Experimental Data Analysis

*in ©MATLAB*

**Lecture 8:**
Introduction to models, regression analysis

Jan Rusz
Czech Technical University in Prague
Motivation

Association
Question: Can be increased blood pressure associated with stress?
Answer: Correlation analysis.

Connection
- Correlation indicates a relationship not causality.
- We need to find a connection to say that relationship is causal (i.e. examine that hormonal response to stress can elevate blood pressure).

Prediction
Question: We interrogate the chief suspect (healthy) and measure his blood pressure. How much was the suspect stressed by our key question?
Answer: Regression analysis.
Linear model

Model specification
\[ y = ax + b \]

Fitted model
\[ y = 2x + 3 \]
Supervised learning

Unsupervised learning

Regression

Classification

Linear models

Nonlinear models

(to be covered later)

(to be covered later)
Supervised learning

Input variables

Model

\[ y \]

Output variable

if \( y \) is continuous
(e.g. 0.5, 1, 1.5, 2, ...)
this is regression

if \( y \) is discrete
(e.g. 0 or 1)
this is classification
**Linear model**

\[ y = ax + b \]
\[ y = w_1 x_1 + w_2 x_2 \]
\[ x_2 = 1 \]

**Linearized model**

\[ y = ax^2 + bx + c \]
\[ y = w_1 x_1 + w_2 x_2 + w_3 x_3 \]
\[ x_3 = 1 \]
Nonparametric nonlinear model

moving average filter

Savitzky-Golay filter

nearest-neighbour regression
## Model characteristics

<table>
<thead>
<tr>
<th></th>
<th>Linear</th>
<th>Parametric</th>
<th>Linear in parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear models</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Linearized models</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Parametric nonlinear models</td>
<td>✗ ✓</td>
<td>✓ ✗</td>
<td>✓ ✗</td>
</tr>
<tr>
<td>Nonparametric linear models</td>
<td>✓ ✗</td>
<td>✗ ✓</td>
<td>✓ ✗</td>
</tr>
<tr>
<td>Nonparametric nonlinear models</td>
<td>✗ ✗</td>
<td>✗ ✗</td>
<td>✓</td>
</tr>
</tbody>
</table>
Matrix representation of linear model

Data

\[
\begin{bmatrix}
3 \\
5 \\
3 \\
1 \\
2
\end{bmatrix}
\]

output variable (n x 1)

Model

\[
\begin{bmatrix}
X_1 & X_2 \\
1 & 1 \\
3 & 1 \\
4 & 1 \\
2 & 1 \\
1 & 1
\end{bmatrix}
\]

input variables (n x p)

Parameters

\[
\begin{bmatrix}
w_1 \\
w_2
\end{bmatrix}
\]

weights (1 x p)

\[y = w_1 X_1 + w_2 X_2\]

adjust parameters to fit data
Matrix representation of linear model

\[ y = Xw + e \]
**Squared error**

\[
\text{squared error} = \sum_{i=1}^{n} (d_i - m_i)^2
\]

- **Data**
- **Model**
- **Residuals**

\[
y = w_1 x + w_2
\]

- \(w_1 = 0\)
- \(w_2 = \text{mean}(x) = 2.8\)

\[
y = 2.8
\]

\[
\text{squared error} = 8.8
\]

\[
y = w_1 x + w_2
\]

- \(w_1 = 0.47\)
- \(w_2 = 1.76\)

\[
y = 0.47x + 1.76
\]

\[
\text{squared error} = 7.29
\]
Ordinary least squares solution

\[ X^T e = 0 \]

\[ X^T(y - Xw) = 0 \]

\[ X^T y - X^T X w = 0 \]

\[ w = (X^T X)^{-1} X^T y \]
Fitting nonlinear model based on local, iterative optimization

\[ y = ax^n \]
Quantifying model accuracy

Squared error = 66.4
(dependent on units, hard to interpret)

$R^2 = 44.6\%$
(independent on units, easy to interpret)

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

Total variance

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

Unexplained variance

\[ R^2 = 100 \times \left( 1 - \frac{\text{SE model fit}}{\text{SE model mean}} \right) \]

Squared error = 119.8

Squared error = 66.4

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

\[ R^2 = 100 \times \left( 1 - \frac{n-1}{n-1} \right) \]
Simple linear regression

Multiple linear regression
Research project: Parkinson’s disease (PD), stuttering & L-dopa

Background:
• Excess dopamine theory of stuttering suggests that stuttering may be related to an excess amount of dopamine within the brain
• Some patients with PD develop stuttering in the course of their illness
• Levodopa is precursor of dopamine used to treat motor manifestations of patients with PD

Hypothesis:
• Stuttering is related to extent of L-dopa doses
• Stuttering is not related to motor speech manifestations in PD
Research project: Parkinson’s disease (PD), stuttering & L-dopa

$r = 0.65, p = 0.02$

$r = -0.33, p = 0.25$
Research project: Parkinson's disease (PD), stuttering & L-dopa

Linear regression model:
\[ y \sim 1 + x_1 + x_2 \]

Estimated Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>tStat</th>
<th>pValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.1987</td>
<td>2.4258</td>
<td>0.49413</td>
<td>0.63093</td>
</tr>
<tr>
<td>x1</td>
<td>0.0043749</td>
<td>0.0015763</td>
<td>2.7755</td>
<td>0.018049</td>
</tr>
<tr>
<td>x2</td>
<td>-0.0097572</td>
<td>0.071821</td>
<td>-0.13586</td>
<td>0.89439</td>
</tr>
</tbody>
</table>

Number of observations: 14, Error degrees of freedom: 11
Root Mean Squared Error: 2.46
R-squared: 0.447, Adjusted R-Squared 0.346
F-statistic vs. constant model: 4.44, p-value = 0.0386

How to report results of regression?

Our case: \([F(2,11) = 4.4, p = 0.04, R^2 = 0.45]\)
Some conclusions? :-)

- ANOVA and linear regression analysis are the “same thing”

- Intercept is the mean of the reference group
- The coefficients for the other two groups are the differences in the mean between the reference group and the other groups