Multiple Regression

Example: Food expenditure and family income

Data:
- Sample of 20 households
- Food expenditure (response variable)
- Family income and family size

```
. regress food income

  Source |       SS       df       MS        Number of obs =      20
        |       TSS     R-squared     Roots
        |       RSS
        |       ESS
Model |   4631.17   1.0000  (-2.4191, 0.4094)
  Residual |  4125.01   19      .216
  Total |  8756.18   19

       Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
income |  .1841099   .0149345   12.33   0.000     .1527336    .2154862
_cons  |  -.4119994   .7637666   -0.54   0.596    -2.016613    1.192615

```

```
. regress food number

  Source |       SS       df       MS        Number of obs =      20
        |       TSS     R-squared     Roots
        |       RSS
        |       ESS
Model |     3281.9    1.0000   (0.4094, -2.4191)
  Residual |  4474.27   19      .228
  Total |  7756.18   19

       Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
number |  2.287334   .4224493    5.41   0.000     1.399801    3.174867
_cons  |  1.217365   1.410627    0.86   0.399    -1.746252    4.180981
```

Income

<table>
<thead>
<tr>
<th>Income</th>
<th>Food Expenditure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>40</td>
<td>8</td>
</tr>
<tr>
<td>60</td>
<td>12</td>
</tr>
<tr>
<td>80</td>
<td>16</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
</tr>
</tbody>
</table>

Family Size

<table>
<thead>
<tr>
<th>Family Size</th>
<th>Food Expenditure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
</tr>
</tbody>
</table>
Multiple Regression

**Multiple regression model**

\[
Y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + \ldots + b_p x_{p,i} + \epsilon_i \quad i = 1, \ldots, n
\]

where

- \(Y_i\) response variable
- \(x_{1,i}, \ldots, x_{p,i}\) predictor variables (fixed, nonrandom)
- \(b_0, \ldots, b_p\) regression coefficients
- \(\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)\) error variable

**Example:** Food expenditure and family income

Fitting multiple regression models in STATA:

```
. regress food income number
```

```
Source | SS   df   MS
--------+------------------ Number of obs = 20
Model   | 386.312865  2 193.156433  F(  2,   17) = 121.47
Resid.  | 27.0326365  17 1.59015509 Prob > F = 0.0000
        +-------------------------------------------------------------
Total   | 413.345502  19 21.7550264 R-squared = 0.9346
        +-------------------------------------------------------------
        | food | Coef.  Std. Err.  t    P>|t|  [95% Conf. Interval]
    +-------------------------------------------------------------
income | .1482117  .0163786  9.05  0.000  .1136558   .1827676
number | .7931055  .2444411  3.24  0.005  .2773798   1.308831
_cons  | -1.118295  .6548524 -1.71  0.106 -2.499913   .2633232
```

Multiple Regression, Mar 3, 2004 - 2 -
Multiple Regression

Example: Food expenditure and family income

Data: \((\text{Food}_i, \text{Income}_i, \text{Number}_i), i = 1, \ldots, 20\)

Fitted regression model:

\[
\hat{\text{Food}} = \hat{b}_0 + \hat{b}_1 \text{Income} + \hat{b}_2 \text{Number}
\]

Fitted model is a two-dimensional plane - difficult to visualize.
Inference for Multiple Regression

Multiple regression model (matrix notation)

\[ Y = X b + \varepsilon \]

where

- \( Y \) \( n \) dimensional vector
- \( X \) \( n \times (1 + p) \) dimensional matrix
- \( b \) \( 1 + p \) dimensional vector
- \( \varepsilon \) \( n \) dimensional vector

Thus the model can be written as

\[
\begin{pmatrix}
Y_1 \\
\vdots \\
Y_n
\end{pmatrix} =
\begin{pmatrix}
1 & x_{1,1} & \cdots & x_{p,1} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{1,n} & \cdots & x_{p,n}
\end{pmatrix}
\begin{pmatrix}
b_0 \\
\vdots \\
b_p
\end{pmatrix} +
\begin{pmatrix}
\varepsilon_1 \\
\vdots \\
\varepsilon_n
\end{pmatrix}
\]

Least squares approach: Minimize

\[ \|Y - \hat{Y}\| = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \]

Results:

\[
\hat{b} = (X^T X)^{-1} X^T Y \sim \mathcal{N}(b, \sigma^2 (X^T X)^{-1})
\]

\[
\hat{Y} = X (X^T X)^{-1} X^T Y \sim \mathcal{N}(X b, \sigma^2 (X^T X)^{-1} X^T)
\]

\[
\hat{\varepsilon} = Y - \hat{Y} = (1 - X (X^T X)^{-1} X^T) Y \sim \mathcal{N}(0, \sigma^2 (1 - X (X^T X)^{-1} X^T))
\]

\[
\hat{\sigma}^2 = s_e^2 = \frac{\|Y - \hat{Y}\|^2}{n - p}
\]

Details \sim course in regression analysis (STAT 22200) or econometrics
Example: Food expenditure and family income

Interpretation of regression coefficients

```
. quietly regress food income
. predict e_food1, residuals
. quietly regress number income
. predict e_num, residuals
. regress e_food1 e_num

|          | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|----------|-------|-----------|------|------|----------------------|
| e_num    | .7931055 | .2375541 | 3.34 | 0.004 | .2940229 1.292188   |
```

```
. quietly regress food number
. predict e_food2, residuals
. quietly regress income number
. predict e_inc, residuals
. regress e_food2 e_inc

|          | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|----------|-------|-----------|------|------|----------------------|
| e_inc    | .1482117 | .0159172 | 9.31 | 0.000 | .114771  .1816525   |
```

Result:

- $b_j$ measures the dependence of $Y$ on $x_j$ after removing the linear effects of all other predictors $x_k$, $k \neq j$.

- $b_j = 0$ if $x_j$ does not provide information for the prediction of $Y$ additionally to the information given by the other predictor variables.
Example: Heart catheterization

Description: A Teflon tube (catheter) 3 mm in diameter is passed into a major vein or artery at the femoral region and pushed up into the heart to obtain information about the heart’s physiology and functional ability. The length of the catheter is typically determined by a physician’s educated guess.

Data:

- Study with 12 children with congenital heart defects
- Exact required catheter length was measured using a fluoroscope
- Patient’s height and weight were recorded

Question: How accurately can catheter length be determined by height and length?
**Multiple Regression**

**Example:** Heart catheterization (contd)

Regression model:

\[ Y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon \]

where
- \( Y \) - distance to pulmonary artery
- \( x_1 \) - height
- \( x_2 \) - weight

STATA regression output:

```
. regress distance height weight

Source | SS df MS Number of obs = 12
-------------+------------------------------ F( 2, 9) = 18.62
Model | 578.81613 2 289.408065 Prob > F = 0.0006
Residual | 139.913037 9 15.545893 R-squared = 0.8053
-------------+------------------------------ Adj R-squared = 0.7621
Total | 718.729167 11 65.3390152 Root MSE = 3.9428
-------------+------------------------------

distance | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+---------------------------------------------------------------
height | .1963566 .3605845 0.54 0.599 -.6193422 1.012056
weight | .1908278 .165164 1.16 0.278 -.1827991 .5644547
_cons | 21.0084 8.751156 2.40 0.040 1.211907 40.80489
```

Note:
- Neither height nor weight seem to be significant for predicting the distance to the pulmonary artery.
- The regression on both variables explains 80% of the variation of the response (length of catheder).
**Example:** Heart cathederization (contd)

Consider predicting the length by height alone and by weight alone:

```
. regress distance height
R-squared = 0.7765

|      | Coef. | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|------|-------|-----------|------|------|----------------------|
| height | .5967612 | .1012558  | 5.89 | 0.000 | .3711492 .8223732   |
| _cons | 12.12405  | 4.247174  | 2.85 | 0.017 | 2.660752 21.58734  |
```

```
. regress distance weight
R-squared = 0.7989

|      | Coef. | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|------|-------|-----------|------|------|----------------------|
| weight | .2772687 | .0439881  | 6.30 | 0.000 | .1792571 .3752804  |
| _cons | 25.63746  | 2.004207  | 12.79 | 0.000 | 21.17181 30.10311  |
```

Note:

- In a simple regression of $Y$ on either height or weight, the explanatory variable is highly significant for predicting $Y$.
- In a multiple regression of $Y$ on height and weight, the coefficients for both height and weight are not significantly different from zero.

**Problem:** Explanatory variables are highly linearly dependent (collinear)
Analysis of Variance

Decomposition of variation:

° $SS_{\text{Total}} = \sum_i (Y_i - \bar{Y})^2$ - total variation

° $SS_{\text{Residual}} = \sum_i (Y_i - \hat{Y}_i)^2$ - variation in regression model

° $SS_{\text{Model}} = SS_{\text{Total}} - SS_{\text{Residual}} = \sum_i (\hat{Y}_i - \bar{Y})^2$ - variation explained by regression

Coefficient of determination: The ratio

$$R^2 = \frac{SS_{\text{Model}}}{SS_{\text{Total}}}$$

indicates how well the regression model predicts the response. $R^2$ is also the squared multiple correlation coefficient - in a simple linear regression we have

$$R^2 = \rho_{XY}^2.$$

Example: Heart cathederization

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>578.81613</td>
<td>2</td>
<td>289.408065</td>
<td>F( 2, 9) = 18.62</td>
</tr>
<tr>
<td>Residual</td>
<td>139.913037</td>
<td>9</td>
<td>15.545893</td>
<td>Prob &gt; F = 0.0006</td>
</tr>
<tr>
<td>Total</td>
<td>718.729167</td>
<td>11</td>
<td>65.3390152</td>
<td>R-squared = 0.8053</td>
</tr>
</tbody>
</table>

The coefficient of determination for these data is

$$R^2 = \frac{578.82}{718.73} = 0.81.$$ 

Regression on height and weight explains 81% of the variation of distance.
Analysis of Variance

Question: Is improvement in prediction (decrease in variation) significant?

Our null hypothesis is that none of the explanatory variables helps to predict the response, that is,

\[ H_0 : b_1 = \ldots = b_p = 0 \quad \text{versus} \quad H_a : b_j \neq 0 \text{ for any } j \in \{1, \ldots, p\}. \]

Under the null hypothesis \( H_0 \) the \( F \) statistic

\[ F = \frac{n - p - 1}{p} \cdot \frac{SS_{\text{Model}}}{SS_{\text{Residual}}} = \frac{n - p - 1}{p} \cdot \frac{SS_{\text{Total}} - SS_{\text{Residual}}}{SS_{\text{Residual}}} \]

is \( F \) distributed with \( p \) and \( n - p - 1 \) degrees of freedom.

The null hypothesis \( H_0 \) is rejected at level \( \alpha \) if \( F > F_{p,n-p-1,\alpha} \).

Example: Heart cathederization

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>578.81613</td>
<td>2</td>
<td>289.408065</td>
<td>( F(2,9) = 18.62 )</td>
</tr>
<tr>
<td>Residual</td>
<td>139.913037</td>
<td>9</td>
<td>15.545893</td>
<td>Prob &gt; F = 0.0006</td>
</tr>
<tr>
<td>Total</td>
<td>718.729167</td>
<td>11</td>
<td>65.3390152</td>
<td>R-squared = 0.8053</td>
</tr>
</tbody>
</table>

The value of the \( F \) statistic is

\[ F = \frac{9}{2} \cdot \frac{578.82}{139.91} = 18.61. \]

The critical value for rejecting \( H_0 : b_1 = b_2 = 0 \) is \( F_{2,9,0.05} = 4.26. \) Thus the null hypothesis \( H_0 \) that both coefficients \( b_1 \) and \( b_2 \) are zero is rejected at significance level \( \alpha = 0.05 \).
Comparing Models

Example: Cobb-Douglas production function

\[ Y = t \cdot K^a \cdot L^b \cdot M^c \]

where

- \( Y \) - output
- \( K \) - capital
- \( L \) - labour
- \( M \) - materials

Regression model:

\[ \log Y = \log t + a \log K + b \log L + c \log M \]
Comparing Models

Example: Cobb-Douglas production function (contd)

Regression model $M_0$ for Cobb-Douglas function:

$$\log Y = \log t + a \log K + b \log L + c \log M$$

```
. regress LY LK LM LL
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1.35136742</td>
<td>3</td>
<td>.450455808</td>
<td>F( 3, 21) = 138.98</td>
</tr>
<tr>
<td>Residual</td>
<td>.068065609</td>
<td>21</td>
<td>.003241219</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>1.41943303</td>
<td>24</td>
<td>.059143043</td>
<td>R-squared = 0.9520</td>
</tr>
</tbody>
</table>

| LY | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|----|-------|-----------|------|------|----------------------|
| LK | .0718626 | .1543912  | 0.47 | 0.646 | -.2492114 .3929366  |
| LM | .7072231 | .3004146  | 2.35 | 0.028 | .0824768 1.331969 |
| LL | .2117778 | .4248755  | 0.50 | 0.623 | -.6717991 1.095355 |
| _cons | .0347117 | .0374354  | 0.93 | 0.364 | -.0431395 .1125629 |

Two variables, $\log K$ and $\log L$, do not improve prediction of $\log Y$.

$\rightarrow$ alternative model $M_1$

$$\log Y = \log t + c \log M$$

```
. regress LY LM
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1.34977753</td>
<td>1</td>
<td>1.34977753</td>
<td>F( 1, 23) = 445.69</td>
</tr>
<tr>
<td>Residual</td>
<td>.069655501</td>
<td>23</td>
<td>.0030285</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>1.41943303</td>
<td>24</td>
<td>.059143043</td>
<td>R-squared = 0.9509</td>
</tr>
</tbody>
</table>

| LY | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|----|-------|-----------|------|------|----------------------|
| LM | .9086794 | .0430421  | 21.11 | 0.000 | .81964 .9977188 |
| _cons | .0512244 | .0189767  | 2.70 | 0.013 | .011968 .0904808 |

Question: Is model $M_0$ significantly better than model $M_1$?
Comparing Models

Consider the multiple regression model with \( p \) explanatory variables

\[
Y_i = b_0 + b_1 x_{1,i} + \ldots + b_p x_{p,i} + \varepsilon_i.
\]

**Problem:**
Test the null hypothesis

\( H_0: \) \( q \) specific explanatory variables all have zero coefficients

versus

\( H_a: \) any of these \( q \) explanatory variables has a nonzero coefficient.

**Solution:**

- Regress \( Y \) on all \( p \) explanatory variables and read \( SS_{\text{Residual}}^{(1)} \) from the output.
- Regress \( Y \) on just \( p - q \) explanatory variables that remain after you remove the \( q \) variables from the model. Read \( SS_{\text{Residual}}^{(2)} \) from the output.
- The test statistic is

\[
F = \frac{n - p - 1}{q} \cdot \frac{SS_{\text{Residual}}^{(2)} - SS_{\text{Residual}}^{(1)}}{SS_{\text{Residual}}^{(1)}}.
\]

Under the null hypothesis, \( F \) is \( F \) distributed with \( q \) and \( n - p - 1 \) degrees of freedom.

- Reject if \( F > F_{q,n-p-1,\alpha} \).
Comparing Models

Example: Cobb-Douglas production function

Comparison of models $M_0$ and $M_1$:

- $M_0$: $SS_{\text{Residual}}^{(0)} = 0.06807$ and $n - p - 1 = 21$.
- $M_1$: $SS_{\text{Residual}}^{(1)} = 0.06966$ and $q = 2$.

$$F = \frac{21}{2} \cdot \frac{0.06966 - 0.06807}{0.06807} = 0.2453$$

- Since $F < F_{2,21,0.05} = 3.47$ we cannot reject $H_0: a = b = 0$.

Using STATA:

```
. test LK LL
   ( 1) LK = 0
   ( 2) LL = 0
   F( 2, 21) = 0.25
   Prob > F = 0.7847

. test LK LL _cons
   ( 1) LK = 0
   ( 2) LL = 0
   ( 3) _cons = 0
   F( 3, 21) = 2.43
   Prob > F = 0.0934
```