Evaluating Regression Models

Data Science Dojo

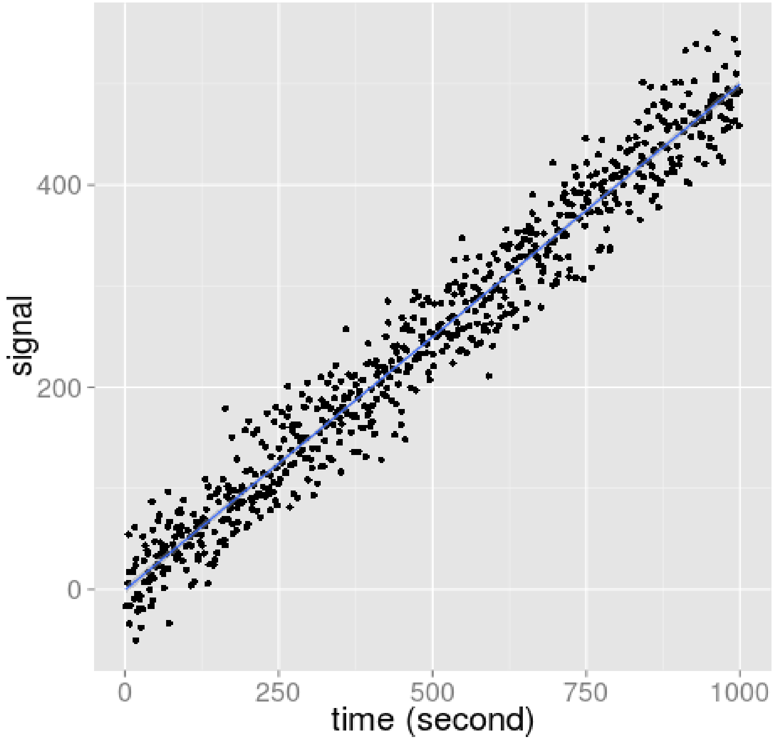
datasciencedojo
unleash the data scientist in you

A simple example

- A **linear regression model** is built based on the training data set:
  signal =
  h(time) = -0.498 + 0.500 x time

  How to evaluate this model?

  What are the evaluation metrics for such regression models?



Basic metrics

Root-mean-square deviation (RMSE)

Mean Absolute Error(MAE)

Coefficient of determination (R Squared)

RMSE

Formula: $$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - h(\boldsymbol{x}_i))^2}{n}}$$

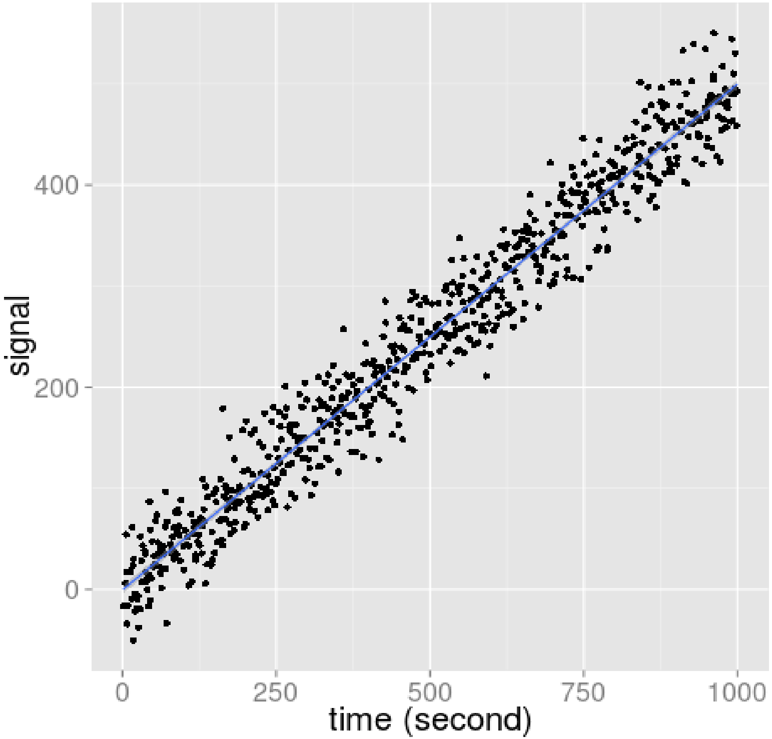
for data set i, $\boldsymbol{x}_i$ is a vector of all the predictors,
y_i is the corresponding response;
$h(\boldsymbol{x}_i)$ is the prediction using a certain model;
n is the number of data in the test data set.

RMSE

- Back to our simple **linear regression model**:

  signal =
  h(time) = -0.498 + 0.500 x time

  700 data items are used to **train** this model;
  Another 300 data items in **testing data set** are used to measure this model.



datasciencedojo
unleash the data scientist in you

RMSE

Using the testing data set with n = 300 data items:

$y_i: \ -5.905, \ 48.261, \ , \ 4.115, \ -8.370, \ 42.222, \ 10.320, \ ...$

$h_{(x_i)}: \ 0.502, \ 1.003, \ 1.503, \ 4.505, \ 5.005, \ 6.507, \ ...$

$h_{(x_i)} - y_i: \ 6.408, \ -47.259, \ -2.612, \ 12.875, \ -37.217, \ -3.814, \ ...$

$(h_{(x_i)} - y_i)^2: \ 41.059, \ 2233.367, \ 6.823, \ 165.766, \ 1385.102, \ 14.546, \ ...$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - h(\boldsymbol{x_i})\right)^2}{n}} = 29.17852$$

MAE

Formula: $$MAE = \sqrt{\frac{\sum_{i=1}^{n} |residual_i|}{n}} = \sqrt{\frac{\sum_{i=1}^{n} |y_i - h(\boldsymbol{x}_i)|}{n}}$$

for data set i, $\boldsymbol{x}_i$ is a vector of all the predictors,
y_i is the corresponding response;
$h(\boldsymbol{x}_i)$ is the prediction using a certain model;
n is the number of data in the test data set.

MAE

Using the testing data set with n = 300 data items:

$$y_i:\ -5.905,\ 48.261,\ ,\ 4.115,\ -8.370,\ 42.222,\ 10.320,\ \ldots$$

$$h_{(x_i)}:\ 0.502,\ 1.003,\ 1.503,\ 4.505,\ 5.005,\ 6.507,\ \ldots$$

$$h_{(x_i)}-y_i:\ 6.408,\ -47.259,\ -2.612,\ 12.875,\ -37.217,\ -3.814,\ \ldots$$

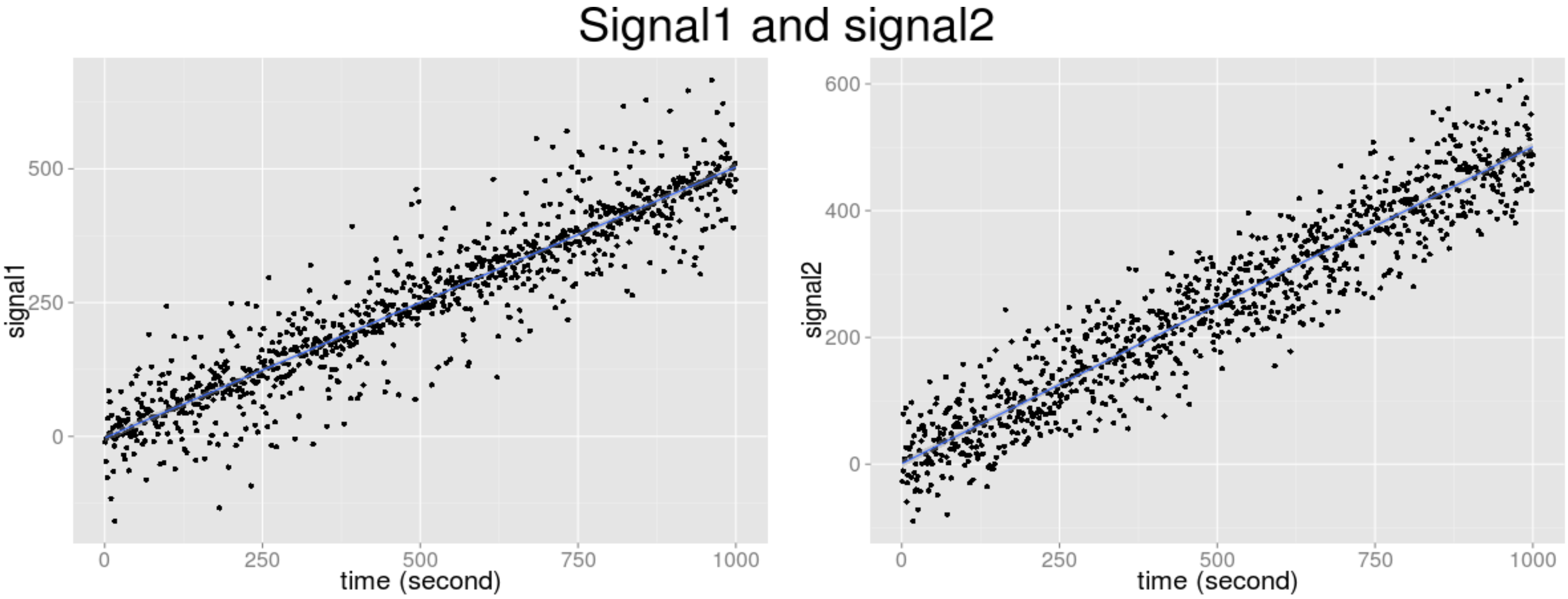
$$|h_{(x_i)}-y_i|:\ 6.408,\ 47.259,\ 2.612,\ 12.875,\ 37.217,\ 3.814,\ \ldots$$

$$MAE = \sqrt{\frac{\sum_{i=1}^{n} |y_i - h(\boldsymbol{x_i})|}{n}} = 23.02094$$

datasciencedojo

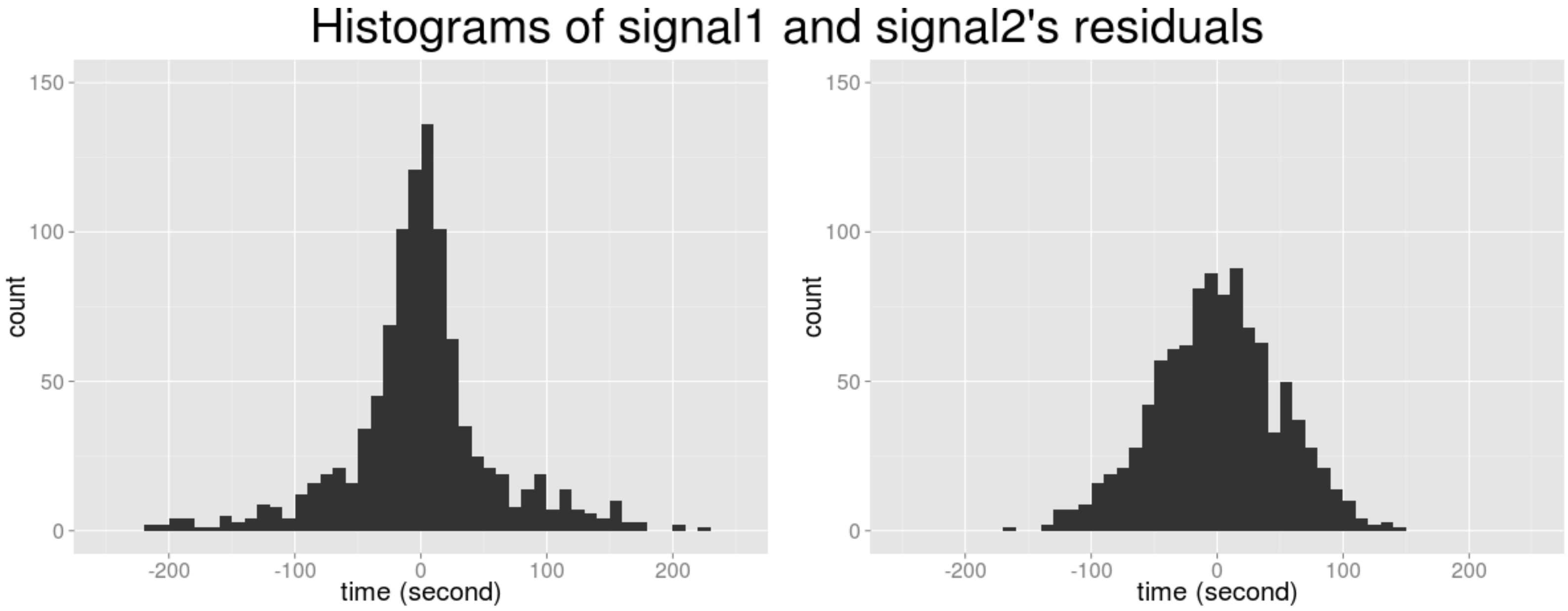
unleash the data scientist in you

MAE *vs.* RMSE

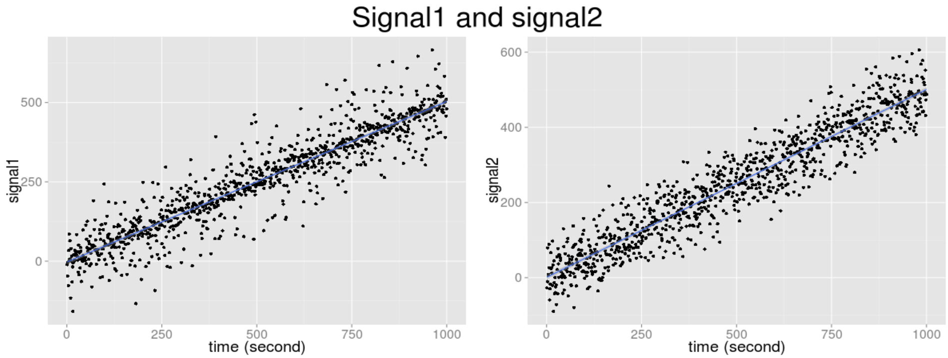


Signal1 and signal2
signal1
500
250
0
0
250
500
750
1000
time (second)
signal2
600
400
200
0
0
250
500
750
1000
time (second)

MAE *vs.* RMSE


Histograms of signal1 and signal2's residuals
150
100
50
0
count
-200
-100
0
100
200
time (second)
150
100
50
0
count
-200
-100
0
100
200
time (second)

MAE *vs.* RMSE



MAE:	41.926 (better)	<	43.199
RMSE:	61.458	>	54.516 (better)

In RMSE, larger deviation cost more

Coefficient of Determination (R squared)

Formula: $R^2 = 1 - \frac{VAR_{res}}{VAR_{tot}}$

(the proportion of variance explained by the model!)

where $VAR_{res} = \sum_{i=1}^{n} \left(y_i - h(\boldsymbol{x_i}) \right)^2 / n$, $VAR_{tot} = \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 / n$

for data set i, $\boldsymbol{x}_i$ is a vector of all the predictors,
y_i is the corresponding response;
$h(\boldsymbol{x}_i)$ is the prediction using a certain model.
$\bar{y}$ is the mean of all responses;
n is the number of data in the test data set.

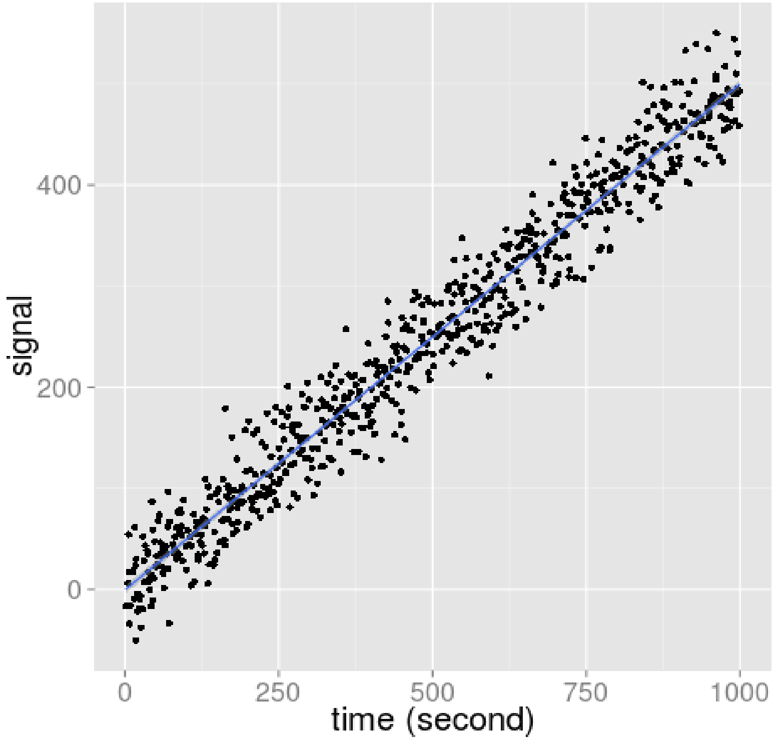
The simple example

$$R^2 = 0.958$$

The model built by real-world data is usually not that good.

$$R^2 = 0.6$$

can be a good model.



Adjusted R squared

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Where p is the total number of predictors in the model (not counting the constant term),
n is the sample size.

Penalized by the more predictors

QUESTIONS