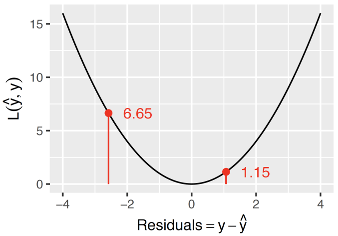
Introduction to Machine Learning

Evaluation
Measures for Regression



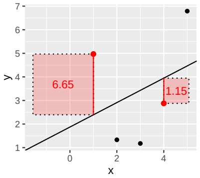
Learning goals

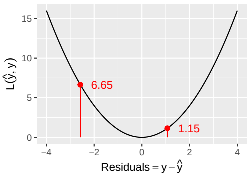
- Know the definitions of mean squared error (MSE) and mean absolute error (MAE)
- Understand the connections of MSE and MAE to L2 and L1 loss
- Know the definition of Spearman's ρ
- Know the definitions of R^2 and generalized R^2

MEAN SQUARED ERROR (MSE)

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

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





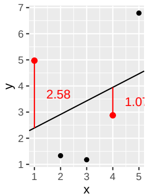
Similar measures:

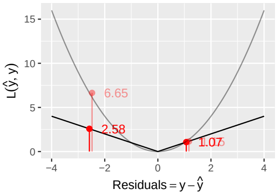
- Sum of squared errors: $\rho_{SSE}(\mathbf{y}, \boldsymbol{F}) = \sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)})^2$
- Root MSE (orig. scale): $\rho_{RMSE}(\mathbf{y}, \boldsymbol{F}) = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)})^2}$

MEAN ABSOLUTE ERROR

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

More robust, less influenced by large residuals, more intuitive than MSE.





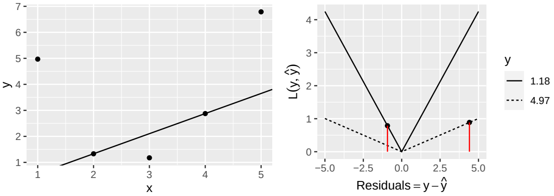
Similar measures:

- Median absolute error (for even more robustness)

MEAN ABSOLUTE PERCENTAGE ERROR

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

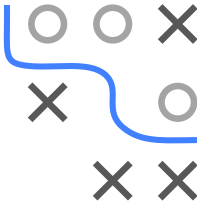
Small $|y|$ influence more strongly. Cannot handle $y = 0$.



Similar measures:

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

R^2

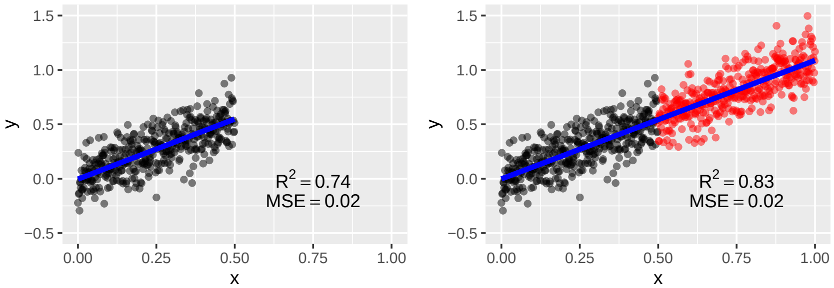


$$\rho_{R^2}(\mathbf{y}, \boldsymbol{F}) = 1 - \frac{\sum\limits_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2}{\sum\limits_{i=1}^{m}(y^{(i)} - \bar{y})^2} = 1 - \frac{SSE_{LinMod}}{SSE_{Intercept}}.$$

- Well-known classical measure for LMs – on train data.
- "Fraction of variance explained" by the model.
- How much SSE of constant baseline is reduced when we use more complex model?
- $\rho_{R^2} = 1$: all residuals are 0, we predict perfectly,
- $\rho_{R^2} = 0.9$: LM reduces SSE by factor of 10.
  $\rho_{R^2} = 0$: we predict as badly as the constant model.
- Is $\in [0, 1]$ on train data; as LM is always better than intercept.

R^2 VS MSE

- Better R^2 does not necessarily imply better fit.
- Data: $y = 1.1x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.15)$.
- Fit half (black) and full data (black and red) with LM.



- Fit does not improve, but R^2 goes up.
- But: Invariant w.r.t. to linear scaling of y, MSE is not.

GENERALIZED R^2 FOR ML

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

- E.g., model vs constant, LM vs non-linear model, tree vs forest, model with fewer features vs model with more, ...
- We could use arbitrary measures.
- In ML we would rather evaluate on test set.
- Can then become negative, e.g., for SSE and constant baseline, if our model fairs worse on the test set than a simple constant.

SPEARMAN'S ρ

Can be used if we care about the relative ranks of predictions:

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

- Very robust against outliers
- A value of 1 or -1 means that $\hat{\mathbf{y}}$ and $\mathbf{y}$ have a perfect monotonic relationship.
- Invariant under monotone transformations of $\hat{\mathbf{y}}$