Contents

Regression performance metrics 1

Regression performance metrics

Now the output variable y is continuously valued.

For each input $\mathbf{x_i}$, the model estimates

$$\hat{y_i} = y_i - \epsilon_i$$

where ϵ_i is an error term, also called the **residual**.

RSS Definition: **Residual sum of squares** (RSS), also called **sum of squared residuals** (SSR) and **sum of squared errors** (SSE):

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

RSS increases with n (with more data).

Relative forms of RSS

- RSS per sample, called the **mean squared error** (MSE):

$$\frac{RSS}{n}$$

- Normalized RSS (divide RSS per sample, by sample variance of y):

$$\frac{\frac{RSS}{n}}{s_y^2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y_i})^2}{\sum_{i=1}^{n} (y_i - \overline{y_i})^2}$$

Ratio of *average error of your model* to *average error of prediction by mean*.

R^2: coefficient of determination

$$R^2 = 1 - \frac{\frac{RSS}{n}}{s_y^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y_i})^2}{\sum_{i=1}^{n} (y_i - \overline{y_i})^2}$$

- What proportion of the variance in y is "explained" by our model?
- $R^2 \approx 1$ - model "explains" all the variance in y
-
  - $R^2 \approx 0$ - model doesn't "explain" any of the variance in y
- Depends on the sample variance of y - can't be compared across datasets

R^2: illustration

```
%matplotlib inline

import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import datasets, linear_model, svm, metrics


x, y = datasets.make_regression(n_features=1, noise=5.0, n_samples=50)
regr = linear_model.LinearRegression()
fit = regr.fit(x, y)
y_hat = regr.predict(x)

im = sns.scatterplot(x=x.flatten(),y=y.flatten(), color='gray');
sns.lineplot(x=x.flatten(), y=y_hat, color='red');
im.text(min(x), max(y), "R^2= %f" % metrics.r2_score(y, y_hat) , horizontalalignment='left',
    size='medium', color='red');
```

MSE: mean squared error