuals
Outliers in $x$ – need not.

Influential observation: If removed, regression line would change markedly.

Extrapolation
Use of a regression line for prediction outside the range of the original $x$-values.
Dangerous, because assumes linear relation valid in a region where there is no data.
Example: Whipp & Ward, *Nature* Jan 1992 World record times for men & women over 100 years Data for 200m, 400m, 800, 1500m and the Marathon, claimed to find broadly consistent extrapolations.

Two difficulties with extrapolation:
(i) the underlying linear relations may not be correct beyond the range of the data,
(ii) your predictions can be sometimes checked and shown to be wrong!
E.g. Notice that the extrapolation for the marathon crosses in 1998. The authors say “Beyond that time current progression rates imply superior performance by women”.

![Graphs showing regression lines and residual plots](image-url)
Men 2:05:42 Khalid Khannouchi MOR Chicago 24.10.99
Women 2:18.47 Catherine Ndereba KEN Chicago 7.10.01

Further mystery: why did the Nature authors omit 100m data? It was surely available.

From another source: Olympic gold medal times for 100M for both women and men. Naïve extrapolation puts intersection out beyond 2100. Perhaps, because of the longer intersection time, it didn’t really support their hypothesis.

**AVERAGES IN VERTICAL STRIPS**
From our previous example of the heights of fathers and sons

Here we are interested in height and weight data on 988 men aged 18-24 from the HANES survey X = height, Y = weight, $\bar{x} = 67.69$ in, $\bar{y} = 68.68$ in, $s_x = 2.74$ in, $s_y = 2.81$ in, $r = 0.5$

Vertical strip: Fathers of 71 in [i.e. 1 SD above average, to nearest inch]

Dotted line: *SD-line* Line through $(\bar{x}, \bar{y})$ that goes up one SD_y for every SD_x it goes along. (If $r = 1$, all points would lie on this line). *Most points in the 73” strip are below the SD line.* It is surprising, but true, that *the average in the strip comes from the regression line:* i.e. Since $x$ is 1 SD above its mean, yhat is 0.47 times 1 SD above its mean.

*On average, each SD increase in x goes with only r SDs increase in y* [the proof comes from the mathematics of the bivariate normal distribution, e.g. Rice, *Mathematical Statistics and Data Analysis* p. 474 and Sec 4.4.1. Math + Bschool Libs]

Note: If you switch the axis, the regression line will look differently.
Graph of averages: (conditional expectations) average son's height at each father's height (in inches). Close to regression line when many sons are averaged together.

Deviations of the averages from the regression line are due to chance variation

The regression line \( \hat{y} = a + bx \) smoothes out the graph of averages.

Graph of averages justification → same answer as LS method!

Regression Effect and Regression Fallacy

A pre-school program attempts to boost children’s IQs. The children are tested when they enter and when they leave. In both tests the scores average out to nearly 100, and SD is about 15. But children who were below average on pretest had average gain of about 5 IQ points, and those above average had average loss of about 5 IQ points.

\[
\begin{align*}
\text{observed} & = \text{true} + \text{chance error} \\
\text{score on test 1} & = \text{real IQ} + \left\{ \begin{array}{c} +5 \text{ (good day)} \\
-5 \text{ (bad day)} \end{array} \right. \\
140 & = 135 + 5 \\
140 & = 145 - 5
\end{align*}
\]

So an observed score of 140 could come from someone with a true score of 135 or 145. Why aren’t the latter both equally likely, if the chance error is half good and half bad?

Because if we suppose that the subjects true IQ is Normally distributed with 100 as its average and 15 as its standard deviation an IQ of 135 will be more likely than an IQ of 145.

The REGRESSION FALLACY consists in thinking that the regression effect must be due to something important, not just the spread around the line. "On test retest situations the reason the retest scores tend to regress towards the mean is due to chance error. If someone scores very high on test 1 it is assumed that it was partly due to luck. On the retest they may not be so lucky so their scores drop on average. The converse is true on very low scores on the first test."

So chance error could be good luck, bad luck, faulty scoring, faulty measuring, mistakes in writing results, mistakes in reading results, anything.

RMS Error

The regression line predicts the average y value associated with a given x value. Note that is also necessary to get a measure of the spread of the y values around that average. To do this, we use the root-mean-square error (r.m.s. error).

To construct the r.m.s. error, you first need to determine the residuals. Residuals are the difference between the actual values and the predicted values. I denoted them by \( \hat{y}_i - y_i \), where \( y_i \) is the observed value for the ith observation and \( \hat{y}_i \) is the predicted value.

They can be positive or negative as the predicted value under or over estimates the actual value. Squaring the residuals, averaging the squares, and taking the square root gives us the r.m.s error. You then use the r.m.s. error as a measure of the spread of the y values about the predicted y value.

\[
\text{RMS Errors} = s_{y|x} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-2}} = \sqrt{\frac{\text{SS(resid)}}{n-2}}
\]
As before, you can usually expect 68% of the y values to be within one r.m.s. error, and 95% to be within two r.m.s. errors of the predicted values. These approximations assume that the data set is football-shaped.

Squaring the residuals, taking the average then the root to compute the r.m.s. error is a lot of work. Fortunately, algebra provides us with a shortcut (whose mechanics we will omit).

The r.m.s error is also equal to $\sqrt{1 - r^2}$ times the SD of y.

$$RMSE = \sqrt{1 - r^2 SD_y}$$

Thus the RMS error is measured on the same scale, with the same units as y. The term $\sqrt{1 - r^2}$ is always between 0 and 1, since r is between -1 and 1. It tells us how much smaller the r.m.s error will be than the SD.

For example, if all the points lie exactly on a line with positive slope, then r will be 1, and the r.m.s. error will be 0. This means there is no spread in the values of y around the regression line (which you already knew since they all lie on a line).

The residuals can also be used to provide graphical information. If you plot the residuals against the x variable, you expect to see no pattern. If you do see a pattern, it is an indication that there is a problem with using a line to approximate this data set.

To use the normal approximation in a vertical slice, consider the points in the slice to be a new group of Y’s. Their average value is the predicted value from the regression line, and their spread or SD is the r.m.s. error from the regression.