Artificial Intelligence, Blockchain, e Criptovalute nello Sviluppo Software

Lezioni 13 e 14: Inferences, Non Parametric Approaches, and Logistic Regression

Giancarlo Succi
Dipartimento di Informatica – Scienza e Ingegneria
Università di Bologna
g.succi@unibo.it

Content

- More on the correlation coefficient
- Non parametric correlations
- Logistic regression

More on the correlation coefficient

Status

- Now we know that the means of samples of a population tend to be distributed normally.
- This is an essential assumption to perform several numeric operations, like Montecarlo simulations, Bootstrap, etc.
- We would like now to understand the distribution of the Pearson momentum correlation coefficient of the sample
- Moreover, we have an open infinite issue on what to do if the data is NOT on a ratio scale

Modeling with linear models (1/2)

Linear regression is dependent on 4 hypothesis:

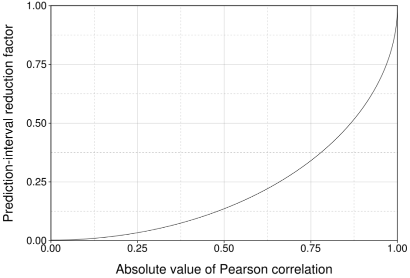
1. Normality
   The dependent variable is normally distributed at each value of the independent variables.
   How to check: histogram of standardized residuals, Q-Q plot
2. Homoscedasticity
   The variability of the standardized residuals is constant and does not depend on dependent variable.
   How to check: plotting the residuals over the mean value of dependent variable

Modeling with linear models (2/2)

1. Independence of error
   Each value of the residual does not depend in some way from the preceding value.
   How to check: Durbin-Watson statistic
2. Linearity
   There is linear dependency between regressors and response
   How to check: linear correlation coefficient

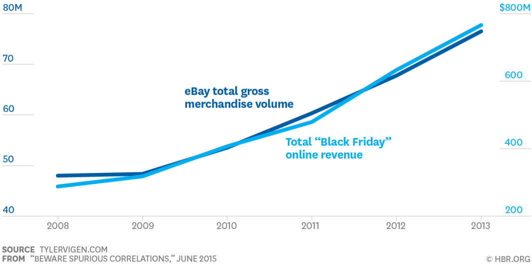
Is the correlation enough for predicting?

- The size of an acceptable correlation depends on the context
- A key question is what is the additional explanation that I get from analysing X vs just using Y
- The following diagram for instance shows how the 95% confidence interval is reduced for increasing values of the correlation



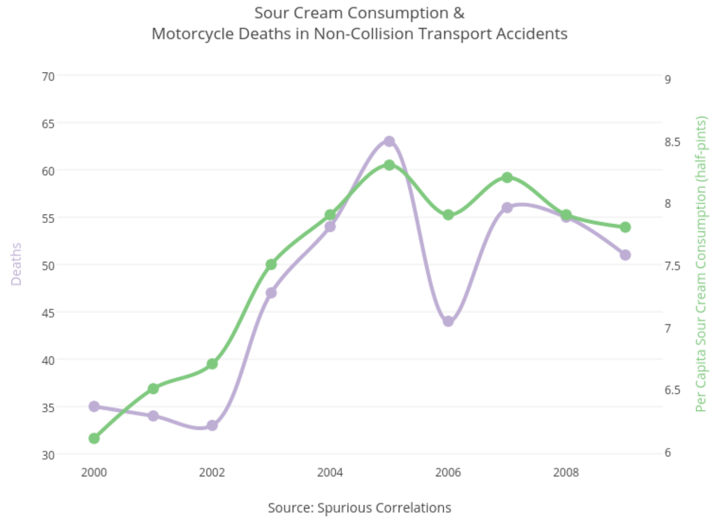
Source with modifications: `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient`

Comparing "Apples and Oranges"
Y axis scales that measure different values may show similar curves that shouldn't be paired. This becomes pernicious when the values appear to be related but aren't.
Example.

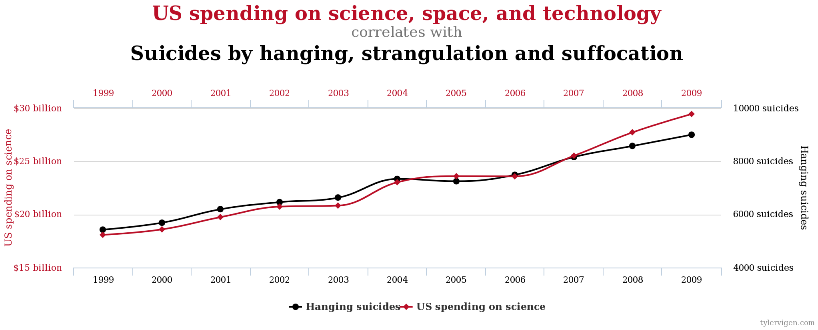


Correlation does not imply causation.

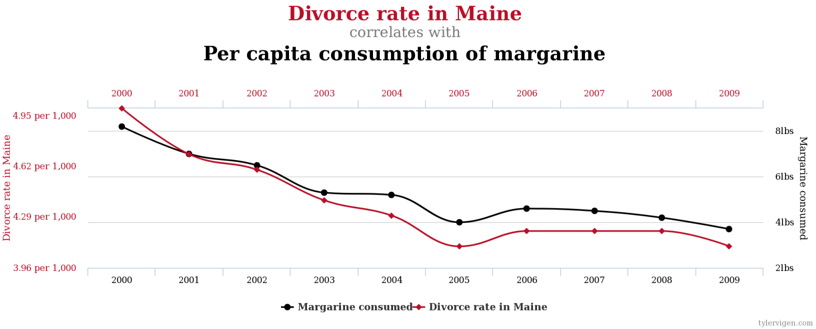
Spurious correlations



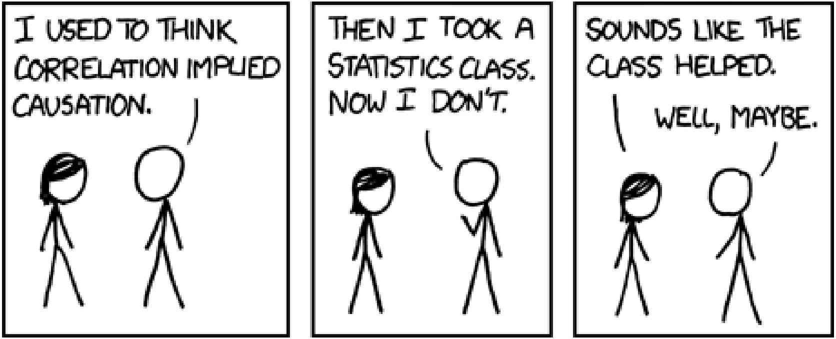
Spurious correlations



Spurious correlations



Spurious correlations



Distribution of r_{XY} of the sample

- Suppose, as usual, that we have two phenomena that we want to measure, X and Y and let us assume:
  - that there is a linear relationship between them
  - that I can express the data I collect as:

    $$\hat{Y} = \theta_0 + \theta_1 X + \epsilon$$

    where ϵ is a stationary Gaussian process $N(0, \sigma^2)$
  - that I have n samples, that is $\mathfrak{n}$ set of random pairs $\mathfrak{S}_\mathfrak{j} = \{(\mathfrak{X}_{\mathfrak{i}_\mathfrak{j}}, \mathfrak{Y}_{\mathfrak{i}_\mathfrak{j}})\}$, with:
    - $\mathfrak{j} \in [1 \ldots \mathfrak{n}]$,
    - $\mathfrak{i}_\mathfrak{j} \in [1 \ldots \mathfrak{m}_\mathfrak{j}]$,
    - $(\forall\ \mathfrak{j})\ \ \mathfrak{m}_\mathfrak{j} \in \mathbb{N}^+$
  - for each $\mathfrak{S}_\mathfrak{j}$ I can compute the Pearson correlation coefficient $\mathfrak{r}_{\mathfrak{X}_\mathfrak{j}\mathfrak{Y}_\mathfrak{j}}$
- What is the distribution of $\mathfrak{r}_{\mathfrak{X}_\mathfrak{j}\mathfrak{Y}_\mathfrak{j}}$?

The Student t (1/3)

- Used to determine the distribution of $\frac{\mathfrak{D}\mathfrak{n}}{\mathfrak{s}_\mathfrak{n}}$ – We know that $\frac{\mathfrak{D}\mathfrak{n}}{\sigma} \xrightarrow{d} N(0,1)$
- Apparently, started in the brewery of Guinness
- The pdf is:

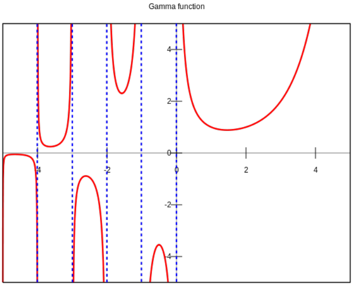
$$f_x(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- We use the Γ function

Source with modifications: `https://en.wikipedia.org/wiki/Student%27s_t-distribution`

- The Γ function, an extension of the factorial to the whole $\mathbb{C}$ set apart from negative integers, that it, it is defined on $(\mathbb{R} - \mathbb{N}^{-}, \mathbb{R})$
- Formally:

$$\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} dx$$



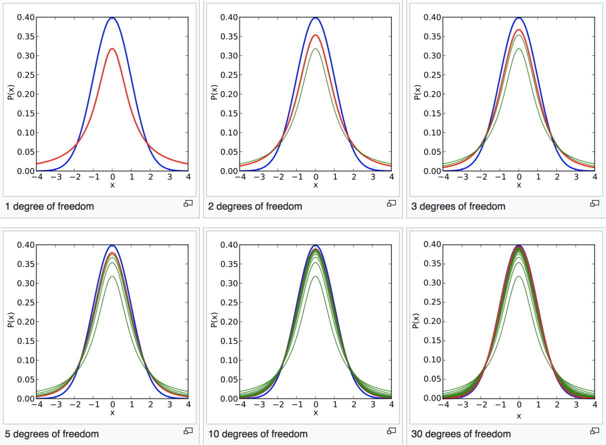
- Recall the pdf:

$$f_x(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- The Student t is symmetric
- ν is the degree of freedom, as it increases the function becomes similar to the Gaussian (in the figure in Slide 18 the t is in red, the Gaussian is in blue, and the previous ts are in green)

Source with modifications: *https://en.wikipedia.org/wiki/Student%27s_t-distribution*

The Student t (3/3)



Source with modifications: *https://en.wikipedia.org/wiki/Student%27s_t-distribution*

Distribution of $\mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j}$ (1/2)

- The $\mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j}$ are approximated by a Student t distribution with $(n-2)$ degrees of freedom, under "good" assumptions
- Under such assumptions **and** the one that we have mentioned before, we have:

$$t = \mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j}\sqrt{\frac{n-2}{1-{\mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j}}^2}}$$

- and conversely:

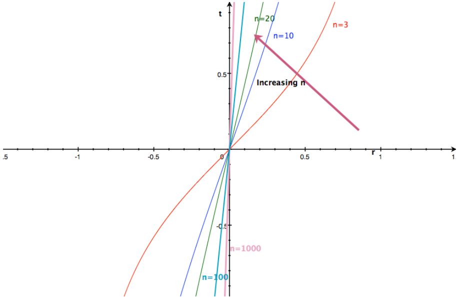
$$\mathfrak{r}_{\mathfrak{X}_j\mathfrak{Y}_j} = \frac{t}{\sqrt{n-2+t^2}}$$

Source with modifications: *https://en.wikipedia.org/wiki/Pearson_correlation_coefficient*

- If we concentrate on:

$$t = \mathfrak{r}_{\mathfrak{X}_j \mathfrak{Y}_j} \sqrt{\frac{n-2}{1 - {\mathfrak{r}_{\mathfrak{X}_j \mathfrak{Y}_j}}^2}}$$

- we notice that for the same value of $\mathfrak{r}_{\mathfrak{X}_j \mathfrak{Y}_j}$ we obtain higher values of t, with increasing values of n



Distribution of $\mathfrak{r}_{\mathfrak{X}_j \mathfrak{Y}_j}$

Claim:

$$t_{n-2} \sim \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Assamptions and Facts:

- $Y = \beta_0 + \beta_1 X + \epsilon.$
- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
- Where $\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R} \setminus \{0\}$ and $\epsilon \sim N(0, \sigma^2)$.

Plan:

- $\hat{\beta}_1 \sim \mathcal{N}$
- $RSS \sim \chi^2$
- $t \sim \frac{\hat{\beta}_1}{RSS}$

Given model $y = \beta_0 + \beta_1 x_i + \epsilon$, β_1 estimator $(\hat{\beta}_1)$ can be derived as follows :

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$$

Remember :

- $s_{xx} = \sum (x_i - \bar{x})^2$
- $s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \sum \frac{(x_i - \bar{x})}{s_{xx}} (y_i - \bar{y}) = \sum \frac{(x_i - \bar{x})}{s_{xx}} y_i - \sum \frac{(x_i - \bar{x})}{s_{xx}} \bar{y}$$

Taken with modifications from https://math.stackexchange.com/questions/787939/show-that-the-least-squares-estimator-of-the-slope-is-an-unbiased-estimator-of-t/788010#788010

Therefore :

$$\begin{aligned}
\hat{\beta}_1 &= \frac{s_{xy}}{s_{xx}} = \sum \frac{(x_i - \bar{x})}{s_{xx}} (y_i - \bar{y}) = \sum \frac{(x_i - \bar{x})}{s_{xx}} y_i - 0 * \bar{y} \\
&= \sum \frac{(x_i - \bar{x})}{s_{xx}} y_i = \sum \frac{(x_i - \bar{x})}{s_{xx}} (\beta_0 + \beta_1 x_i + \epsilon_i) \\
&= \beta_0 \sum \frac{(x_i - \bar{x})}{s_{xx}} + \beta_1 \sum \frac{(x_i - \bar{x})}{s_{xx}} x_i + \sum \frac{(x_i - \bar{x})}{s_{xx}} \epsilon_i
\end{aligned}$$

Note 1:

$$\sum \frac{(x_i - \bar{x})}{s_{xx}} = \frac{1}{s_{xx}} \sum (x_i - \bar{x}) = \frac{1}{s_{xx}} \left(\sum x_i - \sum \bar{x} \right) = 0$$

Taken with modifications from `https://math.stackexchange.com/questions/787939/show-that-the-least-squares-estimator-of-the-slope-is-an-unbiased-estimator-of-t/788010#788010`

Note 2:

$$\sum \frac{(x_i - \bar{x})}{s_{xx}} x_i = \sum \frac{(x_i - \bar{x})}{s_{xx}} (x_i - \bar{x} + \bar{x})$$
$$= \sum \frac{(x_i - \bar{x})}{s_{xx}} (x_i - \bar{x}) + \sum \frac{(x_i - \bar{x})}{s_{xx}} \bar{x}$$
$$= \frac{1}{s_{xx}} \sum (x_i - \bar{x})^2 = 1$$

Putting the simplifications into the original equation:

$$\hat{\beta}_1 = 0 \times \beta_0 + 1 \times \beta_1 + \sum \frac{(x_i - \bar{x})}{s_{xx}} \epsilon_i = \beta_1 + \sum \frac{(x_i - \bar{x})}{s_{xx}} \epsilon_i$$

- Remember that:
  - $\sum_i \frac{(x_i-\bar{x})^2}{s_{xx}} = 1$
  - $\hat{\beta}_1 = \beta_1 + \sum_i \frac{(x_i-\bar{x})}{s_{xx}}\epsilon_i$
  - $\forall i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$
  - $\forall \lambda \neq 0, \quad \lambda\mathcal{N}\left(\mu, \sigma^2\right) = \mathcal{N}\left(\lambda\mu, (\lambda\sigma)^2\right)$
  - $\sum_i \mathcal{N}\left(\mu_i, \sigma_i^2\right) = \mathcal{N}\left(\sum_i \mu_i, \sum_i \sigma_i^2\right)$

- Therefore:
  - $\sum \frac{(x_i-\bar{x})}{s_{xx}}\epsilon_i \sim \mathcal{N}\left(0, \sum_i \left(\frac{\sigma(x_i-\bar{x})}{s_{xx}}\right)^2\right) =$
    $= \mathcal{N}\left(0, \sigma^2 \sum_i \left(\frac{x_i-\bar{x}}{s_{xx}}\right)^2\right) = \mathcal{N}\left(0, \frac{\sigma^2}{s_{xx}} \sum_i \frac{(x_i-\bar{x})^2}{s_{xx}}\right) =$
    $\mathcal{N}\left(0, \frac{\sigma^2}{s_{xx}}\right)$
  - $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right)$
  - $\frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{s_{xx}}} \sim \mathcal{N}(0,1)$

- Note the following:
  - We know that t_n can be rewritten using a χ^2_n distribution:

$$t_n \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_n/n}}$$

  - And also we can connect RSS, σ, and χ^2:

$$\frac{RSS}{\sigma^2} = \frac{\sum \epsilon_i^2}{\sigma^2} = \sum \left(\frac{\epsilon_i}{\sigma}\right)^2 = \sum \left(\mathcal{N}(0,1)\right)^2 \sim \chi^2_{n-2}$$

Taken with modifications from
https://stats.stackexchange.com/questions/204238/why-divide-rss-by-n-2-to-get-rse

- We will now proceed as follows:
  - $\hat{\beta}_1 \sim \mathcal{N}$
  - $RSS \sim \chi^2$
  - $t \sim \frac{\hat{\beta}_1}{RSS}$

Taken with modifications from
https://stats.stackexchange.com/questions/204238/why-divide-rss-by-n-2-to-get-rse

- $r^2 = 1 - \frac{RSS}{SST}$
- $RSS = (1 - r^2) s_{yy}$
- $SST = s_{yy} = \sum (y_i - \bar{y})^2 \frac{RSS}{SST}$
- $r = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2}}$
- $t_{n-2} \sim \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{s_{xx}}}}{\sqrt{RSS/(n-2)\sigma^2}} = \frac{r\sqrt{(n-1)}}{\sqrt{(1-r^2)}}$

Taken with modifications from
https://stats.stackexchange.com/questions/204238/why-divide-rss-by-n-2-to-get-rse

Under null hypothesis $H_0 : \beta_1 = 0$

$$\frac{\frac{\hat{\beta}_1 - 0}{\sigma/\sqrt{s_{xx}}}}{\sqrt{\frac{RSS}{\sigma^2(n-2)}}} = \frac{\hat{\beta}_1\sqrt{s_{xx}}}{\sigma} \cdot \frac{\sqrt{\sigma^2(n-2)}}{\sqrt{RSS}} = \frac{\hat{\beta}_1\sqrt{s_{xx}}}{1} \cdot \frac{\sqrt{(n-2)}}{\sqrt{(1-r^2)s_{yy}}}$$

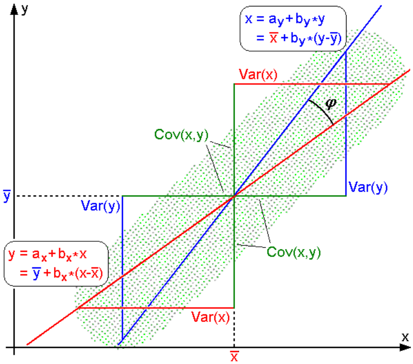
$$= \frac{\hat{\beta}_1\sqrt{s_{xx}}}{\sqrt{s_{yy}}} \cdot \frac{\sqrt{(n-2)}}{\sqrt{(1-r^2)}} = \frac{\frac{s_{xy}}{s_{xx}}\sqrt{s_{xx}}}{\sqrt{s_{yy}}} \cdot \frac{\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

$$= \frac{s_{xy}}{\sqrt{s_{yy}s_{xx}}} \cdot \frac{\sqrt{(n-2)}}{\sqrt{(1-r^2)}} = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

QED

Reasoning on θ_1 (1/2)

$$\theta_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{\sigma_Y}{\sigma_X} r_{X,Y}$$



Source with modifications: *https://en.wikipedia.org/wiki/Pearson_correlation_coefficient*

Reasoning on θ_1 (2/2)

- Remember that:
$$\theta_1 = \frac{\sigma_Y}{\sigma_X} r_{X,Y}$$
- If $r_{X,Y} > 0$, we define the p-value of our correlation the $P(\theta_1 < 0)$, conversely, if $r_{X,Y} < 0$ we define the p-value of our correlation the $P(\theta_1 > 0)$
- In other terms, the p-value of a correlation is the probability that a slope change direction

Source with modifications: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

The power function of a test

Remember that:

- A type 1 error is when we reject the null hypothesis when the null hypothesis is true, that is we think that something is going on, but nothing is really there. The probability of committing a type 1 error is typically referred to as α.
- A type 2 error is when we fail to reject the null hypothesis when actually we should reject it, that is, we fail to perceive a phenomena. The probability of committing a type 2 error is typically referred to as β.

The power function of a test informally is the probability of not committing a type 2 error, that is, $(1 - \beta)$

Power of a test

- In general the non rejection of the null hypothesis H0 does not mean that H0 holds
- The power of a binary test is the probability that the tests rejects the null hypothesis when the alternate hypothesis is true

$$\text{Power(Test)} = \text{P(reject H0} \mid \text{ H1 is true)}$$

- If a test has power of 0.99 in a given situation, it means that the non rejections of H0 means that H0 holds with a p(error) ≤ 0.01
- The power of a test has an essential role in determining the test to select and in interpreting its results

Taken with modifications from https://en.wikipedia.org/wiki/Power_(statistics)

What influences the power of a test

The power of a test is influenced by a variety of factors, such as:

- the size of the datasets
- the magnitude of the effect
- the level of statistical significance
- the intrinsic structure of a test
  - we can use a test only if its hypotheses are all verified
  - informally, the more stringent the hypotheses, the higher the power of the test, since ...
  - ... *we know better the population*

Taken with modifications from `https://en.wikipedia.org/wiki/Power_(statistics)`

Parametric and non parametric tests

We can distinguish two major classes of tests:

- When we **can make assumptions** on the distributions of the two datasets;
  - for this case, we have the *parametric* tests, since we can assume parameters of the underlying distribution
- When we **cannot**
  - for this case, we have the *non parametric* tests, since we cannot make any assumption on any kind of parameter of the underlying distribution

Taken with modifications from `https://en.wikipedia.org/wiki/Power_(statistics)`

The problem of multiple testing

Consider a case where you have **20** hypotheses to test, and a significance level of **0.05**.

The probability of observing at least one significant result just due to chance?

$$\mathbb{P}(at\_least\_1\_signif.\_results) = 1 - \mathbb{P}(no\_signif.\_results) =$$
$$= 1 - (1 - 0.05)^{20} \approx 0.64$$

The Bonferroni correction

So, with 20 tests being considered, we have a **64%** chance of observing at least one significant result, even if all of the tests are actually not significant.

The Bonferroni correction sets the significance cut-off at α/n.

For example, with **20** tests and $\alpha = 0.05$, you'd only reject a null hypothesis if the p-value is less than **0.0025**.

Toward the Bonferroni inequality (1/2)

Claim (Boole Inequality): Let A_1, A_2, …, A_n be n events, then:

$$P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$$

Proof: *By induction*

Base

For n=1 it is trivially verified.

For n=2:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2)$$

since $P(A_1 \cap A_2) \geq 0$.

Toward the Bonferroni inequality (2/2)

Step

Assuming it true for $n \geq 2$, we prove it for $n+1$.

Given the associativity of the $\cup$:

$$P(\bigcup_{i=1}^{n+1} A_i) = P(\bigcup_{i=1}^{n} A_i \cup A_{n+1})$$

Calling B the set $\bigcup_{i=1}^{n} A_i$ and C the set A_{n+1} we can write:

$$P(B \cup C) = P(B) + P(C) - P(B \cap C) \leq P(B) + P(C)$$

Which means:

$$P(\bigcup_{i=1}^{n+1} A_i) \leq P(\bigcup_{i=1}^{n} A_i) + P(A_{n+1}) \leq \sum_{i=1}^{n} P(A_i) + P(A_{n+1}) = \sum_{i=1}^{n+1} P(A_i)$$

QED

The Bonferroni correction - Proof

Claim (Bonferroni correction): In the case of m null hypotheses $H0_1 \cdots H0_m$ sufficient condition to have a probability than a given α of wrongly rejecting a null hypothesis is that $\forall i \in [1 \cdots m], p_i \leq \frac{\alpha}{m}$.

Proof:

From the Boole inequality:

$$P\big(\bigcup_{i=1}^{m}(p_i \leq \frac{\alpha}{m})\big) \leq \sum_{i=1}^{m}\left\{P\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq m\frac{\alpha}{m} = \alpha$$

QED

References

1) http://www.cs.umd.edu/~djacobs/CMSC426/Convolution.pdf

2) https://www.researchgate.net/post/Difference_between_convolution_and_correlation

3) https://www.tutorialspoint.com/signals_and_systems/convolution_and_correlation.htm

Non parametric correlations

Spearman's Rank Correlation Coeff. (1/3)

- What can we do when the data is not normally distributed?
- Or even if the data is not on a ratio scale, just on an ordinal scale?

- *If the data is on a nominal scale, the concept of correlation looses interest; at most we can consider clustering.*

Idea:

- Transform the data into ranks
- Apply the Pearson correlation coefficient to ranks
- Indeed, the values can be different, and also the significance and the mutual relationship
- Remember that:

$$r_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- And also that:

$$\theta_1 = \frac{\sigma_X \sigma_Y}{Var(X)} r_{X,Y}$$

Source with modifications: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Spearman's Rank Correlation Coeff. (3/3)

Definition:

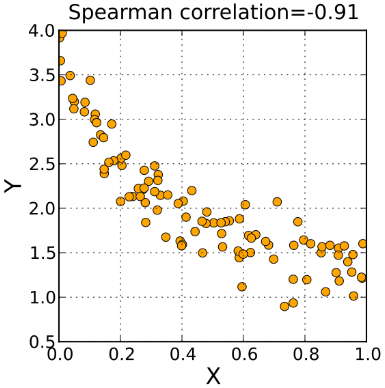
- Let's have two sets $X = \{X_i\}$ and $Y = \{Y_i\}$ of the same size n where $(\forall i) X_i, Y_i \in$ ordinal scale
- Let's consider a set of pairs $P_{X,Y} = \{(X_i, Y_i)\}$
- Let's define
  - $(\forall X_i \in X)\ rk_{X\,i} = \mathrm{rank}(X_i, X)$, $Rk_X = \{rk_{X_i}\}$
  - $(\forall Y_i \in Y)\ rk_{Y_i} = \mathrm{rank}(Y_i, Y)$, $Rk_Y = \{rk_{Y_i}\}$
- We define the Spearman's Rank Correlation Coefficient between X and Y, $r_S(X, Y)$ as:

$$r_S(X,Y) = r(Rk_X, Rk_Y) = \frac{Cov(Rk_X, Rk_Y)}{\sigma_{Rk_X}\sigma_{Rk_Y}}$$

Source with modifications: `https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient`

Visualization of r_S

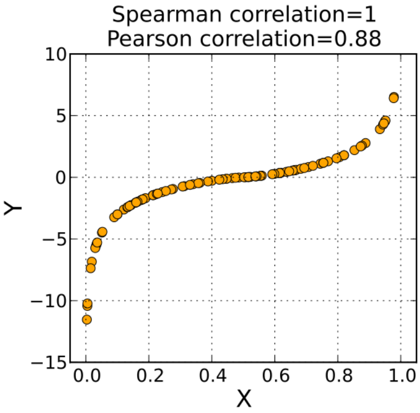
Spearman's Rank Correlation Coefficient is based on monotonicity:

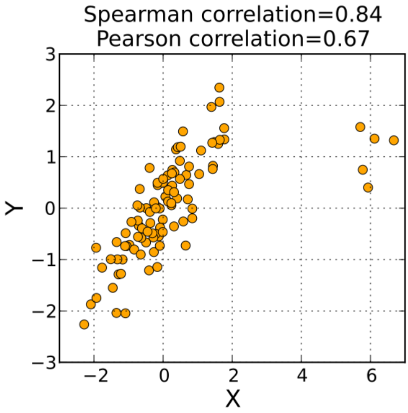


Source with modifications: `https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient`

r and r_S (1/2)

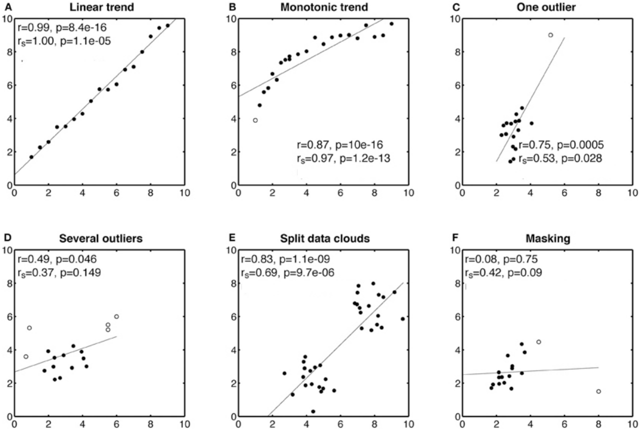
Indeed, the values of r and r_S can be different:





Source with modifications: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

r and r_S (2/2)



Source with modifications: https://www.researchgate.net/figure/Examples-of-Pearson-and-Spearman-correlations-In-each-subplot-r-is-Pearson-correlation_fig7_224915794

Notes about r_S

- If two identical values are assigned their fractional rank
  - So if we have 20, 20, 30, 35, 36, then their ranks should be 1.5 (the average between 1 and 2), 1.5, 3, 4, 5 respectively
- Taking into account that we are dealing with integer ranks, we can simplify the formula as follows if all values are different:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

  - where n is the number of observations and each d_i is equal to the difference in rank between X_i and Y_i:

$$d_i = Rk_{X_i} - Rk_{Y_i}$$

Source with modifications: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Significance of r_S (1/3)

- Being based on ordinals and non assuming anything on the distribution of the underlying populations, the computation of the significance of r_S is based on permutations
- This belong to the family of permutation tests
  - A permutation test (or exact test) is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points
- In our case, since I have sequence of ordinals, we can consider all possible pairs of mutual relationships and, based on this, determine if the monotonic relationship that we have obtained is significantly different from a random order

Source with modifications: https://en.wikipedia.org/wiki/Resampling_(statistics)#Permutation_tests

Significance of r_S (2/3)

- Consider as an example the dataset $\{(X_i, Y_i)\} = \{(10, 2), (15, 0), (20, 4), (21, 50)\}$
- Does it have a significant positive correlation?
- We need to assign ranks the elements, leading to $\{(Rk_{X_i}, Rk_{Y_i})\} = \{(1, 2), (2, 1), (3, 3), (4, 4)\}$
- This leads to $r_S = 0.8$
- To compute the significance, I determine the number of times the comparison $Rk_{Y_i} \leq Rk_{Y_j}$ are true when $i < j$
- These are sequences of Bernoulli trials ...

Source with modifications: `https://en.wikipedia.org/wiki/Resampling_(statistics)#Permutation_tests`

Significance of r_S (3/3)

- It is possible to test for significance also using:

$$w = r\sqrt{\frac{n-2}{1-r^2}}$$

- w follows a t distribution

$$w \sim t$$

Source with modifications: ***https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient***

Kendall's τ (1/2)

An alternative non parametric correlation coefficient is the Kendall's τ

- Let's have two sets $X = \{X_i\}$ and $Y = \{Y_i\}$ of the same size n where $(\forall i) X_i, Y_i \in$ ordinal scale
- Let's consider a set of pairs $P_{X,Y} = \{(X_i, Y_i)\}$
- Let's assume that the two sets X and Y do not contain duplicates
- Let's define
  - a concordant pair, a pair of pairs (X_i, Y_i) and (X_j, Y_j), with $i \neq j$ where either $(X_i > X_j$ and $Y_i > Y_j)$ or $(X_i < X_j$ and $Y_i < Y_j)$
  - a discordant pair, a pair of pairs (X_i, Y_i) and (X_j, Y_j), with $i \neq j$ where either $(X_i > X_j$ and $Y_i < Y_j)$ or $(X_i > X_j$ and $Y_i < Y_j)$

Source with modifications: `https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient`

Kendall's τ (2/2)

- We can define the Kendall's τ as:

$$\tau = \frac{(\#\ \text{concordant pairs}) - (\#\ \text{discordant pairs})}{n(n-1)/2}$$

Source with modifications: *https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient*

Logistic regression

Outline

- Likelihood function, definition
- Maximum likelihood
- Log likelihood
- Logistic regression

Some slides are take from:
`https://www.cs.ox.ac.uk/people/nando.defreitas/`

Likelihood function

Let $X_1, X_2, ..., X_n$ denote a random sample from p.d.f.

$$X_i \sim f_\theta(x),$$

where θ represents one ore more unknown parameters of the distribution.
The joint p.d.f. of $X_1, X_2, ..., X_n$ is $f_\theta(x_1), f_\theta(x_2), ..., f_\theta(x_n)$.
If we consider this joint p.d.f. as a function of θ it is called *likelihood function* of a random sample:

$$L_{x_1,x_2,...,x_n}(\theta) = f_\theta(x_1), f_\theta(x_2), ..., f_\theta(x_n).$$

Maximum likelihood (1/2)

Let's consider an estimator of θ:

$$\hat{\theta} = u(X_1, X_2, ..., X_n).$$

If for every possible θ $L_{x_1,x_2,...,x_n}(\hat{\theta})$ is at least as great as $L_{x_1,x_2,...,x_n}(\theta)$ then $\hat{\theta}$ is called *maximum likelihood estimator*.
Finally:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}(L_{x_1,x_2,...,x_n}(\theta))$$

Maximum loglikelihood (2/2)

Note that, since the likelihood function $L_{x_1,x_2,...,x_n}(\theta)$ and its logarithm $ln(L_{x_1,x_2,...,x_n}(\theta))$, are maximized for the same value θ, either likelihood or its logarithm can be used to find maximum likelihood estimator:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}(ln(L_{x_1,x_2,...,x_n}(\theta)))$$

The concept of regression

Regressions can be of multiple types, so far we have analysed the so called OLS regression:

- quadratic cost function of the kind $\sum_i (\hat{y}_i - y_i)^2$
- linear model of the kind $\hat{y} = \mathbf{Ax} + \eta$

What if:

- we use a different objective function, or
- we use a different model

?

Remember that model is called "the mean function" and its inverse "the link function."

Posing a different problem

Let's suppose to have:

- three iid random variables y_i with $i \in [1 \dots 3]$
- with the same partially unknown pdf, that is
- $(\forall i)\ y_i \sim N(\theta, 1)$
- θ to be determined.

We want to determine the value of θ that maximizes the probability of obtaining y_1 and y_2 and y_3.

In other terms our objective function is the probability of occurrence of y_1 and y_2 and y_3.

We are looking for a maximum likelihood estimator!

Computing the highest probability

Our objective function is therefore:

$$P(y_1, y_2, y_3|\theta) = P(y_1|\theta) \times P(y_2|\theta) \times P(y_3|\theta)$$

We can rewrite this problem as:

$$\max_{\theta}(\prod_{i=1}^{3} P(y_i|\theta))$$

Note that since θ is a *crisp* value:

$$y_i \sim N(\theta, 1) = \text{a shift of } \theta \text{ of } N(0, 1)$$

Using concrete numbers (1/2)

Let us assume that:

- $y_1 = 1$
- $y_2 = 0.5$
- $y_3 = 1.5$

Remember that $N(\theta, \sigma) = \dfrac{1}{\sigma\sqrt{2\pi}}\, e^{-\dfrac{(y-\theta)^2}{2\sigma^2}}$

Therefore, we want to maximize:

$$\prod_{i=1}^{3} P(y_i|\theta) = \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi}}\, e^{-\dfrac{(y_i-\theta)^2}{2}} =$$

$$= \frac{1}{\sqrt{2\pi}}\, e^{-\dfrac{(1-\theta)^2}{2}} \times \frac{1}{\sqrt{2\pi}}\, e^{-\dfrac{(0.5-\theta)^2}{2}} \times \frac{1}{\sqrt{2\pi}}\, e^{-\dfrac{(1.5-\theta)^2}{2}}$$

This is like maximizing:

$$e^{-\frac{(1-\theta)^2}{2}} \times e^{-\frac{(0.5-\theta)^2}{2}} \times e^{-\frac{(1.5-\theta)^2}{2}} =$$

$$= e^{-\frac{(1-\theta)^2}{2}-\frac{(0.5-\theta)^2}{2}-\frac{(1.5-\theta)^2}{2}} =$$

$$= e^{-\frac{(1-\theta)^2+(0.5-\theta)^2+(1.5-\theta)^2}{2}} = e^{-\frac{3.5-6\theta+3\theta^2}{2}}$$

This is like minimizing $g(\theta) = 3.5 - 6\theta + 3\theta^2$.

$$\frac{dg(\theta)}{d\theta} = \frac{d3.5 - 6\theta + 3\theta^2}{d\theta} = -6 + 6\theta$$

Which becomes 0 for $\theta = 1$

What we have discovered

Our solution is therefore $\theta = 1$ and the desired pdf is $N(1, 1)$. But ...

$$mean(1, 0, 5, 1.5) = 1$$

We can try to generalize it...

Generalizing ...

Let's suppose to have:

- n iid random variables y_i with $i \in [1 \dots n]$
- with the same partially unknown pdf, that is
- $(\forall i)\ y_i \sim N(\theta, \sigma)$
- θ and σ to be determined.

We want to determine the value of θ that maximizes the probability of obtaining $(\forall i)\ y_i$.

In other terms our objective is to maximize the probability of occurrence of all y_i, that is a maximum likelihood estimation.

Typically, we would perform a least square estimation, and we know that optimal least square estimator is the Gaussian centered in the average of the points, with their standard deviation.

Maximum likelihood estimator (again)

Let' look for a maximum likelihood estimator!

$$\max_{\sigma,\theta}(\prod_{i=1}^{n} P(y_i|\sigma,\theta)) = \max_{\sigma,\theta}(\prod_{i=1}^{n} P(y_i|\sigma,\theta)) = \max_{\sigma,\theta}(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(y_i-\theta)^2}{2\sigma^2}}) =$$

$$= \max_{\sigma,\theta}\left((\frac{1}{\sigma\sqrt{2\pi}})^n \prod_{i=1}^{n} e^{-\dfrac{(y_i-\theta)^2}{2\sigma^2}})\right) = \max_{\sigma,\theta}\left((\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\sum_{i=1}^{n}\dfrac{(y_i-\theta)^2}{2\sigma^2}}\right) =$$

$$= \max_{\sigma,\theta}\left((\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\dfrac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2}\right)$$

At this point we can take the log of the expression, knowing that the log function is differentiable and monotonically increasing on all $\mathbb{R}$.

Computing the ml estimator

$$log\Big((\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\theta)^2}\Big) =$$

$$= n \times log(\frac{1}{\sigma\sqrt{2\pi}}) + log(e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\theta)^2}) =$$

$$= n \times log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2} \times \sum_{i=1}^{n} (y_i - \theta)^2$$

Taking the partial derivative over θ we obtain:

$$\frac{\partial\Big(n \times log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2} \times \sum_{i=1}^n (y_i - \theta)^2\Big)}{\partial\theta} =$$

Computing the ml estimator - θ

$$= -\frac{\partial\Big(\frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i - \theta)^2\Big)}{\partial\theta} = -\frac{1}{\sigma^2} \times \Big(\sum_{i=1}^{n}(y_i - \theta)\Big)$$

And equating it to 0:

$$-\frac{1}{\sigma^2} \times \Big(\sum_{i=1}^{n}(y_i - \theta)\Big) = 0 \Rightarrow \sum_{i=1}^{n} y_i = n \times \theta \Rightarrow \theta = \frac{\sum_{i=1}^{n} y_i}{n}$$

Oh! θ is the average of the observed y_i!

Computing the ml estimator - σ (1/2)

$$\frac{\partial\left(n \times log(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i - \theta)^2\right)}{\partial\sigma} =$$

$$= \frac{\partial\left(n \times log(\frac{1}{\sigma\sqrt{2\pi}})\right)}{\partial\sigma} - \frac{\partial\left(\frac{1}{2\sigma^2} \times \sum_{i=1}^{n}(y_i - \theta)^2\right)}{\partial\sigma} =$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \times \sum_{i=1}^{n}(y_i - \theta)^2$$

And equating it to 0:

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \times \sum_{i=1}^{n}(y_i - \theta)^2 = 0 \Rightarrow \left(\sum_{i=1}^{n}(y_i - \theta)^2\right) \times \frac{1}{\sigma^3} = \frac{n}{\sigma}$$

Computing the ml estimator - σ (2/2)

Assuming $\sigma \neq 0$:

$$\Rightarrow \left(\sum_{i=1}^{n}(y_i - \theta)^2\right) = n \times \sigma^2 \Rightarrow$$

But we know $\theta = \overline{y_i}$, therefore:

$$\Rightarrow \sigma^2 = \frac{1}{n} \times \left(\sum_{i=1}^{n}(y_i - \overline{y_i})^2\right)$$

Oh! σ is the standard deviation of the observed y_i!

What we have found

We have determined that the maximum likelihood estimator for a sequence of points assumed to be distributed normally is formed by a normal distribution with:

- average equal to the average of the sample,
- standard deviation equal to the standard derivation of the sample.

This coincides with the best quadratic estimator!

We now move forward considering the maximum likelihood estimator for a regression line, meaning, what happens if now we want to model an interdependencies using as objective function the maximum likelihood.

Let's suppose to have:

- $n \times m$ values $x_{i,j}$ with $i \in [1 \ldots n]$, $j \in [1 \ldots m]$ represented in short by a matrix $\boldsymbol{X}$ or a vector $\mathbf{x_i}$, $n > m$ *(why?)*
- n iid random variables y_i with $i \in [1 \ldots n]$ represented in short by a vector $\boldsymbol{y}$
- a linear relationships $\boldsymbol{\theta}$ between $\boldsymbol{X}$ and $\boldsymbol{y}$, *that is, we use the usual link / mean functions*
- each y_i distributed normally with mean $\boldsymbol{x_i^T \theta}$ and standard deviation σ (the same σ for all y_i), that is
- $(\forall i)\ y_i \sim N(\boldsymbol{x_i^T \theta}, \sigma)$
- $\boldsymbol{\theta}$ and σ to be determined.

Ml linear regression - goals

We want to determine the value of $\boldsymbol{\theta}$ and σ that maximizes the probability of obtaining $(\forall i)\ y_i$, that is:

$$\max_{\theta,\sigma}(P(\mathbf{y}|\boldsymbol{X},\boldsymbol{\theta},\sigma)) = \max_{\boldsymbol{\theta},\sigma}(\prod_{i=1}^{n} P(y_i|\boldsymbol{x_i},\boldsymbol{\theta},\sigma))$$

In other terms, our objective function is the conditional probability of occurrence of all y_i.

Computing the optimal θ (1/3)

We can express for simplicity our equation in vectorial form:

$$\max_{\sigma,\boldsymbol{\theta}} \left(\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\dfrac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})}{2\sigma^2}} \right)$$

As mentioned, this is equivalent to maximizing the log:

$$\max_{\sigma,\boldsymbol{\theta}} \left(log \left(\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\dfrac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})}{2\sigma^2}} \right) \right)$$

Which becomes:

$$\max_{\sigma,\boldsymbol{\theta}} \left(n \times log \left(\frac{1}{\sigma\sqrt{2\pi}} \right) + log \left(e^{-\dfrac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})}{2\sigma^2}} \right) \right)$$

Computing the optimal θ (2/3)

$$\max_{\sigma,\boldsymbol{\theta}} \left(n \times log\left(\frac{1}{\sqrt{2\pi}}\right) + n \times log\left(\frac{1}{\sigma}\right) - \frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2} \right)$$

And now we take the partial derivative over $\boldsymbol{\theta}$:

$$\frac{\partial \left(n \times log\left(\frac{1}{\sqrt{2\pi}}\right) + n \times log\left(\frac{1}{\sigma}\right) - \frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2} \right)}{\partial \boldsymbol{\theta}} =$$

$$= -\frac{1}{2\sigma^2} \frac{\partial \left((\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta}) \right)}{\partial \boldsymbol{\theta}} = -\frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{X\theta})$$

And equating it to 0 we obtain:

$$-\frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{X\theta}) = 0 \quad \Rightarrow \quad \boldsymbol{y} = \boldsymbol{X\theta}$$

Computing the optimal θ (3/3)

If $\boldsymbol{X}$ were square, then the solution would be:

$$\boldsymbol{\theta} = \boldsymbol{X^{-1}y}$$

But, as we said, $n > m$, therefore the solution is given by:

$$\boldsymbol{\theta} = (\boldsymbol{X^T X})^{\boldsymbol{-1}}\boldsymbol{X^T y}$$

What a surprise, isn't it?

Computing the optimal σ

Starting from:

$$\max_{\sigma,\boldsymbol{\theta}} \left(n \times log\left(\frac{1}{\sqrt{2\pi}}\right) + n \times log\left(\frac{1}{\sigma}\right) - \frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2}\right)$$

And now we take the partial derivative over σ:

$$\frac{\partial \left(n \times log\left(\frac{1}{\sqrt{2\pi}}\right) + n \times log\left(\frac{1}{\sigma}\right) - \frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{2\sigma^2}\right)}{\partial\sigma} =$$

$$= -\frac{n}{\sigma} + \frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{\sigma^3}$$

And equating it to 0, assuming as usual $\sigma \neq 0$ we obtain:

$$\sigma^2 = \frac{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})}{n}$$

Maximum likelihood estimator - properties

Claim 1: The maximum likelihood estimator of a Gaussian distribution over a set of points coincides with the OLS estimator.
Proof: See above.
QED

Claim 2: The maximum likelihood linear regression coincides with the OLS linear regression.
Proof: See above.
QED

Bernoulli and maximum likelihood

The pdf of a Bernulli distribution can be represented in terms of conditional probability as:

$$P(x|\theta) = \theta^x (1-\theta)^{(1-x)}$$

where clearly x can only be 0 or 1.

We can now introduce the concept of entropy, already hinted in class. Entropy represents the level of uncertainty of a variable.

Entropy (and Bernulli and ml)

Definition (Entropy): Given a random vectorial variable x of n components and a parameter θ, we define entropy of x, $H(x)$ as:

$$H(x) = \sum_{i=1}^{n} p(x_i|\theta) \times log(p(x_i|\theta))$$

We notice that for a Bernulli distribution:

$$H(x) = (1-\theta) log(1-\theta) + \theta log(\theta)$$

Indeed, as θ tends to 0 or to 1 the uncertainty tends to 0, since the likely value of x tend to be 0 or 1 respectively.

From B&B plus ml to LR

We are now ready to move to study a radically different form or regression, the so-called logistic regression.

Our goal is to have a regression that not only represents a relationship between two variables, but is also possible to capture a prediction of probability.

However, the value of a probability is from 0 to 1, so we need a "good" function that can translate any value in such range.

We use often as such function the so-called "sigmoid function." To introduce the sigmoid we start with the definition of a "logistic function."

Logistic

Definition (Logistic function): Given $L, x_0 \in \mathbb{R}$, $k \in \mathbb{R}^+$ a logistic function $f(x)$ is defined as:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Properties (of the logistic function:)

- the domain is all $\mathbb{R}$
- the range is $[0 \dots L]$ if L is positive and $[L \dots 0]$ if L is negative
- $f(x)$ is continuous, monotonically increasing, and differentiable over all its domain
- $f(x)$ is symmetric over x_0
- k is the rate of growth of $f(x)$ and for $k \to +\infty$ $f(x)$ tends to become the step function in x_0

Sigmoid

Definition (Sigmoid): Given $k \in \mathbb{R}^+$, a sigmoid function $sigm(x)$ is defined as a logistic function with $L = 1$ and $x_0 = 0$:

$$sigm(x) = \frac{1}{1 + e^{-kx}}$$

Properties (of the sigmoid function):

- the domain is all $\mathbb{R}$
- the range is $[0 \ldots 1]$
- $sigm(x)$ is continuous, monotonically increasing, and differentiable over all its domain
- $sigm(x)$ is symmetric over 0
- k is the rate of growth of $sigm(x)$ and for $k \to +\infty$ $sigm(x)$ tends to become the step function

Toward a logistic regression (1/2)

Suppose that we want to determine if a given event is going to happen based on a series of n predictors $x_1 \ldots x_n$. We can model the probability of occurrence of the event with a random variable y.

It is as if we have a sequence of flipping of coins each with different values of the possible variables that affect the result, for instance the intensity of the flipping, the temperature, the wind, etc.

Based on such set we want to predict what will be the result of the next flipping, given a set of values assigned to the covariates.

Our question is what is:

$$\text{P(Head} \mid \text{strong toss, strong wind, 60 degrees)}$$

?

Toward a logistic regression (2/2)

Let's try to build a regression line.

As we mentioned, any time we compute a regression we need to determine:

- the function to use as a model, and in this case a linear function would not be suitable, since probabilities range from 0 to 1, for this reason we select a sigmoid function;
- the objective function, and in this case the least square would be inappropriate because it is not a proper metrics space, so we opt for maximizing the conditional probability, that is, we aim at a maximum likelihood estimation.

Logistic regression - HPs

Let

- (y_i, x_i) be a collection of pairs with:
  - $i \in [1 \ldots n]$
  - $y_i \in \{0, 1\}$
  - $x_i \in \mathbb{R}^m$
  - $n > m$
- assume that the y_i are iid random variables
- consider as target mean function the sigmoid
- consider as optimality criteria the maximum likelihood

Logistic regression - goals

We want to determine the values of the parameters that maximize the probability of obtaining $(\forall i)\ y_i$, that is:

$$\max_{Parameters}(P(\mathbf{y}|\boldsymbol{X}, Parameters) = \max_{\boldsymbol{\theta}}(\prod_{i=1}^{n} P(y_i|\boldsymbol{x_i}, Parameters))$$

In other terms, our objective function is the conditional probability of occurrence of all y_i.

Given our link/mean:

$$\max_{\theta}(P(\mathbf{y}|\boldsymbol{X}, \boldsymbol{\theta})) = \max_{\boldsymbol{\theta}}(\prod_{i=1}^{n} P(y_i|sigm(\boldsymbol{x_i}^T\boldsymbol{\theta}))$$

Logistic regression - structure

Since the pdf of a Bernulli distribution is:

$$P(z|k) = k^z(1-k)^{(1-z)}$$

For us the probability k of each event is "approximated" by the sigmoid function (our mean function):

$$\boldsymbol{k} = \frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}$$

And this lead us to

$$P(y_i|\boldsymbol{x_i},\boldsymbol{\theta}) = \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right)^{y_i} \times \left(1-\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right)^{1-y_i}$$

Logistic regression - the problem

Our problem has therefore the form of:

$$\max_{\theta}(P(\mathbf{y}|\boldsymbol{X},\boldsymbol{\theta})) = \max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}} \right)^{y_i} \times \left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}} \right)^{1-y_i}$$

It is like finding an n-dimensional hyperplane dividing the n-dimensional hyperspace in 2 parts, those leading to y being 0 and those leading to y being 1.

Logistic regression - solution (1/3)

Since the log function is continuous, differentiable and monotonically increasing in all $\mathbb{R}^+$, our problem is equivalent to:

$$\max_{\boldsymbol{\theta}} \left(log \left(\prod_{i=1}^{n} \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T \boldsymbol{\theta}}} \right)^{y_i} \times \left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T \boldsymbol{\theta}}} \right)^{1-y_i} \right) \right)$$

And, given the property of logs, this is like maximizing:

$$log \left(\prod_{i=1}^{n} \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T \boldsymbol{\theta}}} \right)^{y_i} \right) + log \left(\prod_{i=1}^{n} \left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T \boldsymbol{\theta}}} \right)^{1-y_i} \right) =$$

$$= \sum_{i=1}^{n} log \left(\frac{1}{1+e^{-\boldsymbol{x_i}^T \boldsymbol{\theta}}} \right)^{y_i} + \sum_{i=1}^{n} log \left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T \boldsymbol{\theta}}} \right)^{1-y_i} =$$

Logistic regression - solution (2/3)

$$= \sum_{i=1}^{n} y_i \times log\left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) + \sum_{i=1}^{n}(1-y_i) \times log\left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) = \cdots$$

A bit of logarithms...

$$log\left(\frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) = log(1) - log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) = -log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right)$$

$$log\left(1 - \frac{1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) = log\left(\frac{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}-1}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) = log\left(\frac{e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}{1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}}\right) =$$

$$= log\left(e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) - log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) = \boldsymbol{x_i}^T\boldsymbol{\theta} - log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) =$$

Logistic regression - solution (3/3)

$$= -\sum_{i=1}^{n} y_i \times log\left(1 + e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right) + \sum_{i=1}^{n}(1-y_i)\times\left(\boldsymbol{x_i}^T\boldsymbol{\theta} - log\left(1 + e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right)\right) =$$

For simplicity let w_i be $log\left(1 + e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right)$.

$$= -\sum_{i=1}^{n} y_i \times w_i + \sum_{i=1}^{n} \boldsymbol{x_i}^T\boldsymbol{\theta} - \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} y_i \times \boldsymbol{x_i}^T\boldsymbol{\theta} + \sum_{i=1}^{n} y_i \times w_i =$$

$$= \sum_{i=1}^{n} \boldsymbol{x_i}^T\boldsymbol{\theta} - \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} y_i \times \boldsymbol{x_i}^T\boldsymbol{\theta} =$$

$$= \sum_{i=1}^{n}(1 + y_i)\boldsymbol{x_i}^T\boldsymbol{\theta} - \sum_{i=1}^{n} w_i$$

Logistic regression - comments

Let $f(\theta)$ be:

$$\sum_{i=1}^{n}(1+y_i)\boldsymbol{x_i}^T\boldsymbol{\theta} - \sum_{i=1}^{n} log\left(1+e^{-\boldsymbol{x_i}^T\boldsymbol{\theta}}\right)$$

Claim: $f(\theta)$ is convex.

Proof: Omitted

Consequence: Optimization algorithms can easily find the maximum.