# Logistic regression instructions

Note that currently, the tool performs only binary logistic analysis with a binary outcome variable (i.e. linguistic variable with two categories).

**1) The tool expects input in the following format (only a small part of the dataset is shown).**

|  | A | B | C | D | F |
|---|---|---|---|---|---|
| 1 | ID | Variety | Genre | Subject | Outcome |
| 2 | 1 | B_BR | D_Press | A_other | A_other |
| 3 | 2 | B_BR | D_Press | A_other | B_must |
| 4 | 3 | B_BR | D_Press | A_other | A_other |
| 5 | 4 | B_BR | D_Press | A_other | A_other |
| 6 | 5 | B_BR | D_Press | A_other | A_other |
| 7 | 6 | B_BR | D_Press | A_other | A_other |
| 8 | 7 | B_BR | D_Press | A_other | A_other |
| 9 | 8 | B_BR | D_Press | C_you | A_other |
| 10 | 9 | B_BR | D_Press | A_other | A_other |
| 11 | 10 | B_BR | D_Press | A_other | B_must |
| 12 | 11 | B_BR | D_Press | A_other | A_other |
| 13 | 12 | B_BR | D_Press | A_other | B_must |
| 14 | 13 | B_BR | D_Press | A_other | A_other |
| 15 | 14 | B_BR | D_Press | A_other | A_other |
| 16 | 15 | B_BR | D_Press | A_other | A_other |
| 17 | 16 | B_BR | D_Press | A_other | A_other |

Each line represents a single observation (i.e. occurrence of the linguistic feature in corpus).

One binary linguistic variable ('Outcome') and three predictor variables.

**2) One of the two options can be selected:**

1. Select what you want to do. For help click here.

> Build a model ▼
> **Build a model**
> Build a model stepwise

The first option (default) selects a block entry of predictors (all relevant predictors at once); the other option selects step-wise predictor entry, which is useful especially for exploratory studies.

3a) Follow the instructions on the screen and enter all relevant pieces of information. Press 'Build model'.

1. Select what you want to do. For help click here.

   Build a model ▼

2. Paste data in the text area.

```
1938   A_AM   B_Fiction   A_other A_other
1939   A_AM   B_Fiction   B_I     A_other
1940   A_AM   B_Fiction   A_other B_must
1941   A_AM   B_Fiction   A_other A_other
1942   A_AM   B_Fiction   B_I     A_other
1943   A_AM   B_Fiction   A_other A_other
1944   A_AM   B_Fiction   C_you   A_other
1945   A_AM   B_Fiction   A_other A_other
1946   A_AM   B_Fiction   A_other A_other
1947   A_AM   B_Fiction   A_other A_other
1948   A_AM   B_Fiction   B_I     A_other
1949   A_AM   B_Fiction   C_you   A_other
1950   A_AM   B_Fiction   C_you   A_other
1951   A_AM   B_Fiction   A_other A_other
```

3. Type in the exact name of the outcome variable. Outcome

4. Type in the exact name(s) of the predictor(s) [use ; as separator]. Variety; Genre

5. Decide if you want to include predictor interactions.

   ○ Yes, include all   ● Yes, include some   ○ No

6. Type in the exact names of the predictors with interactions [use ; as separator]. Variety;Genre

Build model   Clear

> The interactions setting is optional. If interested only in main effects, answer 'No' in Step 5.

4a) The following is an example of the output:

> Summary stats for the whole model.

**Output**

Overall model statistics: Likelihood ratio test (LL): 145.56 (p < 0.0001) -> SIGNIFICANT; C-index: 0.66 -> NOT ACCEPTABLE; Nagelkerke $R^2$: 0.1; AIC: 2515.61

Significance          Effect size

Coefficients:

> Effect of individual predictors

| | Estimate (log odds) | Standard Error | Z value (Wald) | p-value | Estimate (odds) | 95% CI lower | 95% CI upper |
|---|---|---|---|---|---|---|---|
| (Intercept) | 0.542 | 0.200 | 2.714 | 0.007 | 1.720 | 1.168 | 2.561 |
| VarietyB_BR | -0.930 | 0.290 | -3.210 | 0.001 | 0.395 | 0.222 | 0.693 |
| GenreB_Fiction | -1.318 | 0.242 | -5.441 | 0.000 | 0.268 | 0.165 | 0.428 |
| GenreC_General | -0.450 | 0.228 | -1.974 | 0.048 | 0.638 | 0.406 | 0.993 |
| GenreD_Press | -1.191 | 0.268 | -4.445 | 0.000 | 0.304 | 0.179 | 0.511 |
| SubjectB_I | -1.084 | 0.174 | -6.232 | 0.000 | 0.338 | 0.239 | 0.472 |
| SubjectC_you | -0.917 | 0.158 | -5.794 | 0.000 | 0.400 | 0.291 | 0.542 |
| VarietyB_BR:GenreB_Fiction | 1.482 | 0.340 | 4.353 | 0.000 | 4.400 | 2.267 | 8.620 |
| VarietyB_BR:GenreC_General | 0.952 | 0.324 | 2.941 | 0.003 | 2.592 | 1.379 | 4.915 |
| VarietyB_BR:GenreD_Press | 1.490 | 0.379 | 3.927 | 0.000 | 4.438 | 2.118 | 9.384 |

Regression equation (predicted probabilities):

$$\text{probability of } y = \frac{e^{(0.542 + -0.93 \times x_1 + -1.318 \times x_2 + -0.45 \times x_3 + -1.191 \times x_4 + -1.084 \times x_5 + -0.917 \times x_6 + 1.482 \times x_7 + 0.952 \times x_8 + 1.49 \times x_9)}}{1 + e^{(0.542 + -0.93 \times x_1 + -1.318 \times x_2 + -0.45 \times x_3 + -1.191 \times x_4 + -1.084 \times x_5 + -0.917 \times x_6 + 1.482 \times x_7 + 0.952 \times x_8 + 1.49 \times x_9)}}$$

> Equation for calculating predicted probabilities.

3b) Follow the instructions on the screen and enter all relevant pieces of information. Press 'Build model'.

### 2. Paste data in the text area.

```
ID      Variety Genre   Subject Outcome
1       B_BR    D_Press A_other A_other
2       B_BR    D_Press A_other B_must
3       B_BR    D_Press A_other A_other
4       B_BR    D_Press A_other A_other
5       B_BR    D_Press A_other A_other
6       B_BR    D_Press A_other A_other
7       B_BR    D_Press A_other A_other
8       B_BR    D_Press C_you   A_other
9       B_BR    D_Press A_other A_other
10      B_BR    D_Press A_other B_must
11      B_BR    D_Press A_other A_other
12      B_BR    D_Press A_other B_must
13      B_BR    D_Press A_other A_other
14      B_BR    D_Press A_other A_other
```

### 3. Type in the exact name of the outcome variable. Outcome

### 4. Decide on the method of step-wise model buil

⦿ Forward ◯ Backward ◯ Hybrid

> 'Forward' means adding predictors one by one to a null model (with no predictors); 'Backward means deleting predictors one by one from a full model (with all predictors); 'Hybrid' combines these two procedures allowing adding or deleting predictors at each stage.

**Build model**   **Clear**

4a) The following is an example of the output:

**Output**

Summary of the forward step-wise procedure:

> Overview of all the steps taken to arrive at the 'best model'. '+' indicates adding one variable at a time until AIC stops decreasing. '-' indicates deleting one variable at a time until AIC stops decreasing. The smaller the AIC value, the better the model.

| Step | Df | Deviance | Residual df | Residual deviance | AIC |
|------|-----|----------|-------------|-------------------|------|
| 1 | NA | NA | 1950 | 2641.168 | 2643.168 |
| 2+Subject | -2 | 87.446051 | 1948 | 2553.721 | 2559.721 |
| 3+Genre | -3 | 32.571310 | 1945 | 2521.150 | 2533.150 |
| 4+Variety | -1 | 2.554401 | 1944 | 2518.596 | 2532.596 |

Best model - overall statistics: Likelihood ratio test (LL): 122.57 ($p < 0.0001$) -> SIGNIFICANT; C-index: 0.64 -> NOT ACCEPTABLE; Nagelkerke $R^2$: 0.08; AIC: 2532.6

> Summary stats for the whole model.

Coefficients:

| | Estimate (log odds) | Standard Error | Z value (Wald) | p-value | Estimate (odds) | 95% CI lower | 95% CI upper |
|---|---|---|---|---|---|---|---|
| (Intercept) | 0.030 | 0.148 | 0.202 | 0.840 | 1.030 | 0.770 | 1.379 |
| SubjectB_I | -1.060 | 0.173 | -6.120 | 0.000 | 0.346 | 0.245 | 0.483 |
| SubjectC_you | -0.905 | 0.157 | -5.744 | 0.000 | 0.405 | 0.295 | 0.548 |
| GenreB_Fiction | -0.590 | 0.169 | -3.485 | 0.000 | 0.554 | 0.398 | 0.772 |
| GenreC_General | -0.015 | 0.160 | -0.097 | 0.923 | 0.985 | 0.720 | 1.347 |
| GenreD_Press | -0.467 | 0.186 | -2.504 | 0.012 | 0.627 | 0.435 | 0.903 |
| VarietyB_BR | 0.153 | 0.096 | 1.597 | 0.110 | 1.165 | 0.966 | 1.405 |

> Effect of individual predictors (see above).

[Notes]