TO REGRESSION AND CORRELATION

Marianna Bolla, DSc

Institute of Mathematics, BME

April 16, 2020

1 REGRESSION ANALYSIS in general

The so-called supervised learning problem is the following: we want to approximate the random variable Y (called target, response, or dependent variable) with an appropriate function of the random variable X (called predictor, explanatory, or independent variable) with the method of *least squares*. That is,

$$\mathbb{E}(Y - g(X))^2 \to min.$$

over all measurable functions g of X. If the joint distribution of X and Y is known, there is a theoretical solution: $g^{optimal}(X) = \mathbb{E}(Y|X)$, i.e., the *conditional expectation* of Y, given X.

The optimal g is called *regression curve*, and it is linear if the distribution of (X, Y) is bivariate normal. By the Central Limit Theorem, a linear approximation makes sense in case of other continuous bivariate distributions too. Frequently, we estimate the regression parameters from a sample, and use so-called linearizing transformations.

2 Covariance and Pearson Correlation

The notions of covariance and correlation are crucial in linear regression.

Definition 2.1. *The covariance of the rv's X and Y is*

$$c = \mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y).$$

Note that $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$. Also note that if X and Y are independent rv's, then $\mathrm{Cov}(X, Y) = 0$, but usually not vice versa, except if their distribution is bivariate normal.

Definition 2.2. *The (Pearson) correlation (coefficient) of the rv's X and Y is*

$$r = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)} \cdot \sqrt{\mathrm{Var}(Y)}},$$

provided that $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$. Note that $|r| \leq 1$ and it is 1 if and only if there is a linear relation between X and Y as we will see later.

The sample versions are as follows.

Definition 2.3. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. bivariate sample. The empirical covariance of the population variables X and Y is*

$$C = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y},$$

whereas the empirical correlation is

$$R = \frac{C}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}.$$

The numerator is sometimes called product moment and $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2 = nS_X^2$. Also $|R| \leq 1$.

3 Linear regression

3.1 Theoretical solution

We will need the following first and second moments of X and Y:

$$\mathbb{E}(X) = m_1,\ \mathbb{E}(Y) = m_2,\ \mathrm{Var}(X) = \sigma_1^2,\ \mathrm{Var}(Y) = \sigma_2^2,\ \mathrm{Cov}(X, Y) = c,\ \mathrm{Corr}(X, Y) = r,$$

assume that $\sigma_1^2 > 0$. We are looking for the *regression line* $l(x) = \beta x + \alpha$ such that

$$h(\alpha, \beta) = \mathbb{E}(Y - \beta X - \alpha)^2 \to \min. \quad in \quad \alpha, \beta.$$

Taking the partial derivatives and making them equal to zero, we get the following system of linear equations (we also interchanged the differentiation and taking the expectation):

$$\frac{\partial h}{\partial \beta} = -2\mathbb{E}[(Y - \beta X - \alpha)X] = 0$$
$$\frac{\partial h}{\partial \alpha} = -2\mathbb{E}[Y - \beta X - \alpha] = 0,$$

or equivalently,

$$\beta \cdot \mathbb{E}(X^2) + \alpha \cdot \mathbb{E}(X) = \mathbb{E}(XY)$$
$$\beta \cdot \mathbb{E}(X) + \alpha = \mathbb{E}(Y).$$

The cefficient matrix of the unknowns β and α is

$$H = \begin{pmatrix} \mathbb{E}(X^2) & \mathbb{E}(X) \\ \mathbb{E}(X) & 1 \end{pmatrix},$$

the determinant of which is $|H| = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \sigma_1^2 > 0$, so with Cramér's rule, the minimizers are:

$$\beta = \frac{c}{\sigma_1^2} = \frac{r\sigma_1\sigma_2}{\sigma_1^2} = r\frac{\sigma_2}{\sigma_1}, \quad \alpha = \mathbb{E}(Y) - \beta\mathbb{E}(X) = \mu_2 - \frac{c}{\sigma_1^2}\mu_1.$$

(They indeed give a minimum as the Hesse matrix H is positive definite.)

Therefore, the equation of the regression line is

$$y = \frac{c}{\sigma_1^2}(x - \mu_1) + \mu_2$$

or

$$\frac{y - \mu_2}{\sigma_2} = r\frac{x - \mu_1}{\sigma_1}.$$

About te origin of the word 'regression': Sir Francis Galton British doctor, biologists, and mathematician (in the 19th century) investigated the relation between the father–son body-hight. He assumed that $\sigma_1 = \sigma_2 = \sigma$. Then the hight of the son (Y) can be predicted with the hight of the father (X) linearly as

$$Y = m_2 + r(X - m_1),$$

where $|r| \leq 1$, and so,

$$|Y - m_2| \leq |X - m_1|.$$

Therefore, if $r > 0$, then son of a man taller than the average is taller than the population average of his generation, but less tall; likewise, son of a man shorter than the average is shorter than the population average of his generation, but less short. Galton called this phenomenon 'returning to the average', for which the Latin word was 'regression'.

The problem can be described by the following linear model:

$$Y = \ell(X) + \varepsilon = (\beta X + \alpha) + \varepsilon$$

where $\mathbb{E}(\varepsilon^2)$, or equivalently, $\mathrm{Var}(\varepsilon)$ is minimized, since $\mathbb{E}(\varepsilon) = 0$. Because of

$$\mathrm{Cov}(l(X), \varepsilon) = \mathrm{Cov}(\beta X + \alpha, Y - \beta X - \alpha) = \mathrm{Cov}(\beta X, Y - \beta X) =$$

$$= \beta c - \beta^2\sigma_1^2 = \frac{c}{\sigma_1^2}c - \frac{c^2}{(\sigma_1^2)^2}\sigma_1^2 = 0,$$

$$\mathrm{Var}(Y) = \mathrm{Var}(l(X)) + \mathrm{Var}(\varepsilon),$$

where

$$\mathrm{Var}(\varepsilon) = \mathrm{Var}(Y) - \mathrm{Var}(l(X) = \sigma_2^2 - \frac{c^2}{\sigma_1^2} = \sigma_2^2 - \frac{r^2\sigma_1^2\sigma_2^2}{\sigma_1^2} = \sigma_2^2(1 - r^2).$$

This gives rise to the following decomposition of the variance of Y:

$$\mathrm{Var}(Y) = r^2\mathrm{Var}(Y) + (1 - r^2)\mathrm{Var}(Y). \tag{1}$$

The first term on the right hand side is the variance of Y explained by the predictor variable, and the second one is the so-called *residual variance*, that is, the variance of the error term ε. Observe that $r^2 = 1$ is equivalent to $\mathrm{Var}(\varepsilon) = 0$, i.e., there is a linear relation between Y and X with probability 1. The other extreme case $r^2 = 0$ means that $\mathrm{Var}(l(X)) = 0$, i.e., the best linear prediction is constant with probability 1, consequently, $\beta = 0$; in other words, Y is uncorrelated with X, and hence, its best linear approximation is its own expectation.

Note, that with some *linearization* formulas we can use linear regression (sometimes multivariate) in the following models:

- *Multiplicative model*:
$$Y \sim b{X_1}^{a_1} \dots {X_p}^{a_p}.$$
After taking the logarithms, one gets
$$\ln Y \sim \ln b + a_1 \ln X_1 + \cdots + a_p \ln X_p,$$
therefore, we can use linear regression for the log-log data. Whereas the linear regression performs well for data from a multivariate normal distribution, this model favors lognormally distributed data (for example, chemical concentrations).

- *Polynomial regression*: Now we want to approximate Y with a given degree polynomial of X:
$$Y \sim a_1 X + a_2 X^2 + \cdots + a_p X^p + b.$$
The solution is obtained by applying multivariate linear regression for Y with the predictor variables $X_j = X^j$, $j = 1, \dots, p$.

3.2 Estimating the regression coefficients from a sample

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. bivariate sample. Our objective is
$$\sum_{i=1}^{n} (Y_i - \beta X_i - \alpha)^2 \to \min . \quad in \quad \alpha, \beta.$$
The solution is given by the corresponding sample moments:
$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} = R\frac{S_Y}{S_X}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = \bar{Y} - R\frac{S_Y}{S_X}\bar{X},$$
where $S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ and $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$. The sample variance decomposition is as follows.
$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{[\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \sum_{i=1}^{n}(Y_i - \hat{\beta}X_i - \hat{\alpha})^2,$$
or briefly,
$$SST = SSR + SSE = R^2 \cdot SST + (1 - R^2) \cdot SST,$$
where $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is the total variation of the measurements (sum of squares total),
$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}X_i - \hat{\alpha})^2$$
is the *residual sum of squares* (sum of squares due to error), and $SSR = SST - SSE$ is the part of the total variation explained by the regression (sum of squares due to regression). Further, R is the sample correlation coefficient, and R^2 is called *coefficient of determination.*

3.3 The linear model (with deterministic predictors)

Now our model is the following.

$$Y_i = \beta x_i + \alpha + \varepsilon_i \qquad (i = 1, \dots, n),$$

where x_i is the prescribed value of the predictor in the i-th measurement (no error in it). Since the measurement is burdened with the random noise ε_i, the measured value Y_i of the target variable in the i-th measurement is a random variable. We also assume that $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ $(i = 1, \dots, n)$, and the measurement errors are uncorrelated. Because of their equal (but unknown) variance they are called *homoscedastic errors*.

For the estimates of the parameters we have the analogous formulas

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{j=1}^n (x_j - \bar{x})^2} = \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} Y_i = \sum_{i=1}^n k_i Y_i$$

$$\hat{\alpha} = \bar{Y} - \hat{a}\bar{x} = \sum_{i=1}^n \left(\frac{1}{n} - \bar{x} k_i \right) Y_i = \sum_{i=1}^n l_i Y_i,$$

that are linear functions of Y_i's.

Gauss–Markov theorem: In the linear model, the above $\hat{\alpha}, \hat{\beta}$ are linear unbiased estimators of the parameters α and β, respectively; further, among all such estimatotrs, they have the minimal variance. Breafly, they are BLUE (Best Linear Unbiased Estimators).

Note that the unbiased estimator for σ^2 is

$$s^2 := \hat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta} x_i - \hat{\alpha})^2 = \frac{1}{n-2} SSE.$$

If $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \dots, n$ i.i.d., then

$$\hat{\beta} \sim \mathcal{N}(\beta, \frac{\sigma^2}{s_{xx}}), \quad \hat{\alpha} \sim \mathcal{N}(\alpha, \sigma^2 [\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}])$$

and $\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$, independently of $\hat{\alpha}$ and $\hat{\beta}$. With the help of these, we can test hypotheses, like

$$H_0 : \beta = 0 \quad \text{vers.} \quad H_1 : \beta \neq 0.$$

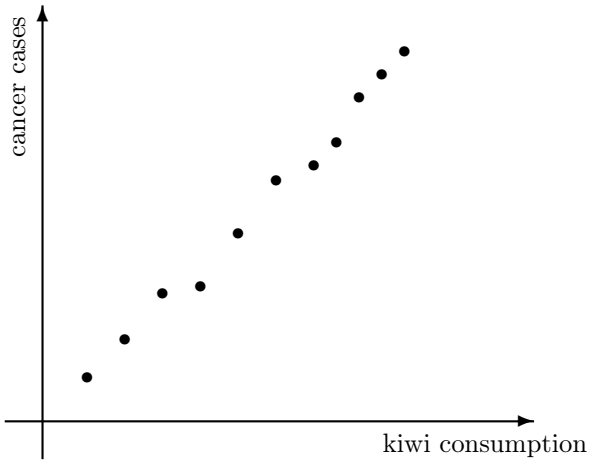
The test statistic is

$$t = \frac{(\hat{\beta} - 0) \frac{\sqrt{s_{xx}}}{\sigma}}{\sqrt{\frac{(n-2)s^2}{\sigma^2} / (n-2)}} = \hat{\beta} \frac{\sqrt{s_{xx}}}{s}$$

that under H_0 follows Student t-distribution with df. $n-2$. Therefore, we reject H_0 with significance ε if $|t| \geq t_{\varepsilon/2}(n-2)$. If it holds with small ε, then 'we can significantly predict Y with X'.

Warning: high correlation and significant regression between two variables, selected from a multivariate data, set can be misleadig. For example, the yearly kiwi consumption and the number of cancer cases in the US between 1970 and

1980 shows a high correlation. However, here other hidden variables, like increasing living standard and possibly increasing consumption of tobacco goods and alcoholic drinks, are eliminated. These issues are treated, e.g., by graphical models.



4 Rank Correlation

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. bivariate sample from a continuous distribution, ties have zero probability. Let $R_1, \dots, R_n$ denote the ranks of the X-bservations, and $S_1, \dots, S_n$ those of the Y-observations between themselves. The *Spearman rank correlation coefficient* between the X and Y sample is

$$r_{sp} = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2}\sqrt{(S_i - \bar{S})^2}} = \frac{\sum_{i=1}^n (R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2})}{n(n^2-1)/12}.$$

As r_{sp} is the Pearson correlation between the ranks, $|r_{sp}| \le 1$, but can be 1 if there is a monotone relation between the X and Y sample.

Bonus exercise 3. Prove that

$$r_{sp} = 1 - \frac{6\sum_{i=1}^n D_i^2}{n(n^2-1)},$$

where $D_i = R_i - S_i$.

This means that the rank correlation depends only on the difference of the ranks, and so, it is unchanged if the same constant is added to the X and Y measurements. For example if this is the rank correlation between the body hights of man–woman couples, it does not change if they go up to a peak, and their hight is measured from the sea level. In contrast, the Pearson correlation gets higher ad higher if they go higher and higher.