Psych 216A: Statistics and data analysis in MATLAB

Lecture 5:
Model accuracy

Kendrick Kay
Quantifying model accuracy

Squared error = 5.4
(dependent on units, hard to interpret)

$R^2 = 75\%$
(independent of units, easy to interpret)
Variance

$$\text{variance} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$
Coefficient of determination ($R^2$)

\[ R^2 = 100 \times \left(1 - \frac{\text{unexplained variance}}{\text{total variance}}\right) \]
Coefficient of determination ($R^2$)

$R^2 = \text{percent explained variance}$

$R^2 = 100 \times \left( \text{fraction explained variance} \right)$

$R^2 = 100 \times \left( 1 - \frac{\text{unexplained variance}}{\text{total variance}} \right)$

$R^2 = 100 \times \left( 1 - \frac{\sum_{i=1}^{n} (d_i - m_i)^2}{n - 1} \right)$

$R^2 = 100 \times \left( 1 - \frac{n - 1}{\sum_{i=1}^{n} (d_i - \bar{d})^2} \right)$

$R^2 = 100 \times \left( 1 - \frac{\sum_{i=1}^{n} (d_i - m_i)^2}{\sum_{i=1}^{n} (d_i - \bar{d})^2} \right)$
Coefficient of determination ($R^2$)

\[ R^2 = 100 \times \left( 1 - \frac{\sum_{i=1}^{n} (d_i - m_i)^2}{\sum_{i=1}^{n} (d_i - \bar{d})^2} \right) \]
Direct calculation of $R^2$ overestimates model accuracy

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>high (80%)</td>
<td>high (79%)</td>
</tr>
<tr>
<td>low (69%)</td>
<td>very high (85%)</td>
</tr>
</tbody>
</table>

Accuracy of fitted model on sample overestimates true accuracy of fitted model.
Overfitting

\[ y = ax^7 + bx^6 + cx^5 + dx^4 + ex^3 + fx^2 + gx + h \]

Maximum polynomial degree = 7

Right graph:
- Underfit
- Overfit
- Optimal

Left graph:
- Data
- True model
- Fitted model

X-axis: \( x \)
Y-axis: \( y \)

Mean squared error

X-axis: Maximum polynomial degree
Y-axis: Mean squared error
Simple models vs. complex models

Simple model (Linear)

Fits well

High noise

Complex model (High-order polynomial)

Fits too much

Low noise

Fits too little

Fits well

Data

True model

Fitted model
Cross-validation

• Goal: estimate true accuracy of a model
• Approach:
  Leave some data out
  Fit model
  Evaluate model on left-out data
Leave-one-out cross-validation

Data

Train
Test

Train and test model

Iteration number

1.
2.
3.
4.
5.
6.

Collect model predictions

Model predictions

Data

Calculate accuracy metric