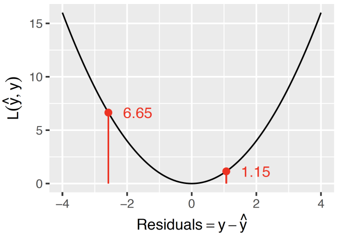
# Introduction to Machine Learning

# Evaluation: Measures for Regression



## Learning goals

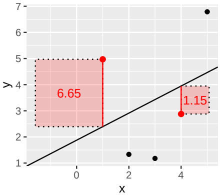
- Know the definitions of mean squared error (MSE) and mean absolute error (MAE)
- Understand the connections of MSE and MAE to L2 and L1 loss
- Know the definition of Spearman's $\rho$
- Know the definitions of $R^2$ and generalized $R^2$

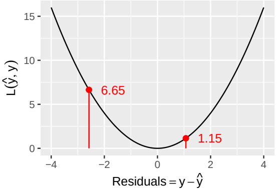
# MEAN SQUARED ERROR

The **mean squared error (MSE)** computes the mean of squared distances between the target variable $y$ and the predicted target $\hat{y}$.

$$\rho_{MSE}(\mathbf{y}, \boldsymbol{F}) = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 \in [0; \infty) \qquad \rightarrow \textit{L2} \text{ loss.}$$

Outliers with large prediction error heavily influence the MSE, as they enter quadratically.





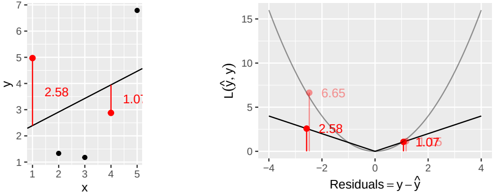
- Sum of squared errors (SSE)
- Root mean squared error (RMSE) → original scale

# MEAN ABSOLUTE ERROR

A more robust (but not necessarily better) alternative is the **mean absolute error (MAE)**:

$$\rho_{MAE}(\mathbf{y}, \boldsymbol{F}) = \frac{1}{n}\sum_{i=1}^{n}|y^{(i)} - \hat{y}^{(i)}| \in [0; \infty) \qquad \rightarrow L1 \text{ loss.}$$

The MAE is less strongly impacted by large errors and maybe more intuitive than the MSE.



Similar measures:

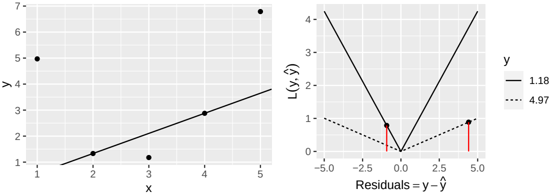
- Median absolute error (for even more robustness)

# MEAN ABSOLUTE PERCENTAGE ERROR

The relative error can be measured with the **mean absolute percentage error (MAPE)**:

$$\rho_{MAPE}(\mathbf{y}, \boldsymbol{F}) = \sum_{i=1}^{n} \left| \frac{y^{(i)} - \hat{y}^{(i)}}{y^{(i)}} \right| \in [0; \infty)$$

The smaller the absolute target values, the stronger they influence the MAPE-optimal model. Cannot handle zero target values.



Similar measures:

- Mean Absolute Scaled Error (MASE)
- Symmetric Mean Absolute Percentage Error (sMAPE)

# $R^2$

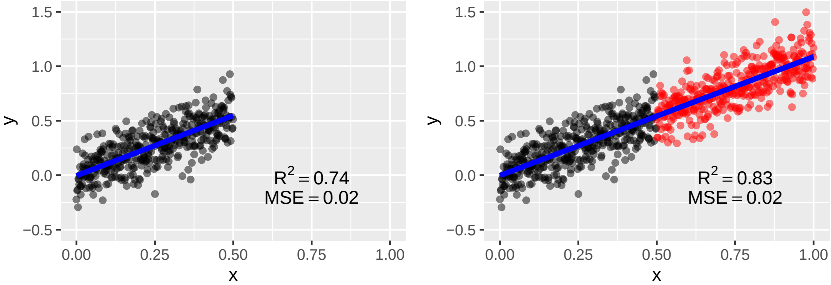
Another well known measure from statistics is $R^2$:

$$\rho_{R^2}(\mathbf{y}, \boldsymbol{F}) = 1 - \frac{\sum\limits_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2}{\sum\limits_{i=1}^{n}(y^{(i)} - \bar{y})^2} = 1 - \frac{SSE_{LinMod}}{SSE_{Intercept}}.$$

- Usually introduced as **fraction of variance explained** by the model.
- Simpler explanation: it compares the SSE of a constant model (baseline) to that of a more complex model (LM) on some data, usually the same as used for model fitting.
- $\rho_{R^2} = 1$: all residuals are 0, we predict perfectly,
  $\rho_{R^2} = 0$: we predict as badly as the constant model.
- If measured on the training data, $\rho_{R^2} \in [0; 1]$, as the LM must be at least as good as the constant, and both SSEs are non-negative.
- On other data $R^2$ can even be negative as there is no guarantee that the LM generalizes better than a constant (overfitting).

# $R^2$ VS MSE

- An improvement in the fraction of variability explained by the model does not necessarily mean a better model fit:



  Here, we generate data with $y = 1.1x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.15)$, and fit the half (black) and the full data set (black and red) with a linear model, respectively. Although the fit does not improve, the $R^2$ value rises.

- While $R^2$ is invarariant with respect to linear scaling of the target values, the MSE is not.

# GENERALIZED $R^2$ FOR ML

A simple generalization of $R^2$ for ML seems to be:

$$1 - \frac{Loss_{ComplexModel}}{Loss_{SimplerModel}}.$$

- This introduces a general measure of comparison between a simpler baseline and a more complex model considered as an alternative.
- Works for arbitrary measures (not only SSE), for arbitrary models, on any data set of interest.
- E.g., feature model vs constant, LM vs non-linear model, tree vs forest, model with fewer features vs model with more, ...
- In ML we would rather evaluate that metric on a hold-out test set – there is no reason not to do that.
- Fairly unknown; our terminology (generalized $R^2$) is non-standard.

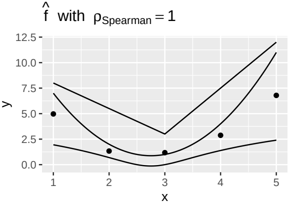
# SPEARMAN'S $\rho$

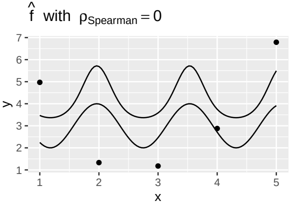
Spearman's $\rho$ measures the rank correlation, i.e.,

$$\rho_{\text{Spearman}}(\mathbf{y}, \boldsymbol{F}) = \frac{\text{Cov}(\mathrm{rg}(\mathbf{y}), \mathrm{rg}(\hat{\mathbf{y}}))}{\sqrt{\text{Var}(\mathrm{rg}(\mathbf{y}))} \cdot \sqrt{\text{Var}(\mathrm{rg}(\hat{\mathbf{y}}))}} \in [-1, 1],$$

where $\mathrm{rg}$ is the ranking function (e.g. $\mathrm{rg}((4, 0.5, 10)) = (2, 1, 3)$).

- It is very robust against outliers, since the correlation is only based on the ranks of $\mathbf{y}$ and $\hat{\mathbf{y}}$, respectively.
- A value of 1 or -1 means that $\hat{\mathbf{y}}$ and $\mathbf{y}$ have a perfect monotonic relationship.
- A value of zero indicates no association between $\mathrm{rg}(\mathbf{y})$ and $\mathrm{rg}(\hat{\mathbf{y}})$.
- It only measures the monotonic relationship, i.e., any strictly increasing transformation applied to $\hat{\mathbf{y}}$ does not alter $\rho_{\text{Spearman}}$.





# ML VS CLASSICAL STATISTICS

- In classical statistics, besides MSE, RMSE and in-sample $R^2$, other metrics are used to evaluate and select regression models.
- They often focus on goodness-of-fit, as measured by (log-)likelihood, rather than predictive accuracy – for example, information criteria:
  - **Akaike's information criterion (AIC)** balances model fit and complexity, penalizing the number of parameters, *p*:

    $$AIC = -2 \cdot \ell(\boldsymbol{\theta}) + 2 \cdot p.$$

  - **Bayesian information criterion (BIC)** is another variant of the AIC with a stronger penalty for more complex models:

    $$BIC = -2 \cdot \ell(\boldsymbol{\theta}) + \log(p).$$

- As both AIC and BIC are based upon a ground-truth distribution, they cannot be used to compare performances across different data sets.
- NB: using the same data for training and evaluation / model selection introduces optimistic bias $\rightarrow$ post-selection inference.