Methods

Ordinary least squares

Ordinary least squares model or OLS, works by creating a line through the data points. Then it calculates the difference between each prediction and observation (residual). And it tries to minimize the squared value of the residuals. The ordinary least squares is defined by:

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

The least squares estimates in this case are given by simple formulas

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Quantile regression

In Koneker's 1978 paper, the θ^{th} Quantile regression is defined as any solution to the following problem:

$$\min_{b \in \mathbf{R}^K} \left[\sum_{t \in \{t : y_t \geqslant x_t b\}} \theta \left| y_t - x_t b \right| + \sum_{t \in \{t : y_t < x_t b\}} (1 - \theta) \left| y_t - x_t b \right| \right] \tag{0.1}$$

where

$$\{x_t : t = 1, ..., T\}$$

denotes a sequence (row) of K-vectors of a known design matrix and

$$\{y_t : t = 1, ..., T\}$$

is a random sample on the regression process $u_t = y_t - x_t \beta$ [1].

Evaluation metrics

Mean absolute error

The mean absolute error (MAE) is the average magnitude of the errors of the values predicted by the regression and the actual observed values for the response variable. Because it is a simple average, all errors have the same weight, there are no penalties for different magnitude deviations [2]. MAE assumes that the errors are normally distributed, if the error distribution was non-normal, the average may not be a good measure of centrality and can paint a false picture of the goodness-of-fit of the regression curve. MAE also assumes that the errors are unbiased. While the average magnitude of the errors is expected to be non-zero (unless the regression is a perfect fit) the average of the residuals, i.e., the deviation of the predicted value from the actual value, considering underestimation and overestimation. This means on average the regression curve does not over or underestimate.

$$\text{MAE } = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

Root mean squared error

It calculates the differences between the predictions and the actual observations (residuals) and then gets their quadratic mean for each. This type of error gives a larger penalty for larger errors [2]. This error also assumes that the errors are unbiased and that they follow a normal distribution. This gives a picture of the size of residuals in comparison to the regression line.

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2}$$

Variance of error

It is a measure of how spread all the errors are from the mean of all errors.

$$\text{Var}(e) = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2$$

Min/max error

A measure of the maximum residual for a prediction and the minimum residual.

$$f : X \to \mathbb{R}, \text{ if } (\forall e \in X_{error}) f(e_i) \geq f(e)$$

$$f : X \to \mathbb{R}, \text{ if } (\forall e \in X_{error}) f(e_i) \leq f(e)$$