MATH60604A
Statistical modelling
§2e - Coefficient of determination

HEC Montréal
Department of Decision Sciences

Pearson's linear correlation coefficient

- The correlation coefficient quantifies the strength of the linear relationship between two random variables X and Y.
- Suppose that we're studying n pairs of observations $(X_1, Y_1), \dots, (X_n, Y_n)$, where (X_i, Y_i) are the values of X and Y for individual i.
- Pearson's correlation coefficient is

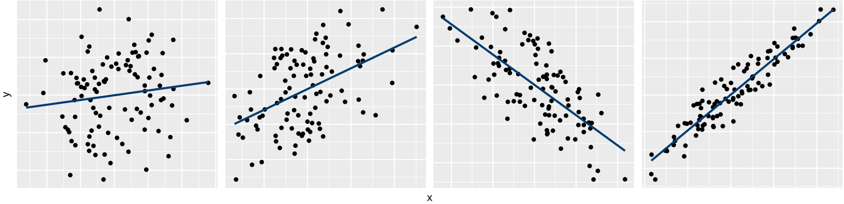
$$R = \frac{\widehat{\mathrm{Co}}(X, Y)}{\sqrt{\widehat{\mathrm{Va}}(X)\widehat{\mathrm{Va}}(Y)}} = \frac{\sum_{i=1}^{n}(\mathrm{X}_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}},$$

where $\overline{X}$ and $\overline{Y}$ are the sample means of X and Y.

Properties of Pearson's linear correlation coefficient

Properties of Pearson's linear correlation coefficient

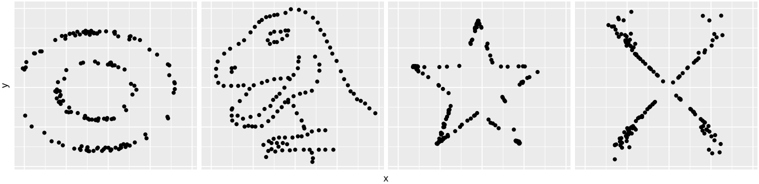
- $-1 \leq r \leq 1$
- $r = 1$ ($r = -1$) if and only if the n observations fall exactly on a positively (negatively) sloped line. In other words, there exist two constants a and $b > 0$ $(b < 0)$ such that $y_i = a + bx_i$ for any i.



From left to right, the fours samples have linear correlation 0.1, 0.5, -0.75 and 0.95.

Pearson's linear correlation coefficient

- If $r > 0$ ($r < 0$), the two variables are positively (negatively) associated, meaning that Y increases (decreases) on average with X.
- The larger $|r|$, the less scattered the points are.
- Independent variables are uncorrelated (not the other way around).
- A correlation of zero does not imply that there is no relationship between the two variables. It only means that there is no **linear** dependence between the two variables.



The four datasets (bullseye, Anscombosaurus, star, cross) have the same correlation of -0.06, yet the variables are clearly not independent.

Coefficient of determination

- Once the model has been fitted, it is be useful to have a measure that will tell us whether the model fits the data well.
- The coefficient of determination, R^2, measures the strength of the linear relationship between $\widehat{Y}$ and Y.
- It is interpreted as the proportion of the variation in Y explained by the $\mathbf{X}$'s.
- R^2 is the squared correlation between the predicted values and the response, $(\widehat{Y}_1, Y_1), \ldots, (\widehat{Y}_n, Y_n)$.

Sum of squares decomposition

- Suppose that we do not use any explanatory variable (i.e., the intercept-only model). In this case, the fitted value for Y is the overall mean and the sum of squared centered observations

$$\mathrm{SS}_c = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

- When we include $\mathbf{X}$, the fitted value of Y_i is rather $\widehat{Y}_i = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_1 \mathrm{X}_{ij}$ and the sum of the squared residuals is

$$\mathrm{SS}_e = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

- The SS_e is non-increasing when we include more variables.

Coefficient of determination (R^2)

- R^2 measures the proportion of the variance in Y explained by the set of predictor variables $X_1, \dots, X_p$,

$$R^2 = \frac{SS_c - SS_e}{SS_c}.$$

- When there is more than one explanatory variable, the square root of R^2 is also called the multiple correlation coefficient.
- R^2 always takes a value between 0 and 1.

Coefficient of determination and interpretation

R-Square	Coeff Var	Root MSE	intention Mean
0.449726	27.41959	2.264401	8.258333

- In the model with all of the explanatory variables, $R^2 = 0.45$. Together, the explanatories explain 45% of the variability in `intention`.
- For the simple linear model with only `fixation` as covariate, $R^2 = 0.182$. That means the variable `fixation` explains 18.2% of the variability in `intention`.

A word of caution regarding R^2

- **Warning**: the more regressors you include in your model, the higher the R^2 (regardless of whether these variables are useful from an inference or predictive perspective).
- R^2 is therefore not a goodness-of-fit criterion.
- Software sometimes report the adjusted R^2, which includes a penalty,

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}.$$

The coefficient loses its interpretability and can be negative.