REGRESSION ANALYSIS

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

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

over all measurable functions g. If the joint distribution of X and Y is known, there is a theoretical solution. The optimal g is called *regression curve*, and it is linear if X, Y is bivariate normal. By the Central Limit Theorem, a linear approximation makes sense. Frequently, we estimate the regression parameters from a sample, and use so-called linearizing transformations.

1 Linear regression

1.1 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

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

The minimizers are:

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

Therefore, the equation of the regression line is

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

or

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

(The name regression comes from Galton: returning to the mean.)

The problem as well can be described by the following linear model:

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

where $\mathbb{E}(\varepsilon^2)$, or equivalently, $\mathrm{Var}(\varepsilon)$ is minimized, since $\mathbb{E}(\varepsilon) = 0$. An easy calculation shows that $\mathrm{Cov}(l(X), \varepsilon) = 0$, therefore,

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

where

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

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 approximation 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:

- *Multivariate regression*:

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

- *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:

$$Y \sim \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \alpha.$$

  The solution is obtained by applying multivariate linear regression for Y with the predictor variables $X_j = X^j$, $j = 1, \dots, p$.

1.2 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:

$$\hat{\beta} = \frac{S_{XY}}{S_X^2} = R \frac{S_Y}{S_X}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = \bar{Y} - R\frac{S_Y}{S_X}\bar{X},$$

where $S_{XY} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$ and $S_X^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$. The sample variance decomposition is as follows.

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \frac{S_{XY}}{S_X^2} + \sum_{i=1}^{n} (Y_i - \hat{\beta} X_i - \hat{\alpha})^2,$$

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, $R = \frac{S_{XY}}{S_X S_Y}$.

1.3 The linear model (with deterministic predictors)

Now our model is the following.

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

where x_i is the prescribed value of the predictor in the i-th measurement. 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 $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ $(i = 1, \ldots, n)$, 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).

Theorem: Assume that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. $(i = 1, \ldots, n,\ n > 2)$. Then the above $\hat{\alpha}$, $\hat{\beta}$ are ML-estimators of the parameters α, β in the linear model. Further, the ML-estimator of the parameter σ^2 is

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

Note that the unbiased estimator for σ^2 is $\frac{1}{n-2}SSE$.