**Hypothesis:** there is a relationship between explained and explanatory variables

**Example**

Hypothesis: expected spendings for food in households of married workers with two children depends on their income. This hypothesis seems to be supported by data (GUS data from household budget survey).

<table>
<thead>
<tr>
<th>income of household (0 -1000)</th>
<th>average spendings for food</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 -1000</td>
<td>442</td>
</tr>
<tr>
<td>1000-1500</td>
<td>534</td>
</tr>
<tr>
<td>1500-2000</td>
<td>608</td>
</tr>
<tr>
<td>2000-2500</td>
<td>657</td>
</tr>
<tr>
<td>2500-3000</td>
<td>717</td>
</tr>
<tr>
<td>3000-</td>
<td>817</td>
</tr>
<tr>
<td>average</td>
<td>644</td>
</tr>
</tbody>
</table>

Average share of expenditure for food in income: 0.33
**Auxiliary hypothesis**: relationship between explained and explanatory variables is linear (linear model):

\[ y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{Ki}\beta_K + \varepsilon_i, \text{ for } i = 1, \ldots, N \]

**Example**

Direction of dependence:

\[
\begin{align*}
\text{expenditure}_i &= \beta_1 + \beta_2 \text{income}_i + \varepsilon_i \\
\text{income}_i &= \alpha_1 + \alpha_2 \text{expenditure}_i + \eta_i
\end{align*}
\]
Example

It was found that there is a positive correlation between the ice-cream consumption and the number of drownings in a given day. Does it imply that after eating ice-cream it is not safe to swim?
More drownings happen in warm days as more people are swimming in such a day. For such days consumption of ice-cream is also higher.
Estimation

- Estimates of parameters for two different samples for a model with known parameters $\beta_1 = 8$ and $\beta_2 = 1.5$ (Monte Carlo method)
Example
(cont.) Estimates of $\beta$ is equal to $\mathbf{b} = \begin{bmatrix} 463 \\ 0.08 \end{bmatrix}$.

Fitted values and residuals for the first observation:

\[
\hat{y}_1 = 463 + 0.08 \times 890.6 = 534.9 \\
e_1 = 639.1 - 534.9 = 104.2
\]
Residuals

\[ y(t) = 8 + 1.5x(t) + e(t) \quad e(t) \sim N(0, 2) \]

\[ y_i - \hat{y}_i = e_i \]
OLS with constant and one explanatory variable

Example

<table>
<thead>
<tr>
<th>variable</th>
<th>average</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>q</td>
<td>644.18</td>
<td>46737</td>
</tr>
<tr>
<td>income</td>
<td>2262.34</td>
<td>1584300</td>
</tr>
</tbody>
</table>

Empirical covariance between variables is equal to 126211. Using derived formulas we obtain:

\[
b_2 = \frac{126211}{1584300} = 0.079664
\]

\[
b_1 = 644.18 - 0.079664 \times 2262.34 = 463.95
\]

Result of regression is usually reported with following table:

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>q</td>
<td>.079664</td>
</tr>
<tr>
<td>income</td>
<td>463.95</td>
</tr>
<tr>
<td>constant</td>
<td>463.95</td>
</tr>
</tbody>
</table>
Example

Estimation of the model explaining the expenditure for food with income of household:

\[ q_i = \beta_1 + \beta_2 inc_i + \varepsilon_i \]
Dependence between expenditure for food and income (data and regression line)
Engle curve (data and estimated curve)

\[ e_{inc} \approx \frac{0.8}{0.33} = 0.24 \]
Example ((cont) In the model for food expenditure)

\[ q_i = \beta_1 + \beta_2 \text{inc}_i + \varepsilon_i \]

\( y \) and \( X \) have following form:

\[
\begin{bmatrix}
639.09 \\
664.47 \\
467.55 \\
\vdots
\end{bmatrix}, \quad
\begin{bmatrix}
1 & 890.6 \\
1 & 2300 \\
1 & 1814.5 \\
\vdots & \vdots
\end{bmatrix}
\]

**Notice**: In model with constant first column of matrix \( X \) is column of ones.
e = y − \hat{y} is perpendicular to X!
Decomposition of the total sum of square

- One of the measures of variation is the sum of squares of variable around its mean
- We define following sum of squares:
  - Total Sum of Squares
    \[ TSS = (\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}}) = \sum_{i=1}^{N} (y_i - \bar{y})^2 \]
    where \( \bar{\mathbf{y}} = \mathbf{l}\bar{\mathbf{y}} \)
  - Explained Sum of Squares
    \[ ESS = (\mathbf{\hat{y}} - \bar{\mathbf{y}})'(\mathbf{\hat{y}} - \bar{\mathbf{y}}) = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 \]
    where \( \bar{\mathbf{y}} = \mathbf{l}\hat{\mathbf{y}} \).
  - Residual Sum of Squares
    \[ RSS = \mathbf{e}'\mathbf{e} = \sum_{i=1}^{N} e_i^2 \]
Decomposition of the total sum of squares in model with constant

\[ TSS = ESS + RSS \]

- Total variation can be decomposed into the part which can be explained with the model and the part which cannot be explained with the model.
### Measures of fit: R²

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>33631553</td>
<td>1</td>
<td>33631553</td>
</tr>
<tr>
<td>Residual</td>
<td>122705284</td>
<td>3344</td>
<td>36694.164</td>
</tr>
<tr>
<td>Total</td>
<td>156336837</td>
<td>3345</td>
<td>46737.4701</td>
</tr>
</tbody>
</table>

Number of obs = 3346

F( 1, 397) = 916.54

Prob > F = 0.0000

R-squared = 0.2151

Adj R-squared = 0.2149

Root MSE = 191.56

| q     | Coef.    | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-------|----------|-----------|-------|------|----------------------|
| inc   | .0796627 | .0026314  | 30.27 | 0.000| .0745034 .0848219    |
| _cons | 463.9612 | 6.812147  | 68.11 | 0.000| 450.6048 477.3176    |
Measures of fit: $R^2$

\[
\frac{156336837}{TSS} = \frac{33631553}{ESS} + \frac{122705284}{RSS}
\]

\[
R^2 = \frac{ESS}{TSS} = \frac{33631553}{156336837} = 0.2151
\]
Constrained maximisation

- result of minimization of \( f(x_1, x_2) \) for unconstrained \( x_1, x_2 \) and unconstrained \( x_2 \) but \( x_1 = 0 \)

\[
f(x_1, x_2) = (x_1 - 2)^2 + (x_2 + 2)^2
\]
Measures of fit cont.

- Totally random variable \( z_i \) was added to regression
- \( R^2 \)

<table>
<thead>
<tr>
<th>Model</th>
<th>RSS</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q_i = \beta_1 + \beta_2 \text{inc}_i + \varepsilon_i )</td>
<td>122705284</td>
<td>.2151</td>
</tr>
<tr>
<td>( q_i = \beta_1 + \beta_2 \text{inc}_i + \beta_3 z_i + \varepsilon_i )</td>
<td>122694775</td>
<td>.2152</td>
</tr>
</tbody>
</table>

- Adjusted measure:

\[
\overline{R^2} = 1 - \frac{N - 1}{N - K} (1 - R^2)
\]

- In previous model

<table>
<thead>
<tr>
<th>Model</th>
<th>( R^2 )</th>
<th>( K )</th>
<th>( \overline{R^2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q_i = \beta_1 + \beta_2 \text{inc}_i + \varepsilon_i )</td>
<td>.2151</td>
<td>2</td>
<td>.2149</td>
</tr>
<tr>
<td>( q_i = \beta_1 + \beta_2 \text{inc}_i + \beta_3 z_i + \varepsilon_i )</td>
<td>.2152</td>
<td>3</td>
<td>.2147</td>
</tr>
</tbody>
</table>