1 Linear regression

It's a very simple approach but yet still effective approach supervised learning, that is used in particular to predict a quantitative response. More complex approaches are extensions or generalizations of linear regression. When doing regression, we ask:

- Is there a relationship between the features and the predicted variables?
- If so, how strong is it?
- How do each individual feature contribute to the predicted variable?
- How accurately can we predict future values?
- Is the relationship linear?
- Is there an interaction among the features?

In statistics, an **interaction** may arise when considering the relationship among three or more variables, and describes a situation in which the effect of one causal variable on an outcome depends on the state of a second causal variable (that is, when effects of the two causes are not additive).

1.1 Simple linear regression

In a simple linear regression, we assume there is a very straighforward approacheto predicting a quantitative approach:

$$Y \approx \beta_o + \beta_1 X$$

We say that Y is approximately modeled by X. β_0 and β_1 are the coefficients, or paremeters, of the model, respectively the intercept and the slope.

1.1.1 Estimating the coefficients

The coefficients are unknown. We use the data $\{(x_1,y_1), \ldots, (x_n, y_n)\}$ to predict their value. We want to find β_0 and β_1 such that $y = \hat{\beta_0} + \hat{\beta_1} x$ is the closest to the n points. Closeness can be assessed in several ways. The most common approach involves minimizing the least squares criterion.

- $\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$: prediction for Y based on the ith value of X.
- $e_i = y_i - \hat{y_i}$: ith residual (difference between the observed response and the predicted response)
- We define the residual sum of squares (RSS): $RSS = \sum_{i=1}^{n} e_i^2$. The goal is the minize this quantity. The quantities we obtain are called the least squares coefficient estimates.

1.1.2 Accuracy of the coefficient estimates

The least squares regression line is an estimation of the population regression line (the truth) The true relationship between Y and X is given by $Y = f(X) + \varepsilon$ where ε is the error term. It's a catch-all quantity representing everything we missed in the regression: the true relationship may not be perfectly linear or some other variables were not taken into account.

We assume that there exist a true linear relationship between Y and X, which we call the population regression line ($f(X) = 2 + 3X$ for instance). We consider that the data set we observe follows this relationship, within a variation described by an error term following a normal distribution with mean zero. Then, the least squares regression line is an estimation of this true population regression line. This approach is very similar to the one we use in classical statistics. The population mean is estimated by taking the sample mean.

The concept of unbiaisedness holds for the least squares method. Over a large number of data sets, we expect that the average of least square coefficient estimates will be very close to the real parameters.

Standard errors allow to assess the accuracy of the coefficient estimates For each individual estimation, we wonder how far it is from the truth. In estimation, we use the standard error (SE) given by the formula $\hat{Var(\mu)} = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$.

Standard deviation The standard deviation quantifies the variation within a set of measurements. It is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range.

Standard error The standard error (SE) of a statistic (usually an estimate of a parameter) is the standard deviation of its sampling distribution or an estimate of that standard deviation. Let's imagine we estimation the mean of a population based on five different samples. We will obtain five different estimates for the same true mean. The standard error is the standard deviation of the means around the real mean.

Sampling distribution The sampling distribution of a population mean is generated by repeated sampling and recording of the means obtained. This forms a distribution of different means, and this distribution has its own mean and variance. Mathematically, the variance of the sampling distribution obtained is equal to the variance of the population divided by the sample size. This is because as the sample size increases, sample means cluster more closely around the population mean.

Similarly for the least squares coefficient estimates, we are able to compute the standard errors of β_0 and β_1 :

$$\hat{SE(\beta_0)}^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where $\sigma^2 = Var(\varepsilon)$. This means that the spread of the sample we observe around the mean is described by the variance of the error term, which is consistent with what we stated before. It is usually not known, but it can be estimated from the data. The estimate of σ^2 is called the residual standard error (RSE) and is given by $RSE = \sqrt{\frac{RSS}{n-2}}$.

Standard errors allow to compute confidence intervals Standard errors are used to compute confidence intervals at a level α. A $1-\alpha$ confidence interval is a range of values such that the true value of the unknown parameter will fall in the interval with probability $1 - \alpha$. A 95% confidence interval of β_1 is $\left[\beta_1 \pm 2\hat{SE}(\hat{\beta_1})\right]$.

Hypothesis testing uses confidence intervals to ensure that a linear relationship actually exists between X and Y Confidence intervals are used to test if there exists a linear relationship between X and Y, i.e. $H_0 : \beta_1 = 0$ (the null hypothesis) is rejected. H_0 is rejected when $\hat{\beta}_1$ is sufficiently far from zero. How far it needs to be is determined by the accuracy of $\hat{\beta}_1$, i.e. by $SE(\hat{\beta_1})$. If the standard error is small, than even a relatively small value of the estimate will ensure that there exists a linear relationships.

Hypothesis testing We perform hypothesis testing on the value of parameters. Testing relies on a procedure which chooses between H_0 and H_1 depending on the observation $X = x$. In the Neyman-Pearson approach, H_0 is choosen to be the dominating starting point while H_1 is choosen to be the risky hypothesis that we want to prove.